CURVE ESTIMATION AND SIGNAL DISCRIMINATION IN SPATIAL PROBLEMS

Christian Rau

A thesis submitted for the degree of Doctor of Philosophy of
The Australian National University

February 2003

(Revised September 2003)

Declaration

Unless otherwise specified in the text, this thesis describes my own work, carried out under the supervision of Professor Peter Hall, who is credited for the formulation of Theorems 2.1, 3.1, 4.1 and 4.2, as well as a major part of the theoretical arguments in Sections 2.4, 3.5 and 4.3. For those parts of the above proofs, and other remarks, which draw on the material in the Appendix, I estimate my contribution at no less than 75 per cent on average. I also acknowledge the contributions of Professors Don Poskitt (Monash University, Melbourne) and Brett Presnell (University of Florida), whose joint work with Peter Hall formed the basis for Subsection A.5.3. A suite of MATLAB© code authored by Brett Presnell enabled me to tackle the dataset of Chapter 6 much more quickly than would otherwise have been possible, and a number of discussions with the previously mentioned individuals influenced the exposition in that chapter. The dataset used there was made available to me by Dr Danny Gibbins and Professor Doug Gray (Cooperative Research Centre for Sensor Signal and Information Processing, Adelaide), who helped me a great deal to understand the nature of the data and its inherent problems during three visits and through other communication.

Christian Rau

Acknowledgements

I would like to start by thanking my supervisor, Peter Hall, whose academic guidance (dating back to the time before I became a student at the Mathematical Sciences Institute, formerly the School of Mathematical Sciences), constant encouragement and vision were indispensable in writing this thesis. Thanks to his financial assistance, I was able to attend a number of conferences, and widen my research focus through participation in courses and interdisciplinary work. I also gratefully acknowledge financial support of an Overseas Postgraduate Research Scholarship, an Australian National University PhD Scholarship, and an ANU Supplementary Scholarship during my candidature. Several postdoctoral fellows, among whom I would particularly like to mention Professor Liang Peng (Georgia Institute of Technology), and visitors, especially Professor Peihua Qiu (University of Minnesota), have inspired my research through generously sharing their ideas.

My sincere thanks go to David Hirst and Reza Pakyari for sharing an office with me during my course, and the staff and other students at the Mathematical Sciences Institute, whose kindness provided a very hospitable study environment. I was particularly glad to have the patient assistance of the IT Unit in the MSI available whenever I needed it. Also, Dr Margaret Kahn and several other staff members of the ANU Supercomputing Facility provided me a lot of help and advice in using these resources.

That my life as an overseas student was such a rewarding one, not only in academic terms, is due to the many kind-hearted people that I was glad to meet inside and outside university. My special thanks go to Yvonne Heslop and the University II Toastmasters Club, to S. and G.C., and among my friends from many places all over the world, to M.T.N. and S.L.W.

Most of all, I would like to express my deepest gratitude to my mother and three siblings for their constant love and manifold support.

Abstract

In many instances arising prominently, but not exclusively, in imaging problems, it is important to condense the salient information so as to obtain a low-dimensional approximant of the data. This thesis is concerned with two basic situations which call for such a dimension reduction. The first of these is the statistical recovery of smooth edges in regression and density surfaces. The edges are understood to be contiguous curves, although they are allowed to meander almost arbitrarily through the plane, and may even split at a finite number of points to yield an edge graph. A novel locally-parametric nonparametric method is proposed which enjoys the benefit of being relatively easy to implement via a 'tracking' approach. These topics are discussed in Chapters 2 and 3, with pertaining background material being given in the Appendix. In Chapter 4 we construct concomitant confidence bands for this estimator, which have asymptotically correct coverage probability. The construction can be likened to only a few existing approaches, and may thus be considered as our main contribution.

Chapter 5 discusses numerical issues pertaining to the edge and confidence band estimators of Chapters 2–4. Connections are drawn to popular topics which originated in the fields of computer vision and signal processing, and which surround edge detection. These connections are exploited so as to obtain greater robustness of the likelihood estimator, such as with the presence of sharp corners.

Chapter 6 addresses a dimension reduction problem for spatial data where the ultimate objective of the analysis is the discrimination of these data into one of a few pre-specified groups. In the dimension reduction step, an instrumental role is played by the recently developed methodology of functional data analysis. Relatively standard non-linear image processing techniques, as well as wavelet shrinkage, are used prior to this step. A case study for remotely-sensed navigation radar data exemplifies the methodology of Chapter 6.

Contents

Chapter 1

Introduction

1.1 Definitions and Notational Framework

In this section, which can be skipped at first reading and referred to later if necessary, we collect notational conventions for referencing purposes, as well as some definitions and facts which will be used in this thesis. Consistency of notation is usually maintained throughout. Exceptions to this should be clear from the context, if not explicitly stated. The numbering of theorems, figures *et cetera* will be according to the chapter in which they appear, while formula numbers refer to chapter and section. For example, equation (2.4.3) is the third equation in Section 2.4 of Chapter 2.

For the number sets we use the notations $\mathbb{N} = \{1, 2, \dots\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\mathbb{R}_+ = [0, \infty)$, and $\mathbb{R}_+^* = (0, \infty)$. The origin in $\mathbb{R}^d$, for $d \geq 1$, is (also) denoted O. For a vector $z \in \mathbb{R}^2$, we let $\arg(z)$ denote its argument as a complex number, taken here in the interval $[-\pi/2, 3\pi/2)$. We write $\subset$ for *strict* set-theoretic inclusion, so that $A \subset B$ means $A \subseteq B$ and $A \neq B$. The indicator function of the set or (probabilistic) event A is defined as

$$I_A(x) = \begin{cases} 1, & x \in A\,, \\ 0, & \text{otherwise.} \end{cases}$$

Sometimes we also write $I(x \in A) = I_A(x)$. A *black-and-white image* is the indicator function of a compact set $A \subset \mathbb{R}^2$. For two sets $A_1, A_2 \subset \mathbb{R}^d$ ($d \in \mathbb{N}$), their symmetric difference is defined as

$$A_1 \triangle A_2 = (A_1 \setminus A_2) \cup (A_2 \setminus A_1)\,.$$

The two morphological operations of *dilation* and *erosion* of a set A by a set B, with

$A, B \subset \mathbb{R}^d$, are defined as the sets

$$A \oplus B = \{a + b : a \in A,\, b \in B\} \qquad \text{and}$$
$$A \ominus B = \{x \in \mathbb{R}^d : \forall\, b \in B,\, x - b \in A\},$$

respectively; in particular, $A \ominus B \subseteq A \subseteq A \oplus B$ if $O \in B$. We write $-A = \{0\} \ominus A$. Through the identification $A \leftrightarrow I_A$, the operators $\oplus, \ominus$ can also be defined on black-and-white images. For $w \in \mathbb{R}$ and $\lambda \in \mathbb{R}_+^*$, the *hard* and *soft thresholding operators* are respectively defined as

$$\eta_{\text{Hard}}(w, \lambda) = \begin{cases} w, & |w| > \lambda\,, \\ 0, & |w| \le \lambda\,, \end{cases}$$
$$\eta_{\text{Soft}}(w, \lambda) = \operatorname{sgn}(w)(|w| - \lambda)_+\,,$$

where $x_+ = \max(x, 0)$.

For $z \in \mathbb{R}$, we write $\langle z \rangle$ for the largest integer less than or equal to z. A real-valued function f is called *Lipschitz continuous* on the compact domain $G \subset \mathbb{R}^d$ (where we are mainly interested in the case $d = 2$) if for any $x, x' \in G$, there exists a constant $L_1 \in \mathbb{R}_+^*$ such that $|f(x) - f(x')| \le L_1 \|x - x'\|$. If instead

$$|f(x) - f(x')| \le L_1 \|x - x'\|^\beta\,, \quad \beta \in (0, 1)\,, \tag{1.1.1}$$

then f is called *Hölder continuous* with index β and Hölder constant L_1. The Hölder continuous functions with index $\beta \in \mathbb{R}_+^* \setminus \mathbb{N}$ are defined as those whose kth derivative $f^{(k)}$ exists, where $k = \langle \beta \rangle$, and satisfies (1.1.1) with $f^{(k)}$ and $\beta - k$ replacing f and β, respectively. If $\beta \in \mathbb{N}_0$ then Hölder continuity with index β means existence and continuity of all partial derivatives of order β. The function $f : G \subseteq \mathbb{R}^d \to \mathbb{R}$ is called *upper semicontinuous* if for each $x \in G$ and sequence $\{x_n\} \subset G$ such that $x_n \to x$ and $x_n \neq x$ for all n, the inequality $\limsup_{n\to\infty} f(x_n) \le f(x)$ holds. If $x = (x_1, \ldots, x_d), y = (y_1, \ldots, y_d) \in \mathbb{R}^d$, the inequality $x \le y$ means that $x_i \le y_i$ for $i = 1, \ldots, d$. The usual scalar (inner) product of x and y is $x \cdot y = \sum_{i=1}^d x_i y_i$. For vectors as well as matrices, the prime symbol $(\cdot)'$ denotes the transpose as usual. Vectors will be regarded as either rows or columns, whatever is notationally more convenient. For a vector $v \in \mathbb{R}^d$, the Euclidean norm is denoted by $\|v\|$, and the area (Lebesgue measure) of a measurable set $A \subset \mathbb{R}^2$ is likewise denoted by $\|A\|$ since no confusion is possible. The tensor product of

two scalar-valued functions $f_i : D_i \subseteq \mathbb{R}$ $(i = 1, 2)$ is defined as

$$\big(f_1 \otimes f_2\big)(x^{(1)}, x^{(2)}) = f_1(x^{(1)})\, f_2(x^{(2)})\,.$$

The topological interior and closure and boundary of a set $A \subset \mathbb{R}^d$ are denoted by A°, $\bar{A}$ and ∂A respectively. A compact set $A \subset \mathbb{R}^d$ is called *star-shaped* if there exists $x_0 \in A$ such that for each point $y \in \partial A$, the line segment $\mathcal{L}(x_0, y) = \big\{\lambda\, x_0 + (1-\lambda)y :\ 0 \le \lambda \le 1\big\}$ lies entirely in A. In the case that one may take $x_0 = O$, the *radial function* (commonly also referred to as the *distance function*) $F : \mathbb{R}^2 \to \mathbb{R}_+$ is defined by

$$F(u) = \inf\big\{\tau > 0 : u/\tau \in A\big\}\,.$$

The compact set A is called *strictly star-shaped* if for one (and hence any) $x \in A^\circ$, the radial function of $A \oplus (-x)$ is continuous. For $x \in \mathbb{R}^2$ and $h > 0$, the set

$$\mathrm{ball}(x, h) = \big\{y \in \mathbb{R}^2 :\ \|y - x\| < h\big\}$$

is the open disc of radius h, centred at x.

From probability theory we shall frequently use standard tools, such as the Borel-Cantelli lemma and Markov's inequality. For brevity, the terms 'random variable,' 'distribution function' and 'independent and identically distributed' will be written as r.v., d.f. and i.i.d., respectively. If the d.f.s of two random variables X, Y coincide, this will be denoted $X \stackrel{\mathrm{d}}{=} Y$. We write $X \to_P Y$ $(X \Longrightarrow Y)$ if X converges in probability (respectively, weakly or in distribution) to Y, assuming in the former case that X and Y are defined on a common probability space. A broader notion of weak convergence that will be useful in this thesis is recapitulated in Section A.3. A family $\{X_\alpha,\, \alpha \in I\}$ of r.v.s with values in a metric space is called *tight* if for every $\epsilon > 0$, there exists a compact set K such that $P(X_\alpha \in K) \ge 1 - \epsilon$ for all α, and *uniformly integrable* if $\sup_\alpha E\big[|X_\alpha| I(|X_\alpha| > c)\big] \to 0$ for $c \to \infty$. It is convenient to give here a version of an exponential inequality for sums of r.v.s, which is used in the proof of Theorem 2.1.

Theorem 1.1. Bernstein's Theorem (Pollard, 1984, p. 193) *Let $\{X_i\}$ denote a sequence of independent scalar random variables with $E(X_i) = 0$ and $|X_i| \le M < \infty$, and let $S_n = n^{-1} \sum_{i=1}^n X_i$ and $V = \sum_{i=1}^n \mathrm{var}(X_i)$. Then, for $\eta > 0$,*

$$P\big\{|S_n| \ge \eta\big\} \le 2 \exp\big\{-\tfrac{1}{2}\,\eta^2/(V + \tfrac{1}{3}M\eta)\big\}\ .$$

In this thesis we shall also need a modicum of the theory of *point processes*. A (planar)

point process is a map N from some probability space $(\Omega, \mathcal{F}, P)$ into the space of locally finite counting measures on $\mathbb{R}^2$. Then $\{\Omega \ni \omega \mapsto \big(N(\omega)\big)(A)\}$ is an a.s. $\mathbb{N}_0$-valued r.v., with A ranging over all compact sets in $\mathbb{R}^2$. It is usual to denote a point process by the entirety of points comprising its support, say $\mathcal{X} = \{X_i\}$, and to write

$$N(A) = \sum_i I(X_i \in A) = \text{card}(\{i : X_i \in A\}) .$$

The arrangement of the X_is can be deterministic, for example triangular, square or hexagonal, with the square lattice or *regular grid*, defined at (1.2.5), being the most important variant in the analysis of digitised images. In the context relevant to this thesis, the process $\mathcal{X}$ is interpreted as representing the loci of observations on the image. An *image estimator* is a function which takes its values in the space of bivariate surfaces defined over the observation area, and measurable with respect to the σ-algebra generated jointly by $\mathcal{X}$ and the system $\mathcal{Y} = \{Y_i\}$, where $Y_i \in \mathbb{R}$ is interpreted as the available observation on the surface at location X_i.

Usually a point process has a random nature, and a special role in the analysis of scatterplots (e.g. Korostelev and Tsybakov, 1993) is played by the (*homogeneous*) *Poisson process*. A Poisson process on Π is a random point process with the following two properties:

1. For each subset $A \subseteq \Pi$ with $\|A\| < \infty$, the r.v. $N(A)$ is Poisson-distributed with parameter $\lambda\|A\|$.

2. For any $k \geq 1$ and disjoint sets $A_1, \ldots, A_k \subset \Pi$, the r.v.s $N(A_1), \ldots, N(A_k)$ are independent.

The parameter $\lambda > 0$ is called the *intensity* of $\{X_i\}$. The celebrated properties of the Poisson process, such as invariance under Euclidean motions and conditional uniformity, will be used without further comment. Replacing the expression $\lambda\|A\|$ in property 1 above by $\int_A \lambda(x)\, dx$ leads to the definition of the general (inhomogeneous) Poisson process with rate function $\{\lambda(x),\, x \in \Pi\}$. Occasionally, we shall use the more general notion of a *Poisson cluster process*, which is defined as follows. Let $\mathcal{X}_1 = \{X_i\}$ be a Poisson process with rate function $\lambda = n\lambda_1$, where λ_1 is a non-random continuous function that is bounded away from zero. Let $\mathcal{X}_2$ be any point process with the property that its expected number of points, μ, is finite. For example, the points in $\mathcal{X}_2$ might be randomly distributed within a fixed region, or placed deterministically at the vertices of a regular lattice in the region, or some combination of these two possibilities. Conditional on $\mathcal{X}_1$,

let $\{\mathcal{X}_{2i} : i \geq 1\}$ be a sequence of independent copies of $\mathcal{X}_2$; and let $\mathcal{X}_\nu$ be the point process represented by the union of the points in $\mathcal{X}_{2i} \oplus X_i$ (for $i \geq 1$). Then the Poisson cluster process $\mathcal{X}_\nu$ has intensity $n\lambda$, where $\lambda = \mu\lambda_1$.

1.2 Introduction

This thesis is primarily concerned with problems emerging in the context of estimating smooth boundary curves in bivariate surfaces, and with methods for discriminating such surfaces on the basis of representations that are likewise 'economical,' albeit in a sense that is less directly related to visual perception. In both cases the estimation problem arises primarily, though not exclusively, through the presence of additive random distortions. The terms *fault lines* or *edges* will also be used interchangeably with 'boundary curves' in the sequel. Bivariate surfaces will often be called *images*, whether sampling takes place on a grid or is governed by an irregular (random) process. Other than for brevity, this convention serves to emphasise the strong links of the topics discussed in this thesis with the field of computer vision. Within the sizeable literature from this and adjacent areas that continually contribute to the problems discussed in this thesis, some salient papers are Marr and Hildreth (1980), Marr (1982), and Huertas and Medioni (1986). Papers reflecting more recent developments are cited in Hall, Qiu and Rau (2002), in addition to other papers cited later in this thesis.

At an early interpretation level, computational vision systems frequently analyse images by detecting and appropriately representing boundary curves of the object(s) present in the scene. Other information, such as the number and size of such objects, or much more complex characteristics which are relevant to recognition tasks, may then be extracted at a higher level of a possibly much more complex scheme of image processing steps.

The philosophy that extracting edges should be the primary task in image processing has largely been influenced by progress in understanding human and animal vision. Work by Marr (1982) and his collaborators (Marr and Poggio, 1979; Marr and Hildreth, 1980) has played a pioneering role in the early development of the field. However, important differences between the approaches from the computer vision and statistics areas do exist, and these will be elaborated upon later. In this thesis, we shall use the term *edge detection* to subsume the notions, and the approaches relevant to the indicated problem, from the sides of either computer vision or statistics. Although one of the aims of this thesis is to suggest how approaches developed from these two views can expeditiously interact, our most salient contributions exclusively relate to the statistical aspects of edge

detection.

Though certainly important enough by itself, the preponderance of work dedicated to edge detection may have obscured the advantages of computing other features from images. The first of these features, namely the estimation of *ridges*, will be seen to provide the framework for the edge estimators proposed in Chapter 3 and 4. A deeper statistical analysis of differential geometric properties of curves in noise-corrupted surfaces is also enabled by our methods, though it is not a major topic here. Secondly, for the situation where a whole dataset of images is available, we consider the estimation of (functional) *principal component scores* associated with a particular image within the dataset. In this context, the interpretation of a 'feature' is more in a data-analytic than in a visual-perceptional sense. In spite of this, all three cases of edge and ridge detection, and principal components analysis, may be regarded as instances of *dimension reduction*, which therefore underpins all the results in this thesis. A further link between the two parts of the thesis consists in our use of the so-called *Karhunen-Loève transform* in the simulation of the probability distribution that pertains to perhaps the most important novel result of this thesis (Theorem 4.2 in Section 4.2), as described in Section A.5.

We proceed with some preliminary remarks on the nature of the problems introduced, putting the topic with the most extensive material last. As for edges, there is also evidence of the psychophysical significance of ridges (see Burbeck and Pizer, 1995). Lindeberg (1998) noted that ridge features are frequently adequate descriptors in exploratory settings. The top of a ridge may be used, after projection onto the image plane, as an estimator of a curve and especially an edge. In addition, the steepness of descent away from the top of the ridge can be used as an important saliency feature that a curve estimate alone can not possibly convey. By using data that are somewhat off the edge curve, the notorious sparseness difficulties in spatial statistics are ameliorated, rendering the edge estimator more stable. This comes at a price of a higher conceptual complexity, which arise through the need for a careful definition of the notion of a ridge. Several definitions of a ridge are plausible (see e.g. Hall, Qian and Titterington, 1992), and in Section A.2 we review the one which is relevant in this thesis. Estimation of edge curves by likelihood ridges will be considered in Chapters 3 and 4.

A first illustration of what lies ahead is given in Figure 1.1. Panel (a) depicts a sample surface in form of an intensity image, obtained from the function $g : [0,1]^2 \to \mathbb{R}$ with equation (throughout this thesis, we use superscripts to denote coordinates, i.e. we write $x = (x^{(1)}, x^{(2)})$ for a generic point in $\mathbb{R}^2$):

$$g\big(x^{(1)}, x^{(2)}\big) = I\big\{x^{(2)} \leq 10(x^{(1)} - 0.1)(x^{(1)} - 0.25)(x^{(1)} - 0.6) + 0.25\big\}\,, \qquad (1.2.1)$$

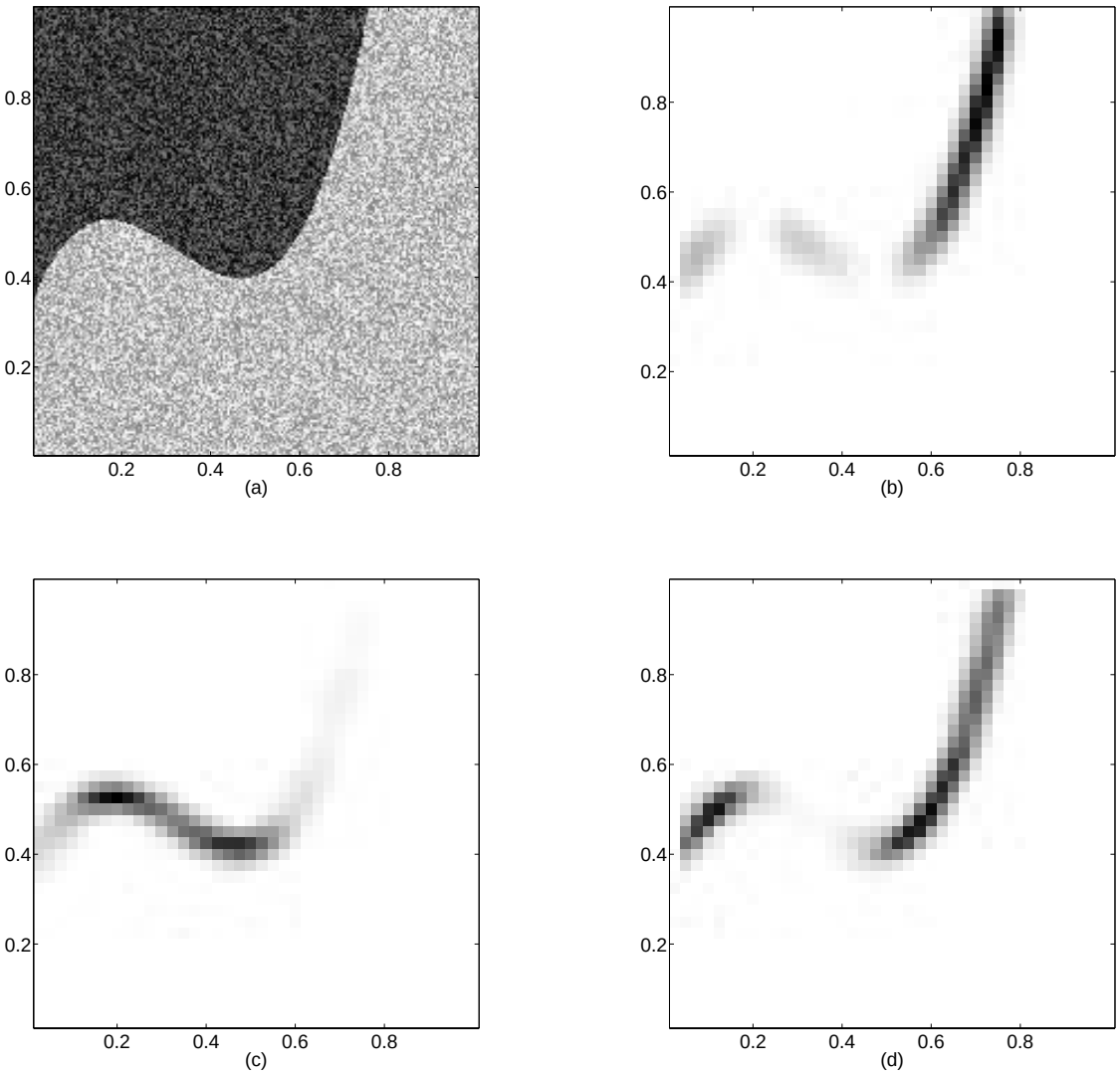

Figure 1.1: Illustration of the local likelihood method. Panel (a) shows the noise-corrupted regression surface, defined by equation (1.2.1). Panels (b)–(d) shows the derived likelihood surfaces, as described in the text.

with superimposed independent and standard Normally distributed noise. Here and in later displays, darker colour corresponds to higher surface elevation. Throughout the thesis, the letter Π will denote the subset of $\mathbb{R}^d$, typically (but not necessarily) a square, from which observations are taken and edge estimators are to be derived. Panels (b)–(d) show the *likelihood surface* $\{\ell(x, \theta_*),\ x \in \Pi\}$, with $\Pi = [0, 1]^2$ computed by discretisation to a 40×40 grid, and for the respective (and in each case fixed) unit vectors $\theta_* = (1, 0)$, $\theta_* = (0, 1)$, and $\theta_* = (1, 1)/\sqrt{2}$. The (log-)likelihood ℓ will be defined in equation (3.2.1).

From Figure 1.1 it is evident, even with the coarse pixelisation, that the direction of

θ_* (horizontal, vertical or diagonal) has significant impact on the strength, or even the mere presence, of a ridge line. The *profiled* likelihood surface $\{m(x), x \in \Pi\}$, defined at (3.2.2), takes all possible directions θ into account, but is expensive to compute over the whole sample space. This is where the local character of our estimation method becomes important from a numerical viewpoint.

The second of the three topics mentioned above consists in the *discrimination* or, synonymously, *classification* of maritime radar imagery. Generally, with the availability of high-resolution data acquisition methods, but even the low-resolution devices used to gather the data studied in this thesis, it is crucial to reduce the size of the input data to manageable proportions.

Next we explain some of the framework that will be adopted in our analysis in Chapter 6. For simplicity, assume first that each datum is not an image matrix but a vector, $\big(X(t_1), \ldots, X(t_p)\big)$ say. The points t_i with $t_1 < t_2 < \ldots < t_p$ are interpreted as the time instances (epochs) at which the continuous process $X = \big\{X(t), t \in \mathcal{T}\big\}$, with $\mathcal{T} \subset \mathbb{R}$ a time interval, is sampled. Thus the primary object of interest is the stochastic process X, and this is what we refer to as the *functional* nature of the discrete data $\big(X(t_1), \ldots, X(t_p)\big)$ that is available for inference. Usually, but not always, $\Delta_i = t_{i+1} - t_i = \Delta$, a constant, and thus the sampling frequency $1/\Delta$ is well-defined. If the datum X represents an object then ideally, one would like to choose the interval $\mathcal{T}$ or $[t_1, t_p]$ as short as possible so as to exclude background noise. The analogous notion for images is the so-called *bounding box* of the object, that is the minimum rectangle which encloses the object and is parallel to the image axes. For the dataset in Chapter 6, the choice of the bounding box was determined automatically but showed considerable variation. This problem remained, albeit to a lesser degree, if background noise was subsequently cut out, so as to let the image come closer in size to the bounding box. Aligning a set of observations in this way leads to a problem known as *curve registration* (see e.g. Gasser and Kneip, 1995, and Ramsay and Silverman, 1997, Chapter 5, pp. 67ff).

Returning to the case of time-indexed signals, the next step routinely consists in subjecting X to an orthogonal transform. Several examples of these are given in Ahmed and Rao (1975); more recently, of course, wavelet transforms have become extremely popular for this task. Transforming the data serves to decorrelate the coefficients in the expansion with respect to the basis vectors or, as explained later, basis *functions* in the analysis of Chapter 6. The decorrelation makes the representation more economical (see e.g. Gerbrands, 1981, p. 376).

Central to the analysis in Chapter 6 is the next step of *dimension reduction*, which

describes how the output vector from the aforementioned orthogonal transform, say $Y = \big(Y_1, \ldots, Y_p\big)$, is mapped into a vector $Z = \big(Z_1, \ldots, Z_m\big)$ with $m \ll p$. (We now use subscript instead of parentheses notation for the index, as the domain of the transform is not temporal or spatial.) It is usual to perform dimension reduction by simply deleting most of the components Y_i. However, care should be taken here because the features that are of potential importance in the subsequent classification step are not necessarily captured by the components Y_i in which most of the 'energy' of the observation X is encoded, which will for most transforms be those with the smallest indices i.

In this thesis, two orthogonal transforms will be used to process the input data: the discrete wavelet transform in order to reduce the influence of noise, and subsequently the *Karhunen-Loève transform* in order to obtain the vector Z used in the classification step. The Karhunen-Loève transform utilises, at an appreciably higher computational expense than other orthogonal transforms, the whole set of input observations simultaneously to obtain a data-driven representation with respect to the overall covariance matrix.

For image data which are the concern of this thesis, both of the dimension reduction and the registration problems are of course much more acute than for 'one-dimensional' signals. Especially the second problem can be expected to be hard; see the pertaining discussion in Subsection 6.2.3. The general methodologies of the linear and spatial cases, however, do not differ to a great extent. In order to apply an orthogonal transform, the image will generally be mapped into a vector, for example by column scanning (Gerbrands, 1981, p. 377). The element x_{kl} in the kth row and lth column of the matrix $X \in \mathbb{R}^{r \times s}$ then appears as the element in position $(l-1)r + k$ of the (column) vector $X_{\text{vec}} \in \mathbb{R}^{rs}$, where $rs = p$ in the earlier notation.

In the context of pattern recognition, and also the more recent field of machine learning, which share important aspects with the statistical classification methods of this thesis, the vector Z is commonly referred to as a *pattern vector* which resides in the *feature space*, which is a subset of $\mathbb{R}^m$, with m having the same meaning as earlier on. We refer to Ahmed and Rao (1975, especially pp. 225ff) for the terminology used here. Later we shall prefer to use terminology from multivariate statistics. The salient features of a pattern recognition system are displayed in Figure 1.2, which is a slightly altered version of the diagram shown in Ahmed and Rao (1975, p. 225). The formal definition of a so-called *discriminant rule* $d(\cdot)$, to which the notation in Figure 1.2 refers, is given in Section 6.1.

It is appropriate to point out here that in the analysis of radar data in Chapter 6, each datum does not have an edge in the previously defined sense, for distinction purposes

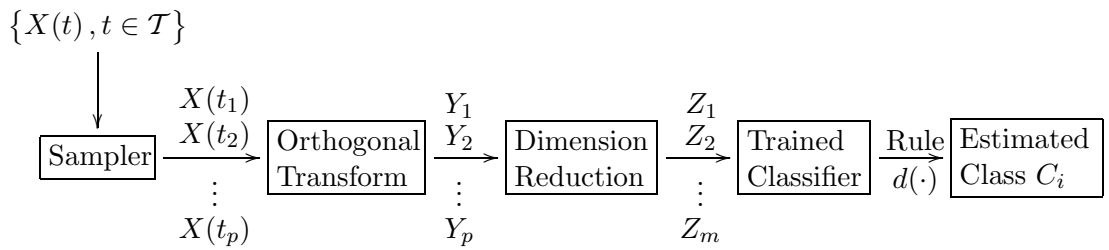

Figure 1.2: A pattern recognition (or classification) system. The output on the right-hand side is chosen from a set of pre-specified classes $\{C_1, \ldots, C_L\}$.

called a *step edge*, but rather a discontinuity in the first derivative, or a *roof edge*, along the 'footprint' which separates the object from its noisy background. (This thesis will only be concerned with the step edges, hence we will only refer to 'edges.') The distinctiveness of this roof edge generally depends on the distance of the object from the radar. These considerations recommend the use of the Karhunen-Loève transform over the alternative of trying to extract the edges, and using these as inputs for the classifier. In order to determine the bounding box of the object, however, we do use an edge detector, based on numerical derivatives. Several simpler such detectors are discussed in Section 1.3. In the analysis of higher-resolution imagery, the more sophisticated edge estimators developed in this thesis may be substituted in the appropriate place in the image processing scheme which is illustrated in Figure 6.5 (see Section 6.3).

In Chapter 6 we consider methodologies connected with *functional data analysis* in the sense of Ramsay and Silverman (1997), and apply them to the task of classification. The 'templates' for this are provided by the respective estimated mean surfaces for each class. In order to build the templates, and thus to provide a basis for classification, it is highly desirable to follow the scheme of Figure 1.2, and to use a low-dimensional vector as a surrogate for the full image data. The Karhunen-Loève expansion is appealing for theoretical reasons, due to its mean-square optimality (cf. Ahmed and Rao, 1975, p. 191, or Subsection A.5.1 of this thesis), as well as practical experience. Another advantage of the Karhunen-Loève expansion is that it "requires no detailed assumptions about the probability structure of the problem" (Kittler and Young, 1973, p. 335), which essentially reduce to the specification of prior probabilities for each class, as at (1.3.5). A pertinent case study will be the subject of Section 6.3. Our objective is to illustrate the methodology especially in those regards in which it extends preceding work of Hall, Poskitt and Presnell (2001). The classification step uses the well-known method of quadratic discriminant analysis, the definition of which is recalled in Section 6.3. Since the latter generally performed well (after dimension reduction), and because of its relative computational simplicity, it shall be our exclusive method of choice.

We now turn to the third of the above topics that will receive the most extensive treatment in this thesis (in Chapters 2–5), namely the estimation of edges. As stated earlier, it is the *statistical* view of this problem that is of main interest here, and some remarks on the significance of this stance are in order. The majority of approaches to edge detection that have been developed in computer vision and statistics have many similarities (sometimes obscured through deviating terminology), but also important differences. Two of the most profound of these lie in the very interpretation of the notion of an edge.

The sole question of what comprises an edge in a real-world image has spawned a great amount of literature in computer vision; see for example Lindeberg (1998) and the references cited therein. Humans seem to be sure enough about recognising edges, although there are famous and remarkably simple examples suggesting that perception can be confounded. Leaving these psychophysical issues aside, a pixeled image has no natural notion of smoothness attached to it. Statistical approaches, through modelling the pixel image as the result of sampling from a regression (or density) surface supported on a continuously-indexed domain, are not plagued by this problem at such a profound level: the edge is simply the set of discontinuity points within the surface. Nevertheless, the notion of *scale*, as used for example by Lindeberg (1998), will be addressed later in this introduction, as well as in Section 5.4, for two principal reasons. First, there is a well known connection between the physical notion of scale, and the statistical notion of the bandwidth of a nonparametric kernel estimator. The estimators that we propose in Chapters 2–4 are precisely of this type. Secondly, we thereby address a problem in computer vision that is closely related to the bias problem associated with such estimators: at a larger scale or bandwidth, the ability to detect the true edge, and in fact the whole noise-corrupted image, will improve up to a certain point. However, this will almost invariably be at the expense of *localisation*, that is the distance between the edge and its estimator, as measured by some criterion.

Though having received less attention, the second of the aforementioned issues, implicit in the remarks in Qiu (2002a), is of hardly less relevance. This concerns the question of whether an edge should be construed as a *curve* or a *point set*. Both of these viewpoints have their relative merits, although the second one has the longer tradition. The studies by Marr and his co-workers (1979, 1980, 1982) have already been mentioned; another paper that has remained very important is Canny (1986), who proposed a derivative-based edge detector that continues to be very frequently used in computer vision. For most of the theory presented in Chapters 2–4, we will narrow our focus to the problem of estimating a *single smooth curve*. Most of the existing literature on this topic assumes that, as in the example function at (1.1), the surface exhibiting the edge has a local

representation of the following type:

$$g(x) = g\big(x^{(1)}, x^{(2)}\big) = \begin{cases} g_0\big(x^{(1)}, x^{(2)}\big)\,, & x^{(2)} < f\big(x^{(1)}\big) \\ g_1\big(x^{(1)}, x^{(2)}\big)\,, & x^{(2)} \geq f\big(x^{(1)}\big)\,, \end{cases} \tag{1.2.2}$$

where g_0, g_1 and g are smooth functions, and $g_1 - g_0$ is bounded away from zero in a neighbourhood of the fault line, which has representation $x^{(2)} = f\big(x^{(1)}\big)$. In contrast, our own approach does not depend on a representation as at (1.2.2); rather, the fault line is allowed to meander almost arbitrarily through the plane. An illustration is given in Figure 1.3 (p. 13). Panel (a) shows the original data, which is a black-and-white image as defined in Section 1.1. Panels (b) and (c) show typical realisations corresponding to $\sigma = 0.5$ and $\sigma = 0.75$, where σ is the standard deviation of Normally distributed noise. A more comprehensive account of the model assumptions and notation will be given later in this section. The regression surface introduced here will be used as an artificial example in Chapter 5.

Another advantage of the 'curve' over the 'point set' view is that the description of the edge is conceptually more concise. By adopting a local approach, as we shall do in Chapters 2–4, a substantial part of the methodology from more conventional curve estimation problems, where a coordinate system is given *a priori*, carries over with few modifications. Considerable difficulties, however, are then encountered in finding an appropriate distance measure between the true curve and its estimate. What 'appropriate' means in this context is not at all easy to determine. The existence of a 'universally true' answer can, to say the least, be seriously doubted. Traditional L^p norms may fail to capture visually striking phenomena; see the discussion and references in Section A.8.

Regarding the edge as a point set, as is customary in the literature, circumvents the technicalities that are associated with the curve representation of an edge. Although the difficulties with distance measures essentially remain, the point-set interpretation makes a wide range of post-processing (e.g., morphological) operations available. On the other hand, in the application of such procedures to curves, care must be taken if their qualitative properties are to be preserved. For instance, if the true edge curve is simple (i.e., does not self intersect), the same should hold true for the estimator. This can often only be achieved by deletion of what are deemed to be spurious edge points. Moreover, procedures that are applied almost automatically in standard one-dimensional settings are plagued with ambiguities, especially for closed curves. As a case in point, consider the common problem of connecting a set of edge points in order to obtain a contiguous estimate. While the point set approach allows practically any rule for this,

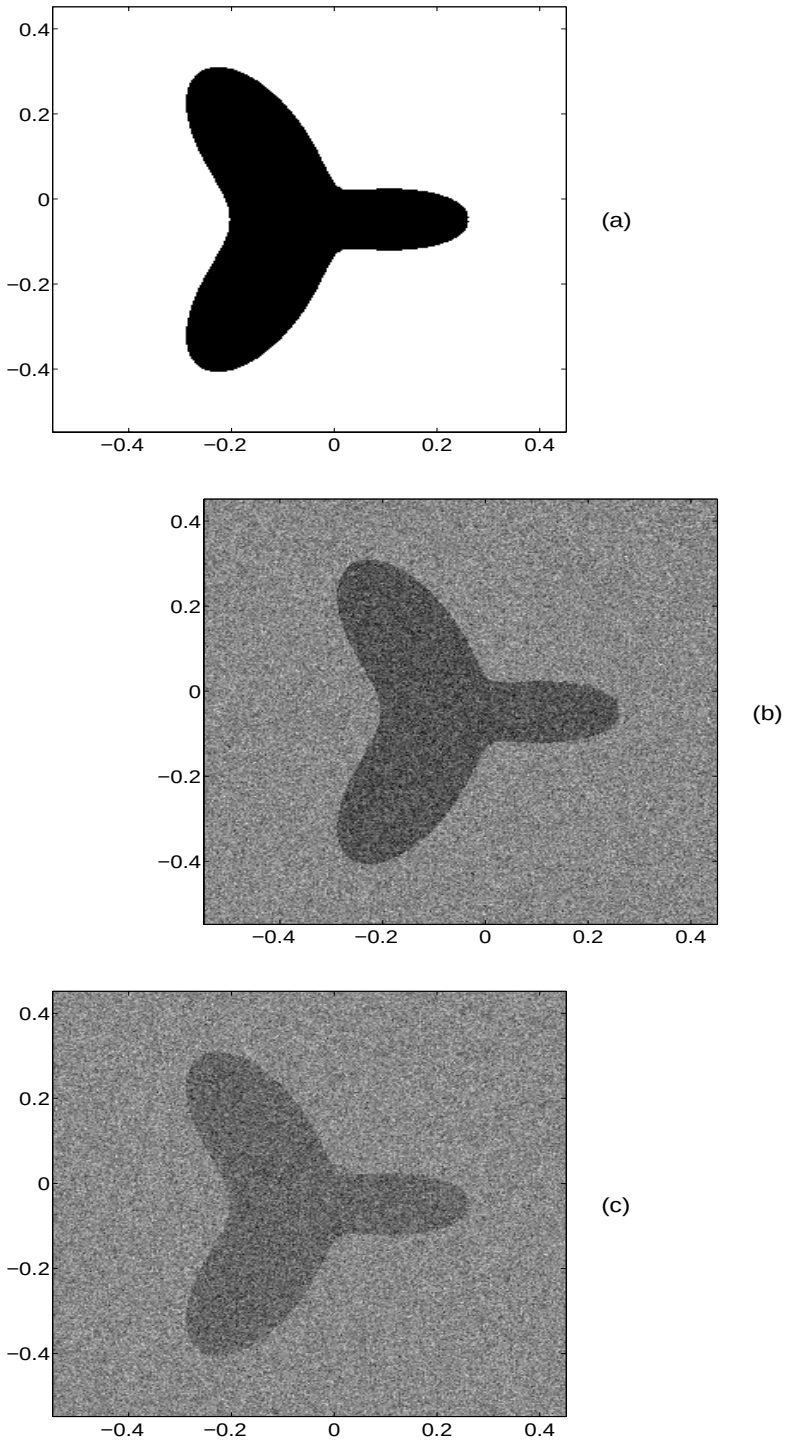

Figure 1.3: Panel (a) shows the original response surface, and panels (b) and (c) depict the superposition of Normal $\mathrm{N}(0, \sigma^2)$ noise with $\sigma = 0.5$ and $\sigma = 0.75$, respectively, in the case of lattice design. The lattice edge width is $1/300$ on the square $\Pi = [-0.55, 0.55]^2$.

the smoothness requirements imposed by the curve estimation approach suggest that a Bezier curve or smoothing spline, say of cubic order, be fitted to the data. If the fault line is closed, it is natural to impose a periodic boundary condition, but the fitted spline will depend on the choice of the starting and end-points on that curve.

The post-processing techniques available for point sets, which have been popular for a long time, include morphological operations (e.g. binary dilation/erosion, defined in Section 1.1), and *thresholding* of some sort, either in the spatial, the Fourier or a wavelet domain. By varying the threshold over a range of values, the differences between edges that are, in common (and somewhat loose) terminology, 'strong' or 'weak,' may be exhibited. This is a much harder task for curves, which — as principally infinitely thin objects — lack such a quantitative descriptor. This problem can be partly overcome by considering the *ridge* surface from which the estimators in Chapters 3 and 4 of this thesis derive, and these issues will be further discussed there.

For practical purposes, where edge detection is used at the front end of an image processing scheme, the differences between the two previously described approaches may be felt only at a later stage. The most important item of note here is the reconstruction of entire regression surfaces containing edges and corrupted by noise, or *edge-sensitive smoothing*. A benefit gained from the curve representation is that edge-sensitive smoothing is principally a straightforward task: standard smoothing procedures can be used on either side of the curve estimate, correcting for boundaries within each of the respective surface patches. Such simplicity does not exist in the point-set approach. If surface reconstruction is the main goal then it may be preferable to take a shortcut by incorporating the possibility of a jump in the model employed by the estimator itself. This is the approach taken in Qiu (2002b).

In spite of the caveats mentioned in the preceding paragraphs, problems in computer vision are a good setting in which to consider a number of fundamental aspects of edge estimation. The following basic quality criteria of an edge detector appeared in Canny (1986):

- Good detection,
- Good localisation,
- Clear response.

Further to our previous remarks, with regard to the first criterion we also recall the perpetual trade-off between errors of Type I or *false positives*, and Type II or *false*

negatives. By a preliminary surface smoothing, it is possible to simultaneously suppress both types of error to some extent, but at the cost of poorer localisation. The last point stipulates that there should be only one response per edge, meaning that a detected edge should not extend more than a single pixel in the direction vertical to it. This raises the problem that when a sloped edge passes through a gridded set of pixel points, it may appear in a 'staircase' fashion, a phenomenon known as *aliasing*. A number of existing approaches in the engineering sciences strive to take this problem into account. Unknown orientation of edges poses another important problem. This points to a natural theoretic *invariance requirement* which Mokhtarian and Mackworth (1986) phrased as follows:

> A reliable representation, suitable for matching, should be essentially invariant with the rotation, scaling, and translation of the curve to make recognition of the curve possible after arbitrary instances of those transformations. Moreover, it should represent the curve at varying levels of detail rather than at only one level.

Invariance with respect to rotation is a property that most clearly distinguishes our edge estimators from the great majority of the currently existing ones. The benefit of analysing a curve embedded in a surface at various levels of detail will be seen in Section 5.4. Our edge detectors, in their basic versions, are not applicable to situations where the fault line is not at least once continuously differentiable. To exclude more complex situations, we limit our attention to cases where singularities occur at a finite number of points only.

The topic of corner detection and our approach to it, though in itself not novel, illustrate two points that should be kept in mind with regard to the material presented in this thesis. The first is that while edges and corners are, in a sense (see Kohlmann, 1996) 'one-dimensional' and 'two-dimensional' features respectively (note that corners may be construed as the intersection of two edges), most corner detectors depend either explicitly or implicitly on the detection of edges (Mokhtarian and Suomela, 1998). Secondly, corner detection appears to provide an important instance of where there is both a possible practical and theoretical preference for using one of the longer-known edge detectors, such as the Canny method. In spite of their shortcomings, several of which have been unravelled since their invention, the 'classical' derivative-based edge detectors are not as stringent with regard to the assumptions on corners.

In images whose orientation is *a priori* unknown, edges that are not parallel to either of the image axes will be the rule rather than the exception. This suggests that beyond the goal of finding candidate points which are likely to lie on an edge, estimation of local slopes at

these points is of valuable assistance, notably in the process of obtaining contiguous fault line stretches from a set of likely boundary points. Unless an edge candidate point is actually a corner point, a locally-linear form of the edge may be justified. Such short line segments, which in their entirety define an edge estimator, were called *edgels* in Nalwa and Binford (1986).

In this thesis, the estimation of fault lines in both regression and density surfaces will be studied. We start with a rough outline of the nature of each of these problems.

For regression surfaces, the explanatory data may be given as sites on a rectangular grid. In computerised images, each design point naturally corresponds to a pixel of the image. However, this regularity assumption is less appropriate in various other important applications. A case in point is the modelling of geostatistical data. For example, the locations of sites where drill cores are taken on the ocean floor are often distributed in a highly irregular fashion. Then it may often be plausible to assume that the design represents a realisation of a (usually homogeneous) planar Poisson process. The favourable role of the Poisson process in modelling design stems from its simplicity, and its compliance with the invariance requirement from earlier.

The problem of estimating fault lines in density surfaces shares several aspects with the regression problem. That the literature on the latter topic dates back longer can be ascribed to the close links that exist between estimating a density surface as a whole, and estimating only a fault line within it. Perhaps the most important estimation technique that has been developed is the excess mass approach; see for example Hartigan (1987) and Polonik (1995). We shall not discuss these approaches further as their methodology is profoundly different from ours, and merely note that in using the excess mass approach, a non-trivial problem is how to determine the right threshold which yields the fault line estimate. It appears only superficially that problems involving *support boundary estimation* can be treated as a special case, since the unavailability of data on one side of the boundary introduces profoundly different aspects. In this thesis, support boundary estimation will not be considered.

In this and the subsequent paragraphs, we make precise the notion of consistency of the statistical estimation methods to be developed. Depending on whether a regression or a density surface is being considered, the sample data are denoted by either $\{(X_i^{(1)}, X_i^{(2)}, Y_i) = (X_i, Y_i)\} \subset \mathbb{R}^3$ or $\{X_i\} \subset \mathbb{R}^2$, where the indices i are taken from some at most countable index set I. The observation area may naturally be taken as a square if $\mathcal{X} \equiv \{X_i\}$, the point process representing the sites of the observations, coincides with a grid, but in principle any compact set can function in this role. This set will always be

denoted by Π, and we shall assume throughout that $0 < \|\Pi\| < \infty$. We further stipulate that our point processes are locally finite, and hence $\mathrm{card}(I) < \infty$ a.s. (This assumption is automatically satisfied for Poisson processes with intensity functions that are bounded away from infinity.) In the regression case, the underlying model will always take the following form:

$$Y_i = g\big(X_i^{(1)}, X_i^{(2)}\big) + \epsilon_i\,, \tag{1.2.3}$$

where $g(x) = E(Y \mid X = x)$ is assumed to be smooth except for the fault line $\mathcal{C}$ running through it, and conditional on the point process $\mathcal{X}$, the errors ϵ_i are independent with mean zero and finite variance $\sigma^2 > 0$. It is also assumed that all moments of ϵ_i exist (even an exponential moment in Chapters 2 and 3). There seems to be a broadly-shared view that even the assumption of uniform boundedness of the errors ϵ_i, that is when there exists a constant $M > 0$ such that $|\epsilon_i| \leq M$, is "quite natural in the context of image analysis" (Gayraud and Tsybakov, 2002, p. 71). This is indeed true for so-called *salt-and-pepper noise*, which occurs when a certain rough proportion of pixels is set to the minimum or maximum value of attainable pixel values. In the case of eight-bit-coding, for example, these values range from 0 to 255.

The assumption that σ is known will be maintained throughout Chapters 2–4 of this thesis, with a single exception made in Remark 4.7. However, it is pertinent to notice that σ "can be known or estimated with sufficient precision, by using the values [...] in a 'background' region of the image, when it is present" (Godtliebsen and Sebastiani, 1994, p. 460). A similar remark applies to the noise statistics that are used for the radar imagery in Chapter 6, where correlations in the signal as well as in the 'noise' (defined there in somewhat loose and not purely probabilistic terms) play an important role.

Unless otherwise stated, we ask that $\mathcal{C}$ be a contiguous and rectifiable curve. Contiguity means that the set of points on $\mathcal{C}$ can be represented in a parametric form, say

$$\mathcal{C} = \big\{x(t) = \big(x^{(1)}(t), x^{(2)}(t)\big) : 0 \leq t \leq T\big\}\,, \quad T \in \mathbb{R}_+^*,$$

where the $x^{(i)}(t)$ are continuous functions for $i = 1, 2$. This assumption mainly serves to ensure that estimators defined by a 'tracking' algorithm (to be introduced in Chapter 2) trace out $\mathcal{C}$ in its entirety, and therefore the extension to multiple fault lines, as long as their number is known in advance, is at least in principle (albeit not necessarily in practice) straightforward. Rectifiability means that for every $0 = t_0 < t_1 < \ldots < t_k = T$, the elementarily defined length of the inscribed polygon with vertices $x(t_i)$ is uniformly bounded, the (finite) least upper bound then being equal to the length of $\mathcal{C}$. We require the existence of two bounded derivatives of a parametrisation of $\mathcal{C}$ as at (1.2), which in

particular implies that $\mathcal{C}$ is rectifiable. A single exception from this assumption is made in Remark 2.9. It is also required that $\mathcal{C}$ not be self-intersecting. However, the case that $\mathcal{C}$ is closed is always permitted; the different assumption made in the theoretical analysis of Section 2.3 (see p. 33) is only made for ease of exposition.

Before we describe the sequence of models at (1.2.3) for both the regression and density cases, we point out that (1.2.3) can naturally be transferred to higher dimensions. Red-Green-Blue, or RGB images consist of three superimposed images corresponding to the primary colours, which may substantially differ, and not reveal the edges in each of the components. In the analysis of multi-band data in signal processing, the number of components, say d, may be even larger than three. Model (1.2.3) is then modified to

$$Y_i^{(k)} = g_k\big(X_i^{(1)}, X_i^{(2)}\big) + \epsilon_i^{(k)}\,, \quad k = 1, \ldots, d\,, \tag{1.2.4}$$

where g_k represents the true regression surface in the k th coordinate, in applied contexts often called *band*. In many applications it may be perfectly acceptable to analyse the data band by band, directly using the proposals of this thesis. Alternatively, if correlations in the errors are deemed to be of enough importance, one could use a k-variate Gaussian model for the vectors $\epsilon_i = (\epsilon_i^{(1)}, \ldots, \epsilon_i^{(k)})$ in (1.2.4). This leads to a weighting of key formulae such as (2.2.2) with the inverse of the error covariance matrix. The sparseness difficulties that are already associated with the case $d = 1$, evidenced by the numerical studies presented in Chapter 5, are then of course aggravated. Moreover, a graphical analysis is also rendered much more difficult. In this thesis, we shall limit ourselves to the univariate case of equation (1.2.3).

As $\mathrm{card}(I) < \infty$ with probability 1, the design is not dense, and thus exact inference on the location of the fault or boundary line is principally impossible. The customary statistical approach of this problem is to derive *asymptotic* inference results as $\mathrm{card}(I) \to \infty$ (Cressie, 1993, p. 350).

As pointed out by Cressie (1993), there are two basic ways in which the sample size $\mathrm{card}(I)$ could tend to infinity:

(a) *Ergodic Limit*: Expansion of the explanatory area Π such that $\|\Pi\| \to \infty$, thus allowing more observations to be taken.

(b) *Infill Asymptotic Limit*: Addition of observations at locations between the existing ones.

Both the regression and the density fault line problems addressed in this thesis are naturally considered in the context of the infill asymptotic.

The following assumptions will be made on the point process $\mathcal{X} = \mathcal{X}_n$ of design points.

1. In the regression surface case, the set $I = I_n$ which enumerates the design points is indexed by the parameter $n \to \infty$, which equals the intensity of the homogeneous Poisson process on Π. If the design is assumed to coincide with a non-random shift of a regular grid, the parameter n will be defined such that the edge width of the mesh is $n^{-1/2}$. Then the design process may be written as $\mathcal{X} = \mathcal{G} \cap \Pi$ where

$$\mathcal{G} = x_0 \oplus \left\{ \left(\frac{j}{\sqrt{n}}, \frac{k}{\sqrt{n}} \right), j, k = \ldots, -1, 0, 1, \ldots \right\}, \tag{1.2.5}$$

   with x_0 a fixed point. Note that in this case, the total sample size is deterministic and asymptotically equal to $n\|\Pi\|$.

2. In the density surface case, the index n in the notation I_n denotes total sample size. It is again assumed that $n \to \infty$. For intensity surfaces, however, we relax this assumption so that we only ask for the existence of a positive sequence $\{c_n\}$ with $c_n \to \infty$, and that the (random) *effective sample size*, $N_n = \text{card}\{\mathcal{X}_n \cap \Pi\}$ satisfies $N_n/c_n \to_P Z$, where $Z > 0$ almost surely.

The Poisson cluster process from the last paragraph of Section 1.1 is an example of an intensity surface that can be treated by our methods. Note that the formulation in the regression surface case does not require that $n \in \mathbb{N}$, although that would typically be the case in image analysis applications (there, in fact, $n^{1/2} \in \mathbb{N}$). If $\mathcal{X}_n$ represents a Poisson process, the increase of the parameter n could represent a sequence of observation stages, say $n = n_k \to \infty$ as $k \to \infty$. At the k th stage, the observations $(X_{k,i}, Y_{k,i})$ are incorporated, where the $X_{k,i}$s come from a Poisson process of intensity $n_{k+1} - n_k > 0$ which is independent of the previous stages. The combined data are then used for inference.

This thesis is organised as follows. Chapter 2 introduces a tracking estimator for a fault line in a regression surface, which works directly on the raw data. In Chapter 3, we propose a class of likelihood-based estimators for boundary curves. These estimators are shown to be capable of dealing with curve estimation problems in the other contexts that have been mentioned, and also of allowing for the construction of pointwise and simultaneous confidence bands. Most of Chapter 4 is devoted to the latter topic. Issues concerning the numerical performance of the edge estimators are studied in Chapter 5. In Chapter 6, we change the focus to demonstrate the use of the Karhunen-Loève expansion in the discrimination of a dataset of radar imagery. An exposition of the theory pertaining to our own contributions from Chapter 2-4 is given in the Appendix.

1.3 Literature Review

In the introduction we have already pointed to the close connections between the detection of edges in pixeled images, and the estimation of fault lines in regression surfaces (for which we retained the term 'edges'). The need to appreciate the distinctions that arise in the course of different interpretations of an edge was also noted. For this reason alone, it is not possible to make a fair comparison of various methods for edge estimation. In view of the plethora of such methods, our exposition below can hardly give more than a flavour of a few salient approaches from the areas which relate most directly to this thesis. We also want to convey the basic philosophy that underlies various edge detectors or estimators, and hence some elements of image processing will also be recapitulated. This will be expedient in later chapters, especially Chapter 5.

Direct methods for edge detection generally follow the same basic scheme. First, some derived image called the *edge map* (see Jain, 1989) is calculated. The pixel values in this image correspond to what is often called the *jump detection criterion* (JDC) for the corresponding pixel in the original image: the larger the absolute value of the JDC, the stronger the evidence for the edge passing through that point (as before, this is to be interpreted modulo the discretisation that comes from finite sampling). In the second step, a thresholding procedure is applied to the edge map pixels in order to decide which of these belong to an edge. However, in a post-processing stage of the edge extraction process, edges that are identified as insignificant may be deleted later. In advance of more detailed comments shortly, we point out that in many practical situations, this threshold may be selected by eye, or from previous experience with similar data. This can be an entirely satisfactory solution, but in the vast majority of cases it underlines the problem that the theory that supports most existing edge estimators is limited. On the other hand, theoretical justifications can "help understand ... [the] strengths and limitations [of edge detectors] so that they can be further improved" (Qiu, 2002b). Consequently, a well-developed theory also helps to alleviate the problem of choosing an appropriate edge detector in a new situation.

The archetype of edge detectors is based on the convolution of an image with a filter of fixed size, which in this context is typically a 3×3 matrix called a *mask*. Considering the noise-free case for the moment, the numerical gradient can be expected to behave in a stable fashion in the smooth regions of the image surface, that is across so-called *texture*. An example of a commonly-used filter for enhancement of whole images (not just edges) is the median filter, and is especially suited to remove salt-and-pepper noise (see the paragraph below (1.2.3)). For a definition of the median filter, see Jain (1989).

In the neighbourhood of an edge, at least one of the first numerical derivatives should reflect the singularity by being large in magnitude.

Methodologically simple edge detectors are based on an approximation of the gradient, of which the Sobel detector is a prominent example. The jump detection procedure proposed by Qiu (2002a) may be regarded as a substantial generalisation of the Sobel edge detector, in that window sizes larger than 3×3 are allowed, making the discretised smoothing operator more sophisticated. At the same time, his procedure does not employ the costly optimisation over the edge direction evinced by (1.3.1) below, and neither do the procedures proposed in this thesis.

Employing second-order derivatives leads to more sophisticated but also more noise-prone edge estimators. The (generalised) Laplacian edge detector (Jain, 1989), which is sometimes also called the Marr-Hildreth edge detector, is the prominent example in this category. An overview of mask-based edge detectors, along with some further discussion, can be found in Jain (1989). They are implemented in various software packages, such as the Image Processing Toolbox of MATLAB©. Even in the presence of mild noise, they may perform poorly at edges which are not approximately horizontal or vertical within the image. Such biases only become more pronounced when increasing the size of the operator mask.

Within the group of derivative-based edge detectors, the more sophisticated ones are those whose origins date back to the Canny algorithm (Canny, 1986). Its more solid theoretical foundations, as compared to the simpler derivative-based edge detectors mentioned before, and overall reliable practical performance have secured Canny's method a place as a building block in many imaging problems. One of these problems is the detection of corners, which will be discussed in Section 5.4.

An inherent weakness of the mask-based edge detectors, as indicated two paragraphs earlier, is that they are mostly geared toward detecting edges in the horizontal and vertical directions. In order to obtain an edge detector that can deal with sloped edges, the computation of the JDC should involve a maximisation over possible (local) angles. In other words, the estimation of fault line points and the estimation of the *tangent direction* should be effected simultaneously, even though that tangent estimate could be considered a nuisance parameter. In fact, tangents are highly useful in order to link edges, identify outliers, and most importantly, to devise efficient algorithms. Nalwa and Binford (1986) is one of the earliest papers to consider such a procedure. By least-squares fitting within each local neighbourhood, they obtained a pilot estimate for the edge tangent. This estimate is then refined through locally fitting a cubic spline. Within

a small region, the judgement for an edge was based on the fit of a univariate surface defined by the hyperbolic tangent function as the local model. They defined a univariate surface to be a surface that was constant along each line parallel to the pilot estimate of the normal direction. An analysis was conducted on coarsely sampled images (128×128 and 256×256) to support the claim that the estimator is capable of good performance. The attempt of these authors to obtain contiguous edge estimates is also an early example of the 'curve view' of edges. Among the drawbacks of their procedure was the number of successive fits required to obtain a satisfactory result; secondly, the plausible but arguable assertion that the hyperbolic tangent function provides a viable local model for edges encountered in practical applications; and thirdly, the tendency of the method to break up edges in places of high curvature.

As Nalwa and Binford (1986) pointed out (p. 707), their statistical analysis of localisation error relies on idealised assumptions on the edge, notably that the noise is Gaussian. At the time of their publication, the development of practical efficiency measures on a theoretical foundation, and for a more inclusive class of images, had not yet penetrated the statistical literature. Korostelev (1991) and Korostelev and Tsybakov (1993, 1994) studied the *minimax* criterion which, although predominantly qualitative rather than quantitative, has proved very useful to assess the performance of edge detectors. Based on minimax theory, several authors have derived convergence rates for diverse spatial problems such as fault line estimation, and also for analogous problems for subsets of the real line $\mathbb{R}$, which are referred to as *change-point estimation*. The approaches that are developed in Chapters 2–4 of this thesis are practical and 'nearly' achieve the known minimax-optimal rates. As described there, a change-point detection might be used at the outset of the procedure. A wide range of methods for this purpose, including some papers on edge detection, is discussed in papers in the conference proceedings edited by Carlstein, Müller and Siegmund (1994); see especially Eubank and Speckman (pp. 130–144 *loc.cit.*). Work of Khmaladze, Mnatsakanov and Toronjadze (2002) should also be mentioned in this context. These authors developed a unifying approach to obtain minimax-optimal rates for change-point problems, as well as black-or-white image problems posed in $\mathbb{R}^d$, for general $d \in \mathbb{N}$.

With the advent of higher computing power, edge detection algorithms could be considered which include the angle fitting as a parameter in the true surface model, rather than through some pilot procedure. Qiu (1997) defined the *rotational difference kernel estimator* (RDKE). Its definition resembles that of our own estimators in important respects, see for example formulae (2.2.2) and (2.4.2) in Chapter 2. The corresponding

jump detection criterion is given by

$$\mathrm{JDC}_{\mathrm{RDKE}}(\theta, x^{(1)}, x^{(2)}) = \frac{1}{n\, h_n\, p_n} \sum_{i=1}^{n} Y_i \times \left[K_2\left(\theta, \frac{X_i^{(1)} - x^{(1)}}{h_n}, \frac{X_i^{(2)} - x^{(2)}}{p_n}\right) - K_1\left(\theta, \frac{X_i^{(1)} - x^{(1)}}{h_n}, \frac{X_i^{(2)} - x^{(2)}}{p_n}\right)\right], \quad (1.3.1)$$

where $K_i(\theta, x)$ is the version of the kernel $K_i(x)$ rotated by angle θ (for $i = 1, 2$), the support sets of K_1 and K_2 equal $[-1/2, 1/2] \times [0, 1]$ and $[-1/2, 1/2] \times [-1, 0]$ respectively, and h_n, p_n are bandwidths. Qiu (1997) assumed that the design points $\{X_i\}$ were gridded, but the estimator could evidently be applied to Poisson-distributed data as well. Special features of the case where the design variables are on a regular lattice have been addressed by Hall and Raimondo (1997a, 1997b, 1998).

The distinction observed earlier between the goals of edge detection and surface estimation also draws a line between the parts of the literature that are devoted to either of these two topics. However, edge estimation was made explicit in some important papers on surface estimation, including the multistage methods of Müller and Song (1994), Qiu and Yandell (1997), Qiu (1998) and Wang (1998). By taking the edge curve explicitly into account, the performance of a kernel smoothed estimate of the regression surface may be improved. Müller (1992) carried out this programme in the context of a one-dimensional change-point problem.

The performance of boundary curve estimation procedures may plausibly be measured by means of (pointwise or simultaneous) confidence envelopes or *bands*. These bands are a random set whose shape is solely determined by the available data, and which covers the true frontier or boundary curve with a given high probability. The bootstrap has proved to be a viable tool for constructing confidence bands; see for example Hall and Owen (1993). For the problem of support boundary estimation, Hall, Park and Stern (1998) and Gijbels, Mammen, Park and Simar (1999) have used arguments with a stronger geometrical flavour, similar to those used in this thesis. In more recent years, due to the interest from the application area of econometrics, numerous other papers have dealt with estimation of support boundaries of point-process intensities or probability densities (e.g. Deprins, Simar and Tulkens, 1984, Carlstein and Krishnamoorthy, 1992, appearing in Carlstein, Müller and Siegmund, 1994, Seiford, 1996, Park, Sickles and Simar, 1998, Kneip, Park and Simar, 1998 and Simar and Wilson, 1998).

However, in the context of fault line estimation and density estimation, which is the main focus of this thesis, there seems to be a scarcity of estimators which are both

widely applicable, and for which limit results beyond mere consistency have been derived. The approaches that have emerged can be grouped into being either local or global in nature. O'Sullivan and Qian (1994) considered the problem of estimating a simple closed curve. As pointed out in the Appendix (see Section A.8), this has the advantage that the estimation problem can be rephrased as pertaining to a connected closed set. The authors define a contrast statistic, and derive the edge estimator through a non-linear optimisation over a function space of boundaries, regularised by using penalty constraints. Our definition of the target function (i.e., the (log-)likelihood; see equation (3.2.1) in Section 3.2) employs a local version of a contrast statistic. The approach was shown to perform well for the set of images considered in that paper, but as the authors note, the need for judicious choice of a starting point for the optimisation requires "some heuristics" (O'Sullivan and Qian, 1994, p. 565).

Müller and Song (1994) also investigated fault line estimation using set-indexed estimators, and showed that a scaled stochastic process induced by the partitions of the image they consider, converges to a set-indexed Brownian motion with linear drift. A more explicit locally-parametric view was adopted in the preceding study by Rudemo and Stryhn (1994). These authors used techniques that are somewhat similar to ours to derive limit distributions for likelihood-band estimators of regression fault lines. Not surprisingly, these limit distributions are likewise related to the Gaussian measures that will be derived here. To illuminate some of the ideas and problems at hand, we now present a brief outline of their model and results.

Rudemo and Stryhn (1994) considered fault line estimation in two related regression models, namely what they called the *upper-lower region model with column estimators* and the *starshaped region model with sector estimators*. Both of these models are amenable to the application of results from change-point analysis, by virtue of an appropriate parametrisation of the edge, and hence we only recapitulate the upper-lower region model. There it is assumed that $\Pi = I_2 = [0,1] \times [0,1]$, and that the set of design points is of pixel form, which is conveniently double-indexed here as

$$S = \{X_{jk}\} = \left\{ \left(\frac{j - \frac{1}{2}}{m}, \frac{k - \frac{1}{2}}{n} \right), \, j = 1, \ldots, m \, , \, k = 1, \ldots, n \right\} ,$$

where $m, n \in \mathbb{N}$. The square I_2 is divided into two connected regions $\mathcal{R}_+$ and $\mathcal{R}_-$ by a continuous path $\mathcal{C} \subset I_2$ without double points. The observations Y_{jk} on both sides of $\mathcal{C}$ (combined) are independent, and distributed with probability densities φ_+ and φ_- depending on whether the corresponding grid site X_{jk} falls below or above $\mathcal{C}$ (for

definiteness, let $\mathcal{C} \subset \mathcal{R}_+$). In a similar fashion to (1.2.2), the true curve is represented by

$$\mathcal{C}_g = \text{graph}(g) = \{(x^{(1)}, x^{(2)}) \in I_2 : \, x^{(2)} = g(x^{(1)}),\, 0 \le x^{(1)} \le 1\}\,, \tag{1.3.2}$$

where $g : [0,1] \to (0,1)$ has at most finitely many discontinuity points, at which the left- and right-hand limits are assumed to exist. There is a one-to-one correspondence between $\mathcal{C}$ and $\mathcal{C}_g$ which is established by joining the limits at the discontinuity points of g by straight vertical lines. If the unit interval $[0,1]$ is divided into $\kappa \le m$ bins of length $1/\kappa$, then for a parameter vector $\theta = (\theta_1, \ldots, \theta_\kappa) \in \mathbb{R}^\kappa$, the piecewise constant *regressogram estimator* $\hat{g}$ is defined by

$$\hat{g}(x|\theta) = \sum_{k=1}^{\kappa} \theta_k \, I\left\{\frac{k-1}{\kappa} \le x < \frac{k}{\kappa}\right\}, \quad 0 \le x \le 1\,. \tag{1.3.3}$$

The parameter vector $\theta = \big(\theta_1, \ldots, \theta_\kappa\big)$ is chosen as a maximiser of the log-likelihood $\ell(\cdot)$, which is

$$\ell(\theta) = \sum_{X_{jk} \in \mathcal{R}_+(\theta) \cap S} \log \varphi_+(Y_{jk}) + \sum_{X_{jk} \in \mathcal{R}_-(\theta) \cap S} \log \varphi_-(Y_{jk})\,, \tag{1.3.4}$$

where $\mathcal{R}_\pm(\theta)$ corresponds to $\mathcal{R}_\pm$ in the same way as described above for $\mathcal{C}$ and $\mathcal{C}_f$. The estimator $\hat{g}$ is called the *column estimator*; notice that $\hat{g} \in D[0,1]$, the Skorokhod space of right-continuous functions on $[0,1]$ with left-hand limits. If $n, m, \kappa \to \infty$ in such a manner that $\kappa \exp(-\epsilon n m/\kappa) \to 0$ for any $\epsilon > 0$, then the estimator $\hat{g}$ is consistent for g in the Skohorod metric, or the supremum metric if $g \in C[0,1]$, the space of continuous functions on $[0,1]$ (Stryhn, 1993, cited in Rudemo and Stryhn, 1994).

The main result of Rudemo and Stryhn (1994) on the boundary estimator in the upper-lower region model is as follows. Assume that $\varphi_+(\cdot) = p(\cdot\,; \psi_0)$ and $\varphi_-(\cdot) = p(\cdot\,; \psi_0 + \Delta)$, where $\{p(\cdot\,;\psi),\, \psi \in \Psi\}$ is a family of probability densities indexed by the parameter vectors taken from $\Psi \subset \mathbb{R}^d$, for some $d \in \mathbb{N}$. Assume further that $I_0 = I(\psi_0)$, the Fisher information matrix of the model evaluated at ψ_0, is positive definite. Let $m, n, \kappa \to \infty$ and $\Delta'\Delta \to 0$ in such a manner that

$$\begin{aligned} n/m \to c \in (0,\infty)\,, \qquad & m/\kappa \equiv a \text{ an integer, or } m/\kappa \to \infty\,, \\ (mn/\kappa)\Delta' I_0 \Delta \to \infty\,, \qquad & (m/\kappa)\Delta' I_0 \Delta \to 0\,, \\ (mn/\kappa^2)\Delta' I_0 \Delta \to \beta \in [0,\infty)\,. & \end{aligned}$$

(The condition $\Delta'\Delta \to 0$, that is the contraction of the jump height, is reviewed in the context of our own methods in Remark 2.6.) Let the column-wise likelihood $\ell(\theta_k)$,

$k = 1, \ldots, \kappa$, be defined by summing in (1.3.4) only over those X_{jk}s which, in addition to the conditions under the summation signs there, are such that their first coordinate lies in the set appearing in the indicator of (1.3.3). Under moment assumptions on the log-likelihood ratio of a change-point problem which serves as an approximant of $\ell(\theta_k)$, Rudemo and Stryhn (1994) proved that if g is continuously differentiable in an η-neighbourhood of a given point $x \in (0, 1)$, for a fixed $\eta > 0$, then

$$(mn/\kappa)\Delta' I_0 \Delta\{\hat{g}(x) - \gamma(x_k)\} \Longrightarrow F_{\text{patr}}\,,$$

where $x_k = (\langle \kappa x \rangle + 1/2)/\kappa$, and F_{patr} is the distribution of the location of the maximum of a two-sided Brownian motion with parabolic-triangular drift. (Thus, the subscript 'patr' in F_{patr} refers to the nature of the drift.) We shall not display the distribution F_{patr} here. Distributions defined through Gaussian processes of a similar (albeit more complicated) nature will be studied, in preparation for our own results, in Subsection A.4.2. Rudemo and Stryhn (1994) do point out (Remark 4 *loc cit.*) that their regressogram binning may be replaced by kernel smoothing, in which case the above asymptotic result simplifies. In view of the remarks on aliasing from the introduction, the use of genuine smoothing methods (rather than binning) seems more efficient.

Although the column estimator (and likewise the estimator for the case of boundaries of star-shaped objects) is fairly easy to construct, and has the additional appeal of directly relating to one-dimensional change-point problems, it suffers from further drawbacks, of which the most important are:

- The approach requires a gridded design, like the ones mentioned previously. However, it seems that the theory may be extended to Poisson design, at the expense of a longer proof. Change-point problems with an irregular (random) design have not received the attention of problems with a regular design. Issues involving the loss of a natural system of axes would also have to be addressed.
- If the fault line doubles back towards itself, so that a representation with $\mathcal{C}_g$ as at (1.3.2) is not available by any rotation of the graph, it is necessary to treat the estimation problem in parts. However, due to limited design density, it might be infeasible to extract the fragments successfully.

In principle, the results of Rudemo and Stryhn (1994) enable the construction of *asymptotic confidence bands* for a fault line, a topic treated extensively in Chapter 4 of this thesis. The interest in the application of confidence bands to more general image models has motivated the research for the results presented in this thesis.

There are also intimate links of the topics of Chapter 4 of this thesis to the problem studied by Gayraud and Tsybakov (2002) who considered the test of the hypothesis that the true fault lines belongs to a parametric class, with a non-parametric alternative. Due to the equivalence of this problem with the construction of confidence sets, a link to the work presented here may be drawn, in the special case of a gridded design, as pointed out in Remark 4.4. In Gayraud and Tsybakov (2002), the null family of distributions was required to comprise only edges with a form of 'near' Lipschitz smoothness, or alternatively that their *Vapnik-Červonenkis dimension* (see Van der Vaart and Wellner, 1996, pp. 85–86 for a definition) be bounded. This seems a rather high price to pay for the greater generality of their other conditions, as compared to those in this thesis, which are formulated in terms of smoothness properties. When used to estimate fault lines with bounded curvature, the methods of this thesis achieve optimal convergence rates to within logarithmic factors. This is even so in settings that are far more general than the locally-linear models (see e.g. formulae (2.2.1) and (2.2.2)) allow. The principle of combining the flexibility of nonparametric, that is qualitative, specification of the estimator with the methodological and computational advantages of locally parametric models has appeared in numerous papers before, for example the approaches to locally parametric density estimation discussed by Hjort (1994), Copas (1995), Hjort and Jones (1996), Loader (1996) and Simonoff (1996, pp. 64ff). Generic properties of smoothing methods such as those in this thesis were studied by Titterington (1985a, 1985b).

As noted earlier, the use of the Karhunen-Loève transform in the context of dimension reduction for pattern recognition and statistical classification stems from its optimality properties with regard to mean-square error, and the mildness of the assumptions on the probability structure of the problem as a whole. Like the Fourier and wavelet transforms, the Karhunen-Loève transform has both a discrete and continuous version. For the definitions of the two versions, we refer to Mendel and Fu (1970). The continuous version is also discussed in Subsections A.5.1 and A.5.3 of this thesis. The connection between the two forms of the Karhunen-Loève expansion can be drawn fairly straightforwardly. Methods for treating multivariate data that can be treated as samples from an intrinsically continuously-indexed stochastic process are nowadays often put under the heading of *functional data analysis* (FDA) in the sense of Ramsay and Silverman (1997). That monograph is an excellent source for methodologies of FDA, and contains many references to earlier work in the same direction, such as Rice and Silverman (1991) and Gasser and Kneip (1995). Further case studies for functional data analysis may be found in Ramsay and Silverman (2002).

We now turn to aspects of the discrimination problem for dimension-reduced signal pro-

cessing data that we shall tackle in Chapter 6. A precise formulation of the problem is given in Definition 6.1 of Section 6.1. For introductory accounts on the topic see, for example, Ahmed and Rao (1975), Lachenbruch (1975), and Mardia, Kent and Bibby (1979). In essence, our methodology does not differ much from the parametric variant of an approach developed in Hall, Poskitt and Presnell (2001), from which most of the material in the remaining paragraphs is taken. The dataset in the analysis of these authors differs from ours in its linear (time-series) type.

From the point of view of decision theory, the most appealing classification method operates by computing the likelihood of a new observation for each of L (seven in our case) candidate types, and then assigns the new observation to the signal type with the highest *posterior* probability. Strictly speaking this requires specification of prior probabilities, $\omega_l \geq 0$, $l = 1, \ldots, L$, satisfying $\sum_l \omega_l = 1$ for the L different types. Given density functions $f_{(1)}, \ldots, f_{(L)}$ for the L different classes, the posterior probability that an incoming signal x is of class l is then

$$p(l|x) = \frac{\omega_l f_{(l)}(x)}{\sum_{j=1}^{L} \omega_j f_{(j)}(x)} \,. \tag{1.3.5}$$

In the analysis of Section 6.3 we shall take the prior probabilities to be uniform, i.e., $\omega_l = 1/L$, $l = 1, \ldots, L$, so that a new observation is assigned to the signal type l for which the likelihood, $f_{(l)}(x)$ is largest.

In practice we must use training data to estimate the class densities for a finite dimensional representation of x, that is the pattern vector $x^{(m)} \in \mathbb{R}^m$ in the discussion around Figure 1.2. (The training and the test data in classification experiments are those whose true class is known or unknown, respectively.) Thus the *true* class densities above, $f_{(l)}(x)$, $l = 1, \ldots, L$, are replaced by estimates $\hat{f}_{(l),m}(x^{(m)})$. One approach, which was employed in Hall, Poskitt and Presnell (2001), is to use kernel methods to estimate the densities of the dimension-reduced training data. Given training data from different signals, they computed the version $\hat{f}_{(l),m}$ of $\hat{f}_m$ for data from the lth type and classified a new datum x as coming from signal type r if $\hat{f}_{(r),m}(x^{(m)}) > \hat{f}_{(l),m}(x^{(m)})$ for all $l \neq r$. This is a fully nonparametric approach that has several advantages, most importantly that it does not require specification of a particular parametric form for the distribution. However, as noted in Hall, Poskitt and Presnell (2001), the storage and especially the computational burden of the kernel approach is a serious drawback, which in fact makes it not nearly fast enough to perform on-line classification of the dataset considered in this thesis, with the present state of technology.

To ameliorate these problems, in the analysis of Section 6.3 we fit an m-variate Gaussian

density $\mathrm{N}(\mu, \Sigma)$ to the features derived from the respective training samples. This results in a procedure that is based on an application of the relatively conventional technique of quadratic discriminant analysis (QDA), after functional data-analytic methods have been used to reduce the dimension of the problem. Arguments for favouring QDA over linear discriminant analysis in the context of the analysis of this thesis, where covariances certainly differ between classes and a substantial deviation from normality is to be expected, are given in Lachenbruch (1975). We note that the method of Hall, Poskitt and Presnell (2001) of employing the Karhunen-Loève expansion, by the subtraction of the overall (inter-class) mean of the data, dates back a long way, at least to Watanabe's "Selfic" method (see Kittler and Young, 1973).

Similarly to the conclusions in Hall, Poskitt and Presnell (2001), functional data-analytic methods provide a basis $\{\psi_r\}$, in the notation of (A.5.2) in Subsection A.5.1, that better reflects the functional properties of real signals, compared to dimension reduction based on canonical variates for example. As we shall see in Subsection 6.3.4, and notwithstanding the substantially greater difficulties with this classification problem as compared to that studied in Hall, Poskitt and Presnell (2001), quadratic discrimination performs quite well in the current context.

Chapter 2

Tracking of Fault Lines in Regression Surfaces

2.1 Introduction

In the present chapter we introduce the prototype of the local kernel-based estimator which will also be the object of study in Chapters 3 and 4. The emphasis in this chapter is on the algorithmic aspect of the estimator. Specifically, it is designed to follow, or 'track,' a fault line once a starting point has been given to the algorithm. With high probability this method produces an estimator which never strays far from the fault line.

There are two principal advantages in using a local, rather than global, approach to fault line estimation. The first of these consists in an appreciable computational saving. As pointed out in Section 1.3, many of the previous approaches, in order to distinguish the smooth regions from those which contain the edge, examine the whole observation area. (However, as we also noted, the main thrust of many of these papers is on surface reconstruction, and because of this more ambitious goal it is then too simplistic to compare the computational expense directly.) On the other hand, the procedure suggested in this chapter involves the numerical minimisation of a function that can be computed at a cost of only $O(nh^2)$, where n is the mean number of points per unit area, and $h > 0$ is the kernel bandwidth. Another advantage consists in the better adaptability of our local method to possible erratic behaviour of the fault line. For the purpose of estimating a single fault line, which is our stipulation, this means that as long as the kernel bandwidth is sufficiently small (less than the reciprocal of curvature, assuming that this is admissible

under the conditions formulated in Section 2.3), the method can cope with any smooth but arbitrarily meandering fault line. This advantage becomes more prominent in the practically especially relevant situation of multiple or terminating fault lines, or fault lines that split at certain points to form an edge *graph*. See Remark 2.9 in Section 2.3.

The salient theoretical result presented in this chapter is the derivation of near minimax-optimality of the estimator, 'near' meaning that the rates are worse than the optimal rates by a factor that can be made as small as $O\{(\log n)^2\}$. By Theorem A.8, the minimax-optimal rate in the case of Poisson-distributed points, when there is no noise and the fault line has bounded curvature, is equal to $O(n^{-2/3})$, with a logarithmic factor in the case of uniform rates. As we have noted in the conclusion of Section A.8, the estimators constructed in the proof are not really practical, for a variety of reasons.

Section 2.2 describes our tracking method. Theoretical properties are presented in Section 2.3, with technical arguments deferred to Section 2.4. The discussion of numerical aspects will be postponed to Chapter 5, after other variants of the estimator have been introduced.

2.2 Methodology

Throughout the sequel, we assume that the general assumptions of Section 1.2 (see pp. 16ff) are satisfied. Suppose the locus of points on $\mathcal{C}$ is determined by the functions $(x^{(1)}(s), x^{(2)}(s))$, where s denotes the distance along $\mathcal{C}$ from a point Q at one end of the curve; and assume that $x^{(1)}(\cdot)$ and $x^{(2)}(\cdot)$ are smooth. No matter whether $\mathcal{C}$ is open or closed we can distinguish left- and right-hand sides of $\mathcal{C}$, with parity determined by tracing $\mathcal{C}$ in the direction of increasing s. By convention, neither the left nor the right side of $\mathcal{C}$ includes any part of $\mathcal{C}$ itself. We employ these data generated from realisations of the model at (1.2.3) to construct a tracking estimator $\widehat{\mathcal{C}}$ of $\mathcal{C}$, as follows.

Assume we have a starting point $\hat{x}_0 = \widehat{Q}$ and a starting estimate $\hat{\theta}$ of the orientation of the curve at $\widehat{Q}$. It may be supposed that all potential estimates of points on $\mathcal{C}$ are confined to a square grid $\mathcal{G}$, and likewise that estimates of the orientation of $\mathcal{C}$ are restricted to a discrete grid $\mathcal{G}'$. (If the process of points X_i was on a lattice then generally $\mathcal{G}$ would be much finer than this. The lattices $\mathcal{G}, \mathcal{G}'$ function as sieves in the interpretation of our definition at (2.2.2) below as that of a *sieved M-estimator*; see Van der Vaart and Wellner, 1996, p. 321.) Suppose we have constructed $\widehat{\mathcal{C}}$ as far as a point $\hat{x}_j$ ($j \in \mathbb{N}_0$), where our estimate of the tangent to $\mathcal{C}$, in the direction of travel along $\mathcal{C}$, is a unit vector $\hat{\theta}_j$. We

show how to construct the next point $\hat{x}_{j+1}$ with tangent estimate in the direction of another unit vector, $\hat{\theta}_{j+1}$.

Let $\mathcal{U}(\hat{x}_{j+1}, \theta_{j+1})$ denote the line through $\hat{x}_{j+1}$ in the direction of the unit vector θ_{j+1}. Locally at $\hat{x}_j$, the model uses the approximating function G, defined by

$$G(x|y, \theta, a, b) = \begin{cases} a\,, & (y-x)\cdot\theta^{\perp} \geq 0 \\ b\,, & \text{otherwise.} \end{cases} \tag{2.2.1}$$

For a bandwidth $h > 0$ and a smooth, radially symmetric bivariate probability density K whose support is the unit disc, choose $(x, \theta) = (\hat{x}_{j+1}, \hat{\theta}_{j+1})$ to minimise

$$S(x, \theta) = \inf_{c_1, c_2 \in \mathbb{R}} \sum_i \{Y_i - G(X_i|x, \theta, c_1, c_2)\}^2 \, K\big\{(\hat{x}_j - X_i)/h\big\}\,, \tag{2.2.2}$$

among those four or five of the eight grid neighbours x of $\hat{x}_j$ which are such that the unit vector in the direction of the line from $\hat{x}_j$ to x has a nonnegative dot product with ω. In order to exclude maximisers that could let the estimator stray off the curve $\mathcal{C}$, the search for θ is restricted to those $\theta \in \mathcal{G}'$ which satisfy the following

Search Cone Condition: For some $\tau > 0$, the angle $\theta \in \mathcal{G}'$ satisfies $h^2 \leq |\arg(\theta) - \arg(\theta_{\text{prev}})| \leq (1-\tau)h$, where θ_{prev} denotes the maximiser from the previous step.

Remark 2.1. *Noise variance estimation.* To estimate σ^2 one may take the variance of residuals corresponding to design points X_i that are near to $\widehat{\mathcal{C}}$, denoted $\delta_i^{\pm}$ in equations (2.4.1). This simple approach requires few extra computations, but its finite-sample performance is negatively affected by inherent bias especially in high-curvature parts of $\mathcal{C}$. Better performance may be achieved by extending the sampling strip further along the direction of an estimate of the normal at $x \in \mathcal{C}$ by an amount of size exactly $O(h^c)$, for any $c \in (1, 2)$.

Remark 2.2. *Robustification.* The least-squares criterion that is evinced by the form of the M-estimator at (2.2.2) yields, as is well known, an efficient estimator only in the case of Normally distributed errors. Ideas of robust M-estimation can expediently be applied in case that outliers are present. The proof in Section 2.4 may be adapted to accommodate outliers which comprise an asymptotically negligible fraction of the Y_i variables. For concepts of robust and efficient estimation, see Huber (1981) and Rousseeuw and Leroy (1987). Some aspects of robust estimation which especially pertain to computer vision can be found in Besl, Birch and Watson (1989) and Xu and Zhang (1996, pp. 93ff). We refer to the latter source for a more comprehensive account of the condensed discussion to follow. In order to estimate the scale parameter which enters

the 'ψ function,' and similar to Remark 2.1, the bias that arises when using local data can be reduced by giving less or no weight to data that is close to the currently-estimated segment of $\mathcal{C}$, defined by the current point x and the angle θ. The scale parameter is commonly estimated using the *median absolute deviation* (MAD) (see Rousseuw and Leroy, 1987, p. 45):

$$\mathrm{MAD}\big(\{Z_1, \ldots, Z_m\}\big) = \operatorname*{median}_i \big\{|Z_i - \operatorname*{median}_j Z_j|\big\}\,, \tag{2.2.3}$$

for $\{Z_1, \ldots, Z_m\}$ a sample of independent r.v.s. The scale parameter is estimated as the ratio MAD/0.6475. Here the factor $1/0.6745 = 1.4826$ is a calibration, making the scale estimator consistent for the standard deviation in the case where the Z_is are Normally distributed. It is pertinent to notice that with the presence of a more substantial number of outliers, identifiability problems arise. Our theory does not cover the case where the suspected source of the outliers is incorrect and instead the tails of the error distribution are heavy.

Our results and their proofs may be generalised to the case of arbitrarily fine grids $\mathcal{G}, \mathcal{G}'$, and even to the continuum. A rather theoretical definition of the estimator, in terms of infinitesimal quantities, that applies in the continuum is given in the paragraph containing (3.2.5). The method discussed theoretically in Section 2.4 is simplified so that technical arguments are relatively transparent, and so that the number of steps needed to traverse the curve is reasonably small without hindering accuracy. Therefore it is applied to relatively coarse grids $\mathcal{G}, \mathcal{G}'$; finer grids produce methods that are closer to being truly rotationally invariant.

2.3 Theoretical Properties

We assume that $\mathcal{C}$ is traced out within a compact rectangle, with its ends at points Q_L and Q_R on the left- and right-hand sides of $\mathcal{R}$, respectively; and that these sides have no other point with $\mathcal{C}$ in common. (Thus, $\Pi = \big[Q_L^{(1)}, Q_R^{(1)}\big] \times \mathbb{R}$.) In conjunction with condition (2.3.3) below, this assumption effectively requires that the fault line continue beyond the confines of Π, although we are only interested in estimating it within Π. Even if this is not the case, and in analogy to local linear techniques for curve estimation (e.g. Fan, 1993), our method of approximation is not significantly affected by the presence of edge effects. Let Q denote the first point on $\mathcal{C}$ that is distant h from the left-hand side of Π. We start tracing our estimator $\widehat{\mathcal{C}}$ at a point $\widehat{Q}$ distant h from the left-hand side of Π, representing an approximation to Q, and stop as soon as we get within h of the

other side of Π. Methods for calculating both $\widehat{Q}$ and a starting orientation are given in Remark 2.8.

The estimator is the piecewise-linear curve defined by joining successive estimates of points on the curve. The candidate points x and candidate unit vectors θ are required to satisfy the regularity conditions

$$\sup_{x\in\mathcal{G}} \inf_{\substack{x'\in\mathcal{G}\\ x'\neq x}} \|x-x'\| = O(h^2)\,, \qquad \sup_{\theta\in\mathcal{G}'} \inf_{\substack{\theta'\in\mathcal{G}'\\ \theta'\neq\theta}} |\arg(\theta')-\arg(\theta)| = O(h)\,. \tag{2.3.1}$$

Since the algorithm moves from one grid point in $\mathcal{G}$ to its immediate neighbour, this means that

$$\frac{\|\mathrm{ball}\big(\hat{x}_{j+1},h\big)\,\cap \mathrm{ball}\big(\hat{x}_j,h\big)\|}{\pi\,h^2} \to 1 \quad \text{as } h=h(n)\to 0\,. \tag{2.3.2}$$

Formula (3.2.4) for the content of the intersection lens is relevant in the present context as well. It should be noted that in the natural case where the grid $\mathcal{G}$ is a regular lattice, the argument to be used in (3.2.4) should be the one corresponding to the points in the corners, at angles $\pm\pi/4$ and $\pm 3\pi/4$.

Next we specify conditions ($\mathrm{C_{rs}}$) on the response surface. Let a general point P on the fault line $\mathcal{C}$ be represented by $(x^{(1)}(s), x^{(2)}(s))$, for $0\le s\le \ell$, where $\ell<\infty$, Q_L and Q_R are represented by $(x^{(1)}(0), x^{(2)}(0))$ and $(x^{(1)}(\ell), x^{(2)}(\ell))$ respectively, and s denotes the distance of P along $\mathcal{C}$ from Q_L. As in the general assumptions of Section 1.2, assume also that the first two derivatives of $x^{(1)}(\cdot)$ and $x^{(2)}(\cdot)$ are uniformly bounded on $[0,\ell]$. We assume too that g and each of its first derivatives is uniformly bounded in the intersection Π_L of Π with the left-hand side of $\mathcal{C}$, and also in the intersection Π_R of Π with the right-hand side; and that if, for $x\in\mathcal{C}$, we define

$$g_L(x) = \lim_{\substack{x'\to x\\ x'\in\Pi_L}} g(x')\,, \qquad g_R(x) = \lim_{\substack{x'\to x\\ x'\in\Pi_R}} g(x')\,,$$

then

$$d_{\min} \equiv \inf_{x\in\mathcal{C}} |g_L(x)-g_R(x)| > 0\,. \tag{2.3.3}$$

Qiu (1998, 2002b) also considered fault lines containing points x_0 such that $g_L(x_0) = g_R(x_0)$, while there exists $h>0$ such that $g_L\neq g_R$ on $\big(\mathrm{ball}(x_0,h)\setminus\{x_0\}\big)\cap\mathcal{C}$. In the cited papers these were termed (one category of) "singular points." Remark 2.6 deals with the behaviour of our estimator with regard to these singular points.

Further to the assumptions in Section 1.2, assume the errors ϵ_i have a distribution that has finite moment generating function in the neighbourhood of the origin; call this condition

(C_{err}). Suppose K is a circularly symmetric nonnegative function supported on $\overline{\text{ball}(0,1)}$, not vanishing on $\text{ball}(0,1)$, and Lipschitz continuous on $\mathbb{R}^2$. (In particular, it follows that K is a kernel of order 2.) Call these conditions (C_{ker}).

In agreement with the framework of Section 1.2, we assume that the point process $\{X_i\}$ is either homogeneous Poisson with intensity n, or is equal to a square lattice with n points per unit area. For ease of reference, call this condition (C_{pp}). Of the bandwidth h we assume the following condition, (C_{bw}): $h = h(n) \to 0$ as $n \to \infty$; in the Poisson case, $(\log n)^2/(nh^3) \to 0$; and in the lattice case, $1/(nh^4) \to 0$. (A logarithmic factor is not required in the lattice case; see Subsection 2.4.1 for discussion.)

Theorem 2.1. *Assume that conditions* (C_{bw}), (C_{err}), (C_{ker}), (C_{pp}) *and* (C_{rs}) *are satisfied, that* $\widehat{Q}$ *is within* C_1h^2 *of* Q*, and that the slope* $\hat{\theta}$ *is within* C_2h *of the slope of* $\mathcal{C}$ *at* Q *for some constants* $C_1, C_2 > 0$*. Then with probability 1, for some* $C_3 > 0$*,* $\widehat{\mathcal{C}}$ *is contained in the envelope of all points that lie within* C_3h^2 *of* $\mathcal{C}$*. Moreover, with probability 1 the algorithm terminates after* $O(h^{-2})$ *steps.*

It should be noted that, as emerges in the conclusion of the proof of the above theorem (see Section 2.4), the constants C_1, C_2 and C_3 are independent of h (and hence n) if h is sufficiently small, that size being determined by the geometry of $\mathcal{C}$ and the exact mesh width of the grids $\mathcal{G}, \mathcal{G}'$.

Corollary 2.1. *In the Poisson process and regular lattice cases, any convergence rate that is slower than* $n^{-2/3}(\log n)^{4/3}$ *or* $n^{-1/2}$*, respectively, is achievable almost surely by the estimator* $\widehat{\mathcal{C}}$*.*

Remark 2.3. *Correcting for location.* It is straightforward to state and prove a version of Theorem 2.1 in which location is re-estimated at each infinitesimal step. There, after moving an amount δ along the tangent we re-fit both location and orientation. The main conclusion remains true — that is, the estimated curve lies within $O(h^2)$ of the true curve, and in fact within $o(h^2)$ of a deterministic approximation, provided it starts within $O(h^2)$ of the true curve. We will elaborate on this interpretation in the paragraph below equation (2.4.10) (see p. 45). As illustrated in Figure 3.1 of Section 3.1, it is essential to use location re-fitting in order to obtain a good estimate even for fault lines with a simple geometry.

Remark 2.4. *Proximity to optimal convergence rates in the deterministic case.* That the convergence rate $O(n^{-2/3})$, modulo logarithmic factors, is optimal in the Poisson case follows from Theorem A.8. Likewise, the rates at which the slope of $\mathcal{C}$ is approximated by the value of θ, computed by least-squares as at (2.2.2), are anything slower than

$n^{-1/3}(\log n)^{2/3}$ or $n^{-1/4}$, in the Poisson and lattice cases respectively. These are within at most logarithmic factors of the best that are possible in a minimax sense. It is essentially this property which guarantees that, as claimed in the theorem, the algorithm concludes after $O(h^{-2})$ steps, since it ensures that the direction in which we move when estimating the curve by a sequence of points on an $O(h^2)$ grid is close to being the correct one — that is, along the curve.

Remark 2.5. *Oversmoothing.* Condition (C_{bw}) is deliberately constructed so as to produce enough oversmoothing to allow a relatively simple asymptotic description of $\widehat{\mathcal{C}}$. We may smooth a little less, producing a slightly faster convergence rate in the Poisson case (although only by a logarithmic factor), at the expense of a more complex asymptotic description of $\widehat{\mathcal{C}}$ and a longer proof. See also the similar comments in Remark 4.5. The oversmoothing also serves to make edge effects less palpable.

Remark 2.6. *Small jump height.* There is at least an occasional interest in the literature (for example Rudemo and Stryhn, 1994) in the case where the jump height, and particularly $d_{\min}$, tends to zero as $n \to \infty$. By considering the scale of the responses Y_i, it can be seen that this has the same effect as letting the error variance $\sigma^2 \to \infty$. Our proof will show that as long as $|g_L(x_0) - g_R(x_0)|$ is of order at least $h^{3/2}$, the conclusion of the theorem remains true. The reciprocal relationship between the jump height and the noise variance to which we have alluded will become evident in the definition of the signal-to-noise ratio SNR in (4.2.5), used in Theorem 4.2.

Remark 2.7. *More general point processes.* The theorem may be generalised to include the cases where $\mathcal{P}_n$ is a Poisson cluster process or a jittered grid process. The latter were discussed by Korostelev and Tsybakov (1993). In all these settings the conditions on the bandwidth, and the conclusions of the theorem and the corollary, are the same as in the Poisson process case. For Poisson and Poisson cluster processes (see the definition on p. 4) the intensity need not be constant. It is sufficient that the process have intensity $n\,\lambda(x)$ at each point $x \in \mathbb{R}^2$, where λ is kept fixed as n diverges, and is bounded away from zero and has bounded first derivatives in Π.

Remark 2.8. *Estimating the starting point and slope.* Assume $\mathcal{C}$ cuts the $x^{(2)}$ axis $\mathcal{A}$, with equation $x^{(1)} = 0$, at a point $Q = (0, q^{(2)})$ where $\mathcal{C}$ is not tangential to the axis; and that Q is unique within $\Pi \cap \mathcal{A}$. Suppose too that $\mathcal{C}$ has two bounded derivatives in a neighbourhood of Q. Let $h = h(n)$ denote a bandwidth sequence that satisfies (C_{bw}) in the Poisson case — that is, $(\log n)^2/(nh^3) \to 0$. Below we describe, in the Poisson case, a method for calculating an estimate $\widehat{Q}$ of Q which lies on $\mathcal{A}$ and, under the above conditions, achieves $O(h^2)$ accuracy a.s. It involves employing a univariate change-point

method to compute a pilot estimator $\widehat{Q}_1$, with an error of order $h^{(3/2)-\epsilon}$ for any given $\epsilon > 0$, and then refining this to $\widehat{Q}$ by using the least-squares criterion at (2.2.2) with x there taken equal to the coordinates of $\widehat{Q}_1$. In addition, the approach we describe below provides a starting orientation which achieves $O(h)$ accuracy with probability 1, so that the estimators obtained by this procedure satisfy the assumptions of the theorem, almost surely.

Let $h_1 = n^{-(1/2)+\Delta}$ where $0 < \Delta < \frac{1}{2}$, and project vertically onto $\mathcal{A}$ all those points $X_i \in \mathcal{P}_n$ within the band $|X_i^{(1)}| \le h_1$. This produces a linear point process $\{X_i^*\}$ with intensity $2nh_1$; if the original process $\{X_i\}$ is Poisson then so is $\{X_i^*\}$. Now any standard change-point estimator (see e.g. Müller, 1992; Eubank and Speckman, 1994; Gijbels, Hall and Kneip, 1999) may be applied to the data $\{(X_i^*, Y_i)\}$. Note that the latter process can also be construed as the independently marked point process where the marks Y_i are distributed according to (location-dependent) laws $F_i(\cdot|t - q^{(2)})$ for $i = 1, 2$, say, with the supports of F_1 and F_2 being at least $d_{\min}$ apart, for all parameter values $t \in \mathbb{R}$. The distribution of $\{Y_i\}$ is a mixture of the two distributions F_1 and F_2 with the weights p and $1 - p$ (or *vice versa*) and

$$p = \frac{h_1 + \min\{h \sin\theta, h_1\}}{2\, h_1}\,, \tag{2.3.4}$$

with $\theta \neq 0$ denoting the inclination angle of $\mathcal{C}$ to the axis $\mathcal{A}$. A suitable change-point algorithm yields an estimator $\widehat{Q}_1$ of Q, with coordinates $(0, \hat{x}^{(2)})$ say, subject to an error of $O\{(nh_1)^{-1+\delta} + h_1\}$ almost surely, for all $\delta > 0$. In view of our choice of h_1, this is of order $n^{-(1/2)+\Delta}$. Since h satisfies ($\mathrm{C_{bw}}$) then the error equals $O(h^{(3/2)-\epsilon})$ for each $\epsilon > 3\Delta$. Note that up to this point, the procedure closely resembles the one suggested by Rudemo and Stryhn (1994) for their regressogram estimator (see Section 1.3).

Choose $\Delta \in (0, \frac{1}{6})$, or alternatively $\epsilon \in (0, \frac{1}{2})$, and substitute the coordinates $(x^{(1)}, x^{(2)}) = (0, \hat{x}^{(2)})$ of $\widehat{Q}_1$ into the definition of $S(x, \theta)$ at (2.2.2). Choosing (x, θ) to minimise S produces the coordinate pair of a point $\widehat{Q}_2$ which, with probability 1, is within $O(h^2)$ of Q, and a starting orientation θ which, a.s., is within $O(h)$ of the slope of $\mathcal{C}$ at Q. (Methods outlined in Subsection 2.4.2 may be used to derive these results.) Taking $\widehat{Q} = (0, \eta)$ and $\hat{\theta} = \theta$, we obtain the desired starting point and starting orientation for the algorithm.

It should be noted that performance deteriorates sharply if the angle θ in (2.3.4) is small. Using a perpendicular direction or a family of dissecting lines, rather than only one, will improve performance. The additional computational cost, however, suggests the use of a conventional edge detection algorithm (see the references in Section 1.3). This may be

either for the purpose of estimating a viable direction for the line search of the change point algorithm, or (if theoretical concerns are not too prominent) by directly using the maximum of the JDC, as defined in Section 5.4, over the sample space Π, as the starting point. If multiple fault lines are anticipated to be present in Π then a modification of these schemes can be devised in a straightforward manner.

Remark 2.9. *Splitting and terminating fault lines.* It should be mentioned that the local estimator at (2.2.2) is constructed to be almost as simple as possible, commensurate with good performance. More elaborate local likelihoods can give improved performance in some settings. The approach can be adapted to situations where the true regression surface exhibits a fault line with a jump in the first derivative (a corner point or "kink"). There, the locally fitted surface would be constant on either side of a line through x that had a kink at x, with the two segments of the line having their normals in respective directions θ_1 and θ_2. In the same way that one determines jump discontinuities in the derivative of a univariate function from threshold exceedences of the differences between one-sided derivative estimates (e.g. Hall and Titterington, 1992), a kink may be deemed present in the neighbourhood of x if the distance between estimated values of θ_1 and θ_2, chosen to maximise the log-likelihood ratio subject to $\theta_1 \cdot \theta_2 \geq 0$, exceeds a threshold. However, estimation of kinks will be difficult unless the density of design points is relatively high, since kinks are high-order features (they relate to derivatives of the fault line, rather than directly to the fault line itself), and so are particularly susceptible to data sparseness problems associated with multivariate contexts. The simpler method suggested earlier, if applied to fault lines with kinks, will estimate the fault line at the same rate as before at all places that are not locations of kinks, and will consistently estimate kinks.

Unlike the testing problem described in the previous paragraph, there is a considerably lesser theoretical difficulty, but high practical relevance, associated with the case when the surface model underlying (2.2.2) is generalised in a somewhat different direction. Hall, Qiu and Rau (2002) consider the case where for a finite number of points $x_1, \ldots, x_p \in \Pi$, and for all $\theta \in [0, 2\,\pi)$, the limits along the rays $x_i + \mathbb{R}_+(\cos\theta, \sin\theta)$, $i = 1, \ldots, p$, exist and comprise a set of $k_i < \infty$ elements. If $\mathcal{C}$ is interpreted not as a single curve but as a mathematical graph then the points x_i can be termed *knots* of respective order k_i. In particular, a knot of order one is a point in whose neighbourhood the condition (2.3.3) is violated. In the present case, the regression surface is locally approximated by the

functions

$$G_k(y|x_0, \theta_1, \ldots, \theta_k, c_1, \ldots, c_k) = \sum_{i=1}^{k} c_i \, I\{\theta^{(i)} \leq \arg(y - x_0) < \theta^{(i+1)}\}$$

for $k \geq 2$, where $(\theta_i, c_i) \in [0, 2\pi) \times \mathbb{R}$ for $i = 1, \ldots, k$, $\theta^{(1)} \leq \theta^{(2)} \ldots \leq \theta^{(k)}$ is the order statistic for $(\theta_1, \ldots, \theta_k)$, and $\theta^{(k+1)} \equiv \theta^{(1)} + 2\pi$. Note that in particular $G_1(\cdot|x_0, c) \equiv c$, and that the function G appearing at (2.2.2) represents only a singleton in the class of functions for G_2, so that the problem of corner detection is rather complementary to the present discussion. Corner detection will be addressed in Section 5.4. It is straightforward to devise a variant of the knot tracking algorithm (with the additional numerical devices to be discussed in Section 5.2) that accommodates knots. Hall, Qiu and Rau (2002) showed that the estimator to which this algorithm gives rise is consistent for the true edge graph, under relatively mild conditions (including those given for the special situation of edges in Section 3.4). These conditions can often be assumed to be satisfied in image processing applications.

Other constraints could also be accommodated in the surface model defined by G_k, such as using response-surface slope. In combination with modelling curvature of $\mathcal{C}$, and with the bandwidth condition adjusted as in Remark 4.5 below, this yields a benefit in terms of convergence rate of the estimator, which is again within a logarithmic factor of the minimax-optimal rate $n^{-4/5}$. However, the price that is to be paid in the number of nuisance parameters (six rather than two) lessens the attractiveness of such a procedure, given the sparseness difficulties which affect the estimator of the present chapter, and even more so those to be constructed in Chapters 3 and 4. Especially if there is reason to believe that the corner arises through the meeting of two straight lines, a more robust low-level method such as cited in Remark 4.8 would usually be preferable for practical purposes.

2.4 Proof of Theorem 2.1

In order to make the arguments more lucid, we first give an outline of the main ideas of the proof in Subsection 2.4.1, before proceeding with the rigorous arguments in Subsection 2.4.2.

2.4.1 Intuitive Outline of Proof

First we consider the Poisson case. Let ball$(x, h) \subset \Pi$ be the disc representing the region of support of the kernel function $K\{(x - \cdot)/h\}$ appearing in formula (2.2.2), and with the points X_i distributed through it. Recall that the line $\mathcal{U} = \mathcal{U}(x, \theta)$ represents the boundary where the jump in the local linear model fitted at (2.2.2) occurs.

Suppose the part of $\mathcal{C}$ which lies within ball(x, h) is distant $C_1 h^2$, at its furthest point, from that part of $\mathcal{U}$ that lies within the disc. (The symbols $C_1, C_2 \ldots$ here and below denote positive constants.) Then within the disc there is a region, with area $O(h^3)$, between the line $\mathcal{U}$ and the curve $\mathcal{C}$. To see why the area is of this size, note that the region can be approximated by a rectangle of which one side is of length $C_2 h$ (the order of the length of a diameter of the disc; the line $\mathcal{U}$ will not be far from being a diameter), and the other is of length $C_3 h^2$ (the order of the distance between $\mathcal{U}$ and $\mathcal{C}$).

The region can be thought of as representing errors arising from two sources: (a) approximating the curve $\mathcal{C}$ by a straight line in our local linear model, and (b) putting the approximating line in the wrong position. When we fit the local linear model we can detect errors of types (a) and (b) if the number of Poisson-distributed points falling within the region is large enough. Since the region has area $O(h^3)$ then the number of points there is $O(nh^3)$, and so we can detect the errors if h is chosen so that nh^3 is sufficiently large.

We can determine what is required by 'sufficiently large' by arguing as follows. For small h, the response surface will be locally constant on either side of the jump, and we can estimate the two constants to within $o_p(1)$ by fitting the local linear model. Since by (2.3.3), the jump is $O(1)$ and not $o(1)$ then we do not have to be very accurate when estimating the constants. Now, the performance of the least-squares fit will deteriorate noticeably as soon as the number of points in the region increases beyond roughly $O(1)$, because then the fitted constant there will either be the one for the high side of the jump when it should be that for the low side, or *vice versa*. Taking account of the need to control moderate deviations, it turns out that 'roughly $O(1)$' has to be interpreted as a power of $\log n$.

Hence, we conclude that nh^3 should be no smaller than a certain power of $\log n$; or equivalently, that the diameter h of the original disc centred at x should be no smaller than $n^{-1/3}$, multiplied by a logarithmic factor. As the argument above suggests, if h is chosen in this way then the least-squares fitting procedure proposed at (2.2.2) can detect departures of up to $O(h^2)$ from the true curve. And, by its nature, having detected the departure the least-squares procedure applies an appropriate correction. It is critical

to this argument that we start the curve-tracking procedure at a point which is within $O(h^c)$ of the true curve, for some $c > 1$. But once that is done, the fact that we can detect each time the curve estimate wanders beyond $C_5 h^2$ away from the true curve (for some $C_5 > 0$), and correct for it, means that we can stay within the range $O(h^2)$ on all subsequent steps.

Similarly it can be shown that, provided the grid of θ values contains at least $O(h^{-1})$ elements, the estimate of the slope of $\mathcal{C}$ which is given by the value of θ defined by minimising $S(x', \theta)$ at (2.2.2), is within $O(h)$ of the true slope. (The $O(h)$ and $O(h^2)$ assertions here and above apply a.s., uniformly along the length of $\mathcal{C}$.)

When points of $\mathcal{P}_n$ are distributed on a lattice, rather than randomly distributed, an extra restriction is necessitated by the fact that regions which are narrower than a constant multiple of $n^{-1/2}$ (the distance between adjacent rows of points) might not receive any points at all. In particular, in order for the region whose width is h^2 to be guaranteed to have enough points to sustain our earlier argument, it is necessary that h^2 be an order of magnitude larger than $n^{-1/2}$. That is, in the case of lattice-distributed design points, h must be an order of magnitude larger than $n^{-1/4}$, as assumed in condition (C_{bw}).

2.4.2 Details of Proof

Let x be a point on the grid $\mathcal{G}$ of potential estimates of points on $\mathcal{C}$, and let x' be any point on $\mathcal{G}$ which is within $h^{3/2}$ of x. (We could replace $h^{3/2}$ by h^c for any $1 < c < 2$.) Let $\mathcal{T} = \mathcal{T}(x')$ denote the disc of radius $2h$ centred at x', and let $\mathcal{U}(x', \theta)$ be the line passing through x' in the direction of the unit vector θ. Then $\mathcal{U}(x', \theta)$ divides $\mathcal{T}$ into two half-discs, $\mathcal{T}_+(x', \theta)$ and $\mathcal{T}_-(x', \theta)$ say. Let $\sum_{i \in \mathcal{I}_\pm}$ denote summation over indices i such that $X_i \in \mathcal{T}_\pm(x', \theta)$. Write $\mathcal{G}_1$ for the set of all possible values of (x, x', θ); because we are working on a grid, this is a finite subset of $\mathbb{R}^5$. (For simplicity of notation we do not indicate dependence on x, and until the paragraph containing (2.4.8) and (2.4.9) we also suppress dependence on (x', θ)). For the respective choices of the plus and minus signs, put

$$K_i = K\big\{(x - X_i)/h\big\}\,, \qquad N_\pm = \sum_{i \in \mathcal{I}_\pm} K_i\,,$$

$$\mu_\pm\, N_\pm = \sum_{i \in \mathcal{I}_\pm} g(X_i)\, K_i\,, \qquad \bar\epsilon_\pm\, N_\pm = \sum_{i \in \mathcal{I}_\pm} \epsilon_i\, K_i\,,$$

$$\bar Y_\pm = \big(N_\pm\big)^{-1} \sum_{i \in \mathcal{I}_\pm} Y_i K_i = \mu_\pm + \bar\epsilon_\pm\,,$$

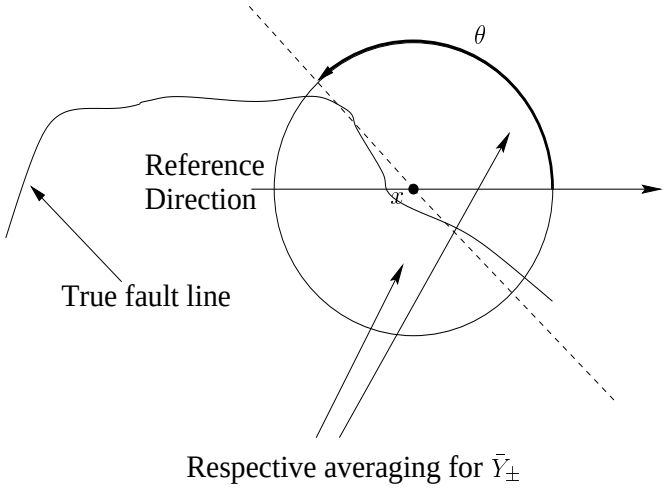

Figure 2.1: Illustration to equation (2.4.2).

$$\begin{aligned}
&\delta_i^\pm = g(X_i) - \mu_\pm\,, && \Delta_\pm = \sum_{i\in\mathcal{I}_\pm} \delta_i^\pm\,\epsilon_i\,K_i\,,\\
&S_1 = \sum_i \epsilon_i^2\,K_i\,, && S_2 = \sum_{i\in\mathcal{I}_+} \left(\delta_i^+\right)^2 K_i + \sum_{i\in\mathcal{I}_-} \left(\delta_i^-\right)^2 K_i\,,\\
&S_3 = \left(\bar{\epsilon}_+\right)^2 N_+ + \left(\bar{\epsilon}_-\right)^2 N_-\,, && S_4 = \Delta_+ + \Delta_-\,,\\
&S_5 = \bar{\epsilon}_+ \sum_{i\in\mathcal{I}_+} \delta_i^+\,K_i + \bar{\epsilon}_- \sum_{i\in\mathcal{I}_-} \delta_i^-\,K_i\,. && \qquad (2.4.1)
\end{aligned}$$

Note that the constants $M_\pm$ are Nadaraya-Watson kernel estimators of the regression surface (e.g. Korostelev and Tsybakov, 1993), locally on the '$\pm$' sides of x' and as determined by θ. In this notation, the quantity $S(x',\theta)$ defined at (2.2.2) is given by

$$\begin{aligned}
S(x',\theta) &= \sum_{i\in\mathcal{I}_+} \left(Y_i - \bar{Y}_+\right)^2 K_i + \sum_{i\in\mathcal{I}_-} \left(Y_i - \bar{Y}_-\right)^2 K_i\\
&= S_1 + S_2 - S_3 + 2\,S_4 - 2\,S_5\,. \qquad (2.4.2)
\end{aligned}$$

In passing from (2.2.2) to (2.4.2) we have found the values of c_1 and c_2 that minimise the sum of squares on the right-hand side of (2.2.2), and substituted them back into the formula. See Figure 2.1 for an illustration.

By the Cauchy-Schwarz inequality, $|S_5| \le (S_2 S_3)^{1/2}$, and so

$$|S(x',\theta) - (S_1 + S_2)| \le S_3 + 2|S_4| + 2\,(S_2\,S_3)^{1/2}\,. \qquad (2.4.3)$$

Next we develop bounds to $T_1 \equiv \bar{\epsilon}_\pm N_\pm$ and $T_2 \equiv \Delta_\pm$. Both T_1 and T_2 may be written in the form $T = \sum_i \epsilon_i w_i$, where the weights w_i are such that $w = \sup_i |w_i|$ and $W^2 = \sum_i w_i^2$ are both finite. Since, by ($\mathrm{C_{err}}$), the distribution of the errors ϵ_i has a finite moment generating function γ, say, in a neighbourhood of the origin, then there exist constants $D_1, D_2 > 0$ such that $|\log \gamma(t)| \le D_1 t^2$ whenever $|t| \le D_2$. Given $u > 0$, put $z = uW$ and $t = \min\{u/(2D_1W), D_2/w\}$. Then, by Markov's inequality, using condition ($\mathrm{C_{err}}$) and noting that $|tw_i| \le D_2$ for each i, we have that $P'(T \ge z) \le E'(e^{tT-tz}) \le \exp(D_1 t^2 W^2 - tz)$, where P' and E' denote probability and expectation conditional on the points of $\mathcal{P}_n$ (this is only relevant if $\mathcal{P}_n$ represents a Poisson process). Considering separately the cases $u \le 2D_1D_2W/w$ and $u > 2D_1D_2W/w$ we may show that for all $u > 0$, $D_1 t^2 W^2 - tz \le -C_1 u \min(u, W/w)$, where $C_1 > 0$ depends only on D_1 and D_2. Furthermore, in the cases $T = T_1$ and $T = T_2$ we have respectively $w \le \|K\|_\infty$ and $w \le 2\,\|K\|_\infty \|f\|_\infty$. From these results, and their analogues in the opposite tail, we deduce that in either case,

$$P'(|T| > u\,W) \le 2 \exp\big\{ - C_2\, u \min(u, W)\big\}\,. \tag{2.4.4}$$

Note that in the case of T_k, $W = W_k$ where $W_1^2 = \sum_{i \in \mathcal{I}_\pm} K_i^2$ and $W_2^2 = \sum_{i \in \mathcal{I}_\pm} (\delta_i^\pm)^2\, K_i^2$. When ambiguity could otherwise occur we shall write T_k as $T_k^\pm$ and W_k as $W_k^\pm$. In this notation, and using (2.4.4) and an obvious notation for the sum over the $\pm$ sides of $\mathcal{C}$, we have

$$\begin{aligned}
P'\big\{(S_2 S_3)^{1/2} &\ge (2S_2 \sup K)^{1/2}\, u\big\} \\
&= P'\big\{S_3 \ge 2\,(\sup K)\, u^2\big\} \le \sum_{j \in \{+,-\}} P'\big\{\big(T_1^{(j)}\big)^2 / N_\pm > (\sup K)\, u^2\big\} \\
&\le \sum_{j \in \{+,-\}} P'\big(T_1^{(j)} \ge W_1^{(j)}\, u\big) \le 2 \sum_{j \in \{+,-\}} \exp\big\{ - C_2\, u \min\big(u, W_1^{(j)}\big)\big\}\,.
\end{aligned}$$

Also, provided $u \ge 1$,

$$\begin{aligned}
P'\big\{|S_4| > 2\,u^2 \max\big(S_2^{1/2}, 1\big)\big\} &\le \sum_{j \in \{+,-\}} P'\big\{\big|T_2^{(j)}\big| > u^2 \max\big(W_2^{(j)}, 1\big)\big\} \\
&\le 2 \sum_{j \in \{+,-\}} \exp\Big[- C_2\, u^2 \max\big(1, 1/W_2^{(j)}\big) \\
&\qquad \times \min\big\{u^2 \max\big(1, 1/W_2^{(j)}\big), W_2^{(j)}\big\}\Big] \\
&\le 4 \exp\big(- C_2\, u^2\big)\,.
\end{aligned}$$

Therefore, by (2.4.3), if $u \geq 1$,

$$|S(x', \theta) - (S_1 + S_2)| \leq C_3 \, u^2 \, \max\left(S_2^{1/2}, 1\right) Z_1 \,, \tag{2.4.5}$$

where $C_3 > 0$ depends only on $\|K\|_\infty$, and the nonnegative random variable $Z_1 = Z_1(u)$ satisfies

$$P'(Z_1 > 1) \leq 6 \sum_{j \in \{+,-\}} \exp\left\{ - C_2 \, u \, \min\left(u, W_1^{(j)}\right)\right\} . \tag{2.4.6}$$

If the points in $\mathcal{P}_n$ are Poisson-distributed then, for all $C > 0$,

$$P\left[\inf_{(x,x',\theta) \in \mathcal{G}_1} \left\{ \min\left(W_1^+, W_1^-\right)\right\}^2 \geq C_4 \, n \, h^2\right] = 1 - O\big(n^{-C}\big) \,, \tag{2.4.7}$$

where $C_4 > 0$ depends only on K. To appreciate why, use Markov's inequality to prove that $P\{|(W_1^\pm)^2 - E(W_1^\pm)^2| > \epsilon n h^2\} = O(n^{-C})$ for all $\epsilon, C > 0$, uniformly in $(x, x', \theta) \in \mathcal{G}_1$. Direct calculation shows that $E(W_1^\pm)^2/(nh^2)$ is bounded away from 0 uniformly in $(x, x', \theta) \in \mathcal{G}_1$. From these properties and the fact that $\mathcal{G}_1$ has $O(n^C)$ elements for some $C > 0$, we obtain (2.4.7).

When the points in $\mathcal{P}_n$ are on a square lattice, condition ($\mathrm{C_{bw}}$) implies that for all sufficiently large n, $(W_1^\pm)^2/(nh^2)$ is bounded away from 0 uniformly in $(x, x', \theta) \in \mathcal{G}_1$. In this case, (2.4.7) holds in a degenerate form, with its right-hand side replaced by 1, for some $C_4 > 0$.

From (2.4.5), (2.4.6) and (2.4.7), and the fact that $nh^2/\log n \to \infty$ under condition ($\mathrm{C_{bw}}$), we may show that for any $C_5 > 0$ there exists $C_6 > 0$ such that, uniformly in choices of $(x, x', \theta) \in \mathcal{G}_1$,

$$\begin{aligned} &\big|S(x',\theta) - \{S_1 + S_2(x', \theta)\}\big| \\ &\qquad \leq C_6 \, (\log n) \, \max\left\{S_2(x', \theta)^{1/2}, 1\right\} Z_2(x', \theta) \,, \end{aligned} \tag{2.4.8}$$

$$P\{Z_2(x', \theta) > 1\} \leq C_6 \, n^{-C_5} \,. \tag{2.4.9}$$

(Take u in (2.4.5) and (2.4.6) to equal a sufficiently large constant multiple of $(\log n)^{1/2}$.) From this point we explicitly state the dependence of S_k on (x', θ).

Since the number of elements of $\mathcal{G}_1$ is only polynomially large in n, and since the constant C_5 at (2.4.9) may be taken arbitrarily large, then by (2.4.8),(2.4.9) and the Borel-Cantelli

lemma we have that a.s., for all sufficiently large n,

$$|S(x',\theta) - \{S_1 + S_2(x',\theta)\}| \leq C_7 \,(\log n)\, \max\left\{S_2(x',\theta)^{1/2}, 1\right\} \quad \text{for all} \quad (x,x',\theta) \in \mathcal{G}_1\,. \tag{2.4.10}$$

We note in passing that in case that if an intrinsically 'continuous' viewpoint of the algorithm is adopted (cf. the closing remark of Section 2.2), the proof can be maintained by introducing the grid of points $\left(x_i,\theta_i\right) \in \mathcal{G}\times\mathcal{G}'$ only at this stage, with $\mathcal{G},\mathcal{G}'$ satisfying the same conditions (2.3.1). The arguments up to and leading to formula (2.4.18) will remain valid. What is crucial here is that there exists a triplet $(x'',\theta'') \in \mathcal{G}\times\mathcal{G}'$ such that $\operatorname{card}\left\{\mathcal{X}\cap\left\{\mathcal{T}_+(x',\theta)\mathbin{\triangle}\mathcal{T}_+(x'',\theta'')\right\}\right\} = O(1)$. We also point out that the idea of this symmetric difference approximation naturally leads to a deterministic tracking algorithm in the continuum, pretending that the regression surface is observed without noise, and with the bandwidth parameter held fixed. At each tracking step, the tangent is placed in such a way that the kernel-weighted area assigned to the 'wrong' side of the kernel disc, related to the symmetric difference from earlier in this paragraph, is minimised. That is, the target function is the area enclosed by the diameter of the kernel disc and the fault line; and an infinitesimal element da of that area which is distant u from the diameter contributes an amount of $K(u)\,da$ to the target function. Ties may be broken by minimising the distance to the previously-estimated slope of the disc diameter. As will be of relevance in Chapter 4, a correction of location as well as slope (cf. Remark 2.3) may also be incorporated. The assessment of the performance of the estimator in terms of the L^1 norm will be discussed later in Remark 4.3; cf. also the numerical results in Section 5.3.

Returning to the main proof, and in a similar manner as for obtaining (2.4.10), by going back to the bound at (2.4.4) and taking u equal to a sufficiently large constant multiple of $(\log n)^{1/2}$, we may prove that a.s., for all sufficiently large n,

$$\left|\epsilon_\pm(x',\theta)\right| N_\pm(x',\theta)^{1/2} \leq C_7\,\log n \quad \text{for all} \quad (x,x',\theta) \in \mathcal{G}_1\,. \tag{2.4.11}$$

This concludes the part of the proof that is concerned with the randomness induced by the additive errors ϵ_i, while the following part addresses the effects of the randomness of

design. Let $D \equiv \|\text{grad}\, f\,\|_\infty$ (the supremum norm is taken over Π) and

$$b = b(x') = \begin{cases} \text{length}\big(\mathcal{C} \cap \mathcal{T}(x')\big)^{-1} \displaystyle\int\limits_{\mathcal{C}\cap\mathcal{T}(x')} |f_L(x) - f_R(x)|\, dx\,, & \mathcal{C} \cap \mathcal{T}(x') \neq \emptyset \\ 0 & \text{otherwise}\,, \end{cases}$$

in other words, b equals the integral mean of $|f_L - f_R|$ over $\mathcal{C} \cap \mathcal{T}(x')$; note that by (2.3.3), $b \geq d_{\min} > 0$. Let

$$\mathcal{R}_{\pm,+} = \Pi_L \cap \mathcal{T}_\pm(x', \theta)\,, \quad \mathcal{R}_{\pm,-} = \Pi_R \cap \mathcal{T}_\pm(x', \theta)\,,$$

with Π_L and Π_R as the left- and right-hand sides of $\mathcal{C}$ respectively as in the paragraph containing (2.3.3), and with the $\pm$ signs taken respectively, and let

$$N_{\pm,+} = N_{\pm,+}(x', \theta) = \sum_{i: X_i \in \mathcal{R}_{\pm,+}} K_i\,, \qquad N_{\pm,-} = \sum_{i: X_i \in \mathcal{R}_{\pm,-}} K_i\,.$$

Then, if $X_i \in \mathcal{T}(x')$ is on the left-hand side of $\mathcal{C}$,

$$b\, N_{\pm,-} \big(N_\pm\big)^{-1} - 4\, D\, h \leq \big|\delta_i^\pm\big| \leq b\, N_{\pm,-} \big(N_\pm\big)^{-1} + 4\, D\, h\,, \tag{2.4.12}$$

and if X_i is on the right-hand side of $\mathcal{C}$ then the same is true provided $N_{\pm,-}$ is replaced by $N_{\pm,+}$. (Here and in (2.4.13) below we suppress dependence of $N_\pm$, $N_{\pm,+}$, $N_{\pm,-}$ and $\delta_i^\pm$ on (x', θ).) To prove (2.4.12), and skipping the simpler case $b = 0$, choose $z \in \mathcal{C} \cap \mathcal{T}(x')$ such that $|g_L(z) - g_R(z)| = b$. The assertion follows by applying the triangle inequality to each of the paired terms in the expression

$$\begin{aligned} N_\pm\, \delta_i^\pm = &\sum_{m: X_m \in R_{\pm,+}} \big\{[g(X_i) - g_R(z)] + (g_R(z) - g(X_m)]\big\} \\ &+ \sum_{m: X_m \in R_{\pm,-}} \big\{[g(X_i) - g_R(z)] + [g_R(z) - g_L(z)] + [g_L(z) - g(X_m)]\big\}\,. \end{aligned}$$

Returning to the main proof, squaring, multiplying by K_i, adding over i such that $X_i \in \mathcal{T}_\pm(x', \theta)$, and noting that

$$\big(N_{\pm,-}/N_\pm\big)^2\, N_{\pm,+} + \big(N_{\pm,+}/N_\pm\big)^2\, N_{\pm,-} = N_{\pm,+}\, N_{\pm,-}/N_\pm\,,$$

we deduce that

$$b\,(b - 16\,Dh)\,N_{\pm,+}\,N_{\pm,-}\,\big(N_\pm\big)^{-1} \le \sum_{i \in \mathcal{I}_\pm} \big(\delta_i^\pm\big)^2\,K_i \le b\,(b + 16\,Dh)\,N_{\pm,+}\,N_{\pm,-}\,\big(N_\pm\big)^{-1} + 16\,N_\pm\,D^2\,h^2\,. \tag{2.4.13}$$

Next we prove that for a constant $C_8 > 0$, and for $j = 1, 2$,

$$\big(N_\pm - C_8 nh^2\big)\,\big(nh^2\big)^{-1} = o(1)\,, \tag{2.4.14}$$

where the random variable represented by '$o(1)$' is of that size uniformly in $(x, x', \theta) \in \mathcal{G}_1$, a.s. To derive (2.4.14), consider first the case where $\mathcal{P}_n$ is a Poisson process. Let $A_\pm$ denote the number of nonvanishing terms in the series defining $N_\pm$, and define

$$\pi\big(C, A_\pm\big) \equiv P\big\{\big|N_\pm - E\big(N_\pm\big|A_\pm\big)\big| > C\,\mathrm{var}\big(N_\pm\big|A_\pm\big)^{1/2}\log n\big|A_\pm\big\}\,.$$

We consider this conditional expectation separately on the sets $\mathcal{F}_1 = \big\{A_\pm \le (\log n)^2\big\}$ and $\mathcal{F}_2 = \big\{A_\pm > (\log n)^2\big\}$. Use Bernstein's inequality (Theorem 1.1) to show that it holds generally true that if X is a Poisson-distributed random variable with mean $\mu = \mu(n)$, satisfying $\mu(n)/\log n \to \infty$ as $n \to \infty$, then $P(|X - \mu| > \epsilon\mu) = O(n^{-C})$ for all $\epsilon, C > 0$. We apply this to the probability of the set $\mathcal{F}_1$. A second application of Bernstein's theorem is used to deal with the conditional probability on $\mathcal{F}_2$, bounding its probability multiplier by 1. In this context, observe too that

$$0 < C_9\,A_\pm \le \mathrm{var}(N_\pm|A_\pm) \le \|K\|_\infty^2 A_\pm,$$

for a constant C_9 depending only on K. Hence it follows that if $C > 0$ is given then there exists $C' > 0$ such that

$$P\big\{\big|A_\pm - E\big(A_\pm\big)\big| > C'\,\big(\mathrm{var}A_\pm\big)^{1/2}\log n\big\} = O\big(n^{-C}\big)\,.$$

Combining these results with the Borel-Cantelli lemma, and exploiting the fact that the number of elements of $\mathcal{G}_1$ is only polynomially large in n, we obtain (2.4.14) in the Poisson case.

The context where $\mathcal{P}_n$ is a lattice may be treated more directly, since there $N_\pm$ is a deterministic sum. It is asymptotic to a constant multiple of nh^2, uniformly in (x, x'). In the lattice case the assertion 'a.s.' is not required when describing (2.4.14).

Let $\mathcal{I} = \mathcal{I}(x', \theta)$ be the line segment of length $4h$ defined as the diameter of $\mathcal{T} = \mathcal{T}(x')$

that is aligned in direction θ, and put $\mathcal{C}_{\mathcal{T}} = \mathcal{C} \cap \mathcal{T}$. In what follows, we shall temporarily make use of the following

'Distance' Assumption: $\vec{d}_H(\mathcal{C}_{\mathcal{T}}, \mathcal{I}) \leq h^{3/2}$. (The directed Hausdorff distance $\vec{d}_H$ is defined at (A.8.1)).

When the 'distance' assumption holds, one of $\mathcal{R}_{\pm,+}$ and $\mathcal{R}_{\pm,-}$ is of area not exceeding $4h^{5/2}$; let it be $\mathcal{R}_{\pm,k_\pm}$. The other region is of area at least $2\pi h^2 - 4h^{5/2}$. (Recall that $\mathcal{T}(x')$ is of radius $2h$.) Let this region be $\mathcal{R}_{\pm,\ell_\pm}$. The argument used to derive (2.4.14) may be employed to show that

$$N_{\pm,k_\pm}\left(nh^2\right)^{-1} = o(1) \tag{2.4.15}$$

uniformly in $(x, x', \theta) \in \mathcal{G}_1$ for which the 'distance' assumption holds, with probability 1. Combining this result with (2.4.14) we conclude that $N_{\pm,\ell_\pm}/N_\pm = 1 + o(1)$, where the '$o(1)$' term has the same interpretation as that at (2.4.15). Hence,

$$\frac{N_{+,+}N_{+,-}}{N_+} + \frac{N_{-,+}N_{-,-}}{N_-} = \{1 + o(1)\}\left(N_{+,k_+} + N_{-,k_-}\right), \tag{2.4.16}$$

again with the same interpretation of '$o(1)$.'

Let $\mathcal{J} = \mathcal{J}(x', \theta) = \mathcal{R}_{+,k_+} \cup \mathcal{R}_{-,k_-}$ denote the closed region within $\mathcal{T}(x')$ that is bordered by $\mathcal{C}_{\mathcal{T}}$, $\mathcal{I}$ and the perimeter of $\mathcal{T}(x')$. Thus, $\mathcal{J}$ has a vertex at any point where $\mathcal{C}_{\mathcal{T}}$ intersects $\mathcal{I}$; it follows from ($\mathrm{C_{rs}}$) that the number of such vertices does not exceed two, for h small enough. In the case where $\mathcal{P}_n$ is Poisson, let $A = A(x', \theta)$ equal the integral of $K\{\|x - y\|/h\}$ over $y \in \mathcal{J}$, and in the case where $\mathcal{P}_n$ is a lattice, let A be the average of this kernel over all grid points y that fall within $\mathcal{J}$. In the first case, $A(x', \theta)$ can be computed via the Poisson process $\mathcal{P}_{n,K}$ which is a (deterministic) transformation of $\mathcal{P}_n$, and is defined for Borel sets $B \subseteq \mathbb{R}$ as follows:

$$\mathcal{P}_{n,K}(B) = \mathcal{P}_{n,K}(B|x, x', \theta) \equiv \sum_i I\big(K_i(x) \in B\big)\,.$$

The rate function of this process is obtained by calculating

$$\begin{aligned} E(\mathcal{P}_{n,K}(B)) &= \sum_{r=0}^{\infty} E(\mathcal{P}_{n,K}(B) \mid \mathcal{P}_n(B) = r)\, P(\mathcal{P}_n(B) = r) \\ &= \sum_{r=0}^{\infty} \frac{r}{\|B\|} \int_B K\left(\frac{x-y}{h}\right) dy\, \frac{(n\|B\|)^r}{r!}\, \exp\big(-n\|B\|\big) \end{aligned}$$

$$= n \int_B K\left(\frac{x-y}{h}\right) dy\,, \tag{2.4.17}$$

and taking $B = \mathcal{R}_{+,k_+} \cup \mathcal{R}_{-,k_-}$ shows that $E(N_+, k_+ + N_-, k_-) = nA$. We claim that, with probability 1, uniformly in values $(x, x', \theta) \in \mathcal{G}_1$ for which the 'distance' assumption holds,

$$N_{+,k_+} + N_{-,k_-} = \{1 + o(1)\}\, n\, A(x', \theta) + O\big\{(\log n)^2\big\}\,, \tag{2.4.18}$$

whence it follows from (2.4.16) that

$$\frac{N_{+,+}N_{+,-}}{N_+} + \frac{N_{-,+}N_{-,-}}{N_-} = \{1 + o(1)\}\, n\, A(x', \theta) + O\big\{(\log n)^2\big\}\,, \tag{2.4.19}$$

a.s., uniformly in the same range of values of x, x', θ.

To appreciate why (2.4.18) holds, consider first the case where $\mathcal{P}_n$ is Poisson. Let $A_1 = A_1(x', \theta) = \mathcal{J}(x', \theta) \cap \mathrm{ball}(x, 1)$. Using the argument leading to (2.4.14) and (2.4.15) we may prove that, given $C > 0$, we can choose $C' > 0$ so large that for $j = 1, 2$,

$$P\big\{\big|N_{+,k_+} - E\big(N_{+,k_+}\big)\big| > C'\,(nA_1)^{1/2}\,\log n\big\} = O\big(n^{-C}\big)$$

uniformly in $(x, x', \theta) \in \mathcal{G}_1$ for which the 'distance' assumption holds and $nA_1 \geq (\log n)^2$. To treat the case where $nA_1 < (\log n)^2$ we note that when this inequality holds, the set of which A_1 is the area is contained within a set of which the area is $A_2 = n^{-1}(\log n)^2$. Hence, given $C > 0$, we may choose $C' > 0$ so large that

$$P\big[\big|N_{\pm,k_\pm} - E\big(N_{\pm,k_\pm}\big)\big| > C'\,\big\{(nA_1)^{1/2} + \log n\big\}\log n\big] = O\big(n^{-C}\big)$$

uniformly in $(x, x', \theta) \in \mathcal{G}_1$ for which the 'distance' assumption holds, this time without regard for the sign of $nA_1 - (\log n)^2$. Result (2.4.18) follows from this property and the observations preceding (2.4.17).

The case where $\mathcal{P}_n$ is defined on a lattice may be treated by a counting squares argument, specifically by Lemma 2.1.1 in Huxley (1996, p. 27). It suffices to consider the case where h is small enough so that A consists of at most three connected (solid) components. Also, the con-convex parts of A, which we may assume to be the set for the point count N_{-,k_-}, can be treated as the difference of the relevant half-disc (a convex set) and the part of the interior of the parabola that is in the same half-disc. The aforementioned lemma by Huxley (1996) bounds the number of squares forming an inner or outer cover of a convex set. We also need to account for the deviations incurred by the kernel K, where we need the Lipschitz smoothness condition. Clearly, the length of the curve delineating the set

of area N_{+,k_+} is $O(h)$. By similar reasoning for the '$-$' side and the previous remarks, it follows in the context of gridded design that

$$N_{+,k_+} + N_{-,k_-} = nA + o(nA) + (O\sqrt{n}h)\,.$$

From (2.4.13), (2.4.14), (2.4.19) and the fact that $(\log n)^2/nh^3 \to 0$ (or $1/(nh^4) \to 0$ in the case of gridded design) we see that a.s., uniformly in values $(x, x', \theta) \in \mathcal{G}_1$ for which the 'distance' assumption holds,

$$S_2(x', \theta) = \sum_{i \in \mathcal{I}_+} \big(\delta_i^+\big)^2 K_i + \sum_{i \in \mathcal{I}_-} \big(\delta_i^-\big)^2 K_i = \{1+o(1)\}\, n\, b(x')^2\, A(x', \theta) + o\big(nh^3\big)\,. \quad (2.4.20)$$

From (2.4.10) and (2.4.20) we conclude that a.s.,

$$S(x', \theta) = S_1 + \{1 + o(1)\}\, n\, b(x')^2\, A(x', \theta) + o\big(nh^3\big)\,, \quad (2.4.21)$$

uniformly in $(x, x', \theta) \in \mathcal{G}_1$ satisfying the 'distance' assumption.

Note that while S_1 depends on x it does not depend on (x', θ). Therefore, (2.4.21) implies that if we minimise $A(x', \theta)$ rather than $S(x', \theta)$, with respect to (x', θ), we merely neglect terms in the target function which are of order $o(A + h^3)$, uniformly in (x, x', θ). Although these functions will generally have multiple minimisers, the following arguments are valid for any two respective choices, as assumption ($\mathrm{C_{pp}}$) ensures that all minimisers are contained within a rectangle of size $O(h^2) \times O(h)$, a.s. For the purposes of the next paragraph we view the choice of $(\xi_0, \eta_0, \theta_0) \in \{\operatorname{argmin} A(x', \theta)\}$ as a function of h, and proceed to show that the choice of (ξ_0, η_0) and θ_0 differs from the 'correct' choices only by terms of order $O(h^2)$ and $O(h)$, respectively. This discussion is essentially the discrete version of the considerations in the proof of Theorem 4.1. Indeed, the statements in the next paragraph can be deduced by a combination of Lemma 4.1 and a Riemann sum approximation to the integrals (area calculations) appearing there.

We start with noticing that assumption ($\mathrm{C_{rs}}$) implies that $\mathcal{C}$ is locally quadratic. For greater ease of exposition we first consider the case when the area is not given kernel weights, although we continue to use the notation $A(x', \theta)$ for this modified target function. Geometric arguments may then be employed to show that if x' is distant further than $O(h^2)$ from $\mathcal{C}$, with θ arbitrary, then $A(x', \theta)$ is strictly larger than $A(\xi_0, \eta_0, \theta_0)$, for sufficiently small h. Similarly it can be shown that if the sequence x' is within $O(h^2)$ of the true curve but the slope component θ is such that there exists $C > 0$ with $\|\theta - \theta_0\| \geq Ch$, then for h small enough there exists a grid point (ξ', η') between the line and the parabola

with the property $A(\xi', \eta', \theta_0) < A(x', \theta)$. (All foregoing statements hold a.s.) The previous arguments may be transferred to the case of kernel-weighting, if we too observe that as K is assumed to be Lipschitz continuous, the function $K\{\|x - \cdot\|/h\}$ converges uniformly to $K\{\|x' - \cdot\|/h\}$ as $x \to x'$. In summary, we thus have established that up to terms of sufficiently small order, the algorithm minimises $A(x', \theta)$ a.s., and we claim that the theorem follows from this property.

To explain why, we note the following four properties. In points (a),(b) below we refer to the location component of the minimiser of $A(x', \theta)$, while (a′),(b′) refer to its slope. (a) There exist constants $0 < C_1 < C_2$ such that, if the point P with coordinates x at which we are situated at a given step is further than C_2h^2 from $\mathcal{C}$, then the grid point x' that results from minimising $A(x', \theta)$ is less than C_1h^2 from $\mathcal{C}$. (b) If $C_3 > 0$ is given then there exists $C_4 = C_4(C_3) > 0$ such that, provided x is distant no more than C_3h^2 from $\mathcal{C}$, the value of x' that results from minimising $A(x', \theta)$ is distant no more than C_4h^2 from $\mathcal{C}$. In (a) and (b), distance of a point to $\mathcal{C}$ denotes shortest distance. The slope component θ which minimises $A(x', \theta)$, has the following properties (a′),(b′) which respectively correspond to (a),(b). (a′) There exist constants $0 < C_5 < C_6$ such that, if the slope estimate ω from the previous step is further than C_5h from the true slope of the curve at the point of $\mathcal{C}$ that is closest to x', then the grid point θ that results from minimising $A(x', \theta)$ is less than C_6h from that slope. Here and in in (b′), we represent the true slope of the point on $\mathcal{C}$ to which we refer, by a unit vector such that the dot product with θ is non-negative. (b′) If $C_7 > 0$ is given then there exists $C_8 = C_8(C_7) > 0$ such that, provided ω is distant no more than C_7h from the true slope on the corresponding point on $\mathcal{C}$ as given in (a′), the value of θ that results from minimising $A(x', \theta)$ is distant no more than C_8h from that slope. (The constants $C_1, C_2, C_4(C_3), C_5, C_6, C_8(C_7)$ depend on the maximum curvature of $\mathcal{C}$ and on the constants B_1, B_2 in the definition of the grid of points x' and θ, but for h sufficiently small they do not depend on h.)

Properties (a),(a′) and result (2.4.21) imply that: (c) if the grid vertex x is further than C_2h^2 from $\mathcal{C}$ then at the next step we move to a grid vertex x' that is less than $C_1h^2 + o(h^2)$ from $\mathcal{C}$. Properties (b),(b′) imply that: (d) if at the nth step we are at a point x which is distant no more than C_3h^2 from $\mathcal{C}$, then x' will be distant no more than $C_4(C_3)h^2 + o(h^2)$ from $\mathcal{C}$. Furthermore, in view of the uniformity of the remainder at (2.4.21), the $o(h^2)$ remainders here are uniformly small. Now, $\widehat{Q}$ and Q, and the corresponding slopes, are distant $O(h^2)$ and $O(h)$ apart, respectively. (The last assertion would have to be qualified 'a.s.' if Q and the slope at Q were estimated by a procedure such as described in Remark 2.8, but the conclusions, being subject to the same restriction, would not be affected thereby.) These initial conditions, and properties (c) and (d), imply that at each

step the point estimate must uniformly lie within $O(h^2)$ of the curve. Also, due to the same initial conditions, the initial triplet (x', θ') (be it deterministic, or the result of an estimation procedure such as described in Remark 2.8) satisfies the 'distance' assumption a.s. Then the fact that subsequent points are uniformly distant $O(h^2)$ from $\mathcal{C}$ implies that this is valid throughout, a.s.

Similarly, using (a$'$) and (b$'$) from two paragraphs earlier, it may be proved that the estimate of the slope of $\mathcal{C}$, provided by the value of θ obtained from minimising with respect to (x', θ'), is uniformly within $O(h)$ of the true slope of the curve at the point of $\mathcal{C}$ that is closest to x'. Therefore, the sign convention for passing from a slope estimate θ (a unit vector) at x to the slope estimate θ' at x' (see the paragraph containing (2.2.2)) ensures that successive point estimates progress steadily along the curve, on the grid $\mathcal{G}$ with edge width equal to a constant multiple of h^2, so that only $O(h^{-2})$ steps are required until termination within h of the opposite side of Π.

Chapter 3

Local Likelihood Tracking of Fault Lines

3.1 Introduction

In Chapter 2 we have introduced a tracking estimator based on kernel-weighted local least squares. The procedure suggested in Section 2.2 limited itself to the estimation of the tracking direction for the edge. Notwithstanding the fact that this estimator is theoretically sound, computational drawbacks even for relatively simple curves should be recognised. One of these arises when curvature of the edge is large in comparison to the step width of the tracking procedure.

Figure 3.1 shows a simple example where the true fault line is a circle. The true regression surface, with equation

$$g\big(x^{(1)}, x^{(2)}\big) = \begin{cases} 1 - 8\big\{(x^{(1)} - 0.5)^2 + (x^{(2)} - 0.5)^2\big\}, & (x^{(1)} - 0.5)^2 + (x^{(2)} - 0.5)^2 \leq \frac{1}{16} \\ 0 & \text{otherwise,} \end{cases} \tag{3.1.1}$$

is shown in Figure 1(a). This surface, which is very similar in structure to that used in Qiu (2002), has a jump of constant magnitude 0.5 along the circle $\mathcal{S} = \big\{x : \|x - (0.5, 0.5)\| = 0.25\big\}$. Based on the description in the paragraph following equation (2.4.10), and even in the noise-free setting, it is not hard to prove that regardless how small the step size, the tracked points will, except for degenerate cases, spiral away from the true curve until the latter has been lost completely. This phenomenon is illustrated in panel (b), where

we checked $m = 300$ angles in the minimisation of (2.2.2). (If we let $m \to \infty$, the ensuing estimator would be asymptotically equal to a circle that is concentric with $\mathcal{S}$, and with slightly larger radius. On the other hand, the likelihood surface ridge studied in Chapter 4 would consist of an *inscribed* circle. This can be derived from Theorem 4.1.) In Figure 3.1, tracking takes place in both directions from the starting point $Q = (0.75, 0.5)$. If the surface at (3.1.1) was modified such that it took the value 1 inside $\mathcal{S}$, it would emerge more clearly that the algorithm produces an estimator that has at each point a similar derivative as in the projection on that point onto $\mathcal{C}$.

The performance of the tracking estimate is slightly improved when a small amount of noise is added. Panel (c) shows the original data with the superposition of independent and $\mathrm{N}(0, 0.25^2)$ distributed noise. On this occasion the data are shown in form of an image, because the plot of type (a) would not reproduce well. Figure (d) shows the outcome in this situation.

In view of these problems, it seems natural to follow the suggestion of Remark 2.3 and to consider the likelihood as a function of both slope and location. Benefits gained from the study of the arising likelihood *surface* were already mentioned in Section 1.2. In the present chapter we exploit this idea and point out again how in many practically relevant situations that exhibit fault lines of moderately complex nature, the computational workload can be held within acceptable limits. From the viewpoint of image analysis, the results of this chapter show that basic local likelihood ideas can be successfully transferred into a purely statistical framework.

A major advantage of the local-likelihood principle is that it can be adapted to a range of problems that are often considered in an *ad hoc* fashion. We begin with the previously-studied setup of fault lines in regression surfaces in Section 3.2, before moving on to the case of fault lines in the support of densities in Section 3.3. The main results are given in Section 3.4, with the theoretical arguments being deferred to Section 3.5.

3.2 Fault Lines in Response Surfaces

In this section we again adopt the standard assumptions of Section 1.2. Somewhat differently from Chapter 2, we are now constructing the estimator $\widehat{\mathcal{C}}$ from a *local likelihood*. The critical parameter in this technique is the bandwidth which, as in similar problems in density function estimation studied earlier (Copas, 1995; Hjort and Jones, 1996), is a sort of tuning parameter for choosing between a simple parametric model, which is in general only approximately true, and a fully nonparametric model.

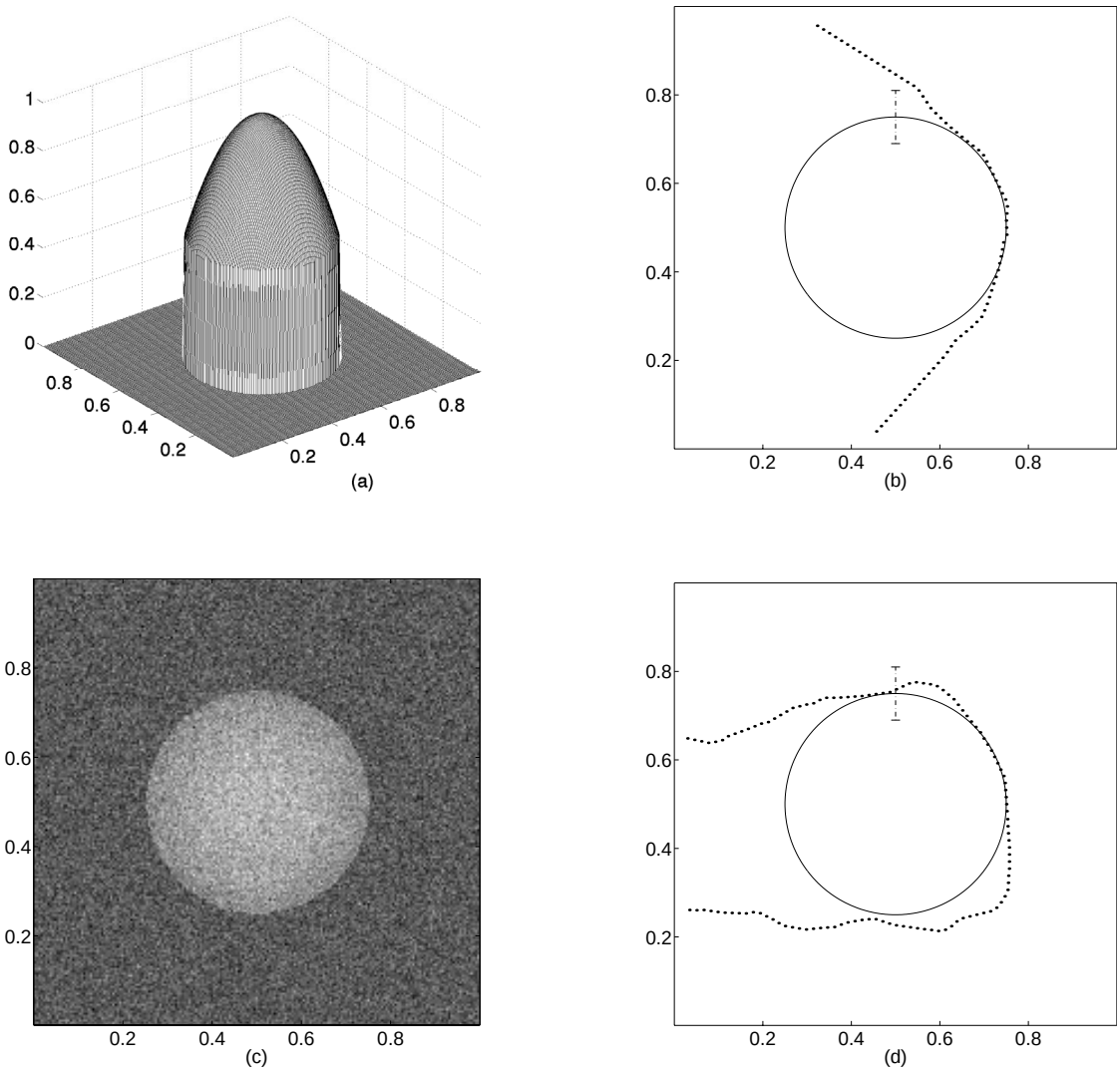

Figure 3.1: Behaviour of the tracking estimate from Chapter 2. Panel (a) shows the non-corrupted regression surface defined by equation (3.1.1). Panel (b) shows the result of tracking with starting point $x_0 = (0.75, 0.5)$, kernel bandwidth $h = 0.08$, and step size $\delta = 0.2\,h$. Panel (c) depicts the image corresponding to (a), observed with superposition of $\mathrm{N}(0, 0.25^2)$ distributed noise. Panel (d) shows the performance of the algorithm for the data in (c).

The local likelihood is thresholded as indicated in the next paragraph, to yield points in the sample space Π which are deemed edge points. In the standard reformulation of the estimation problem as one of testing, thresholding amounts to determination of critical points. It should be noted that it is here that we pay the prize for the locally-parametric character of our model, due to the presence of bias in implicit estimates of both the fault line and the response surface. Only if the fault line is exactly linear, and the response surface exactly constant on either side of the fault line, does bias not complicate the use of a formal testing procedure. Thus, the application of a fully data-driven approach in any formal way seems to be made impractical. However, from a practical viewpoint this does not heavily subtract from the feasibility of the approach, at least as long the parametric model is a reasonable approximation. The more complex case where an additional testing is employed in order to find points where edges split (see Remark 2.9) was shown to be tractable by this method in two numerical examples in Hall, Qiu and Rau (2002).

The estimator $\widehat{\mathcal{C}}$ is defined using concepts that are closely related to *hysteresis* thresholding of edges in image processing, for example the Canny edge detector (see pp. 107f of this thesis), and comprises all those points $x \in \Pi$ which have either

- a log-likelihood ratio above a pre-determined threshold $t_{\text{upper}} > 0$, or
- are connected to any such point via the tracking algorithm from Chapter 2, and have a log-likelihood ratio above a second pre-determined threshold $t_{\text{lower}} < t_{\text{upper}}$.

Thresholds t_{lower} and t_{upper} would usually be chosen from prior experience with the data concerned. Choosing a relatively small value of t_{lower}, the notion of 'small' being dependent on the nature of the data that we are working with, yields longer ridge lines, but usually also some spurious ridges. As these will typically comprise only a small piece of a curve, they can easily be eliminated with methods used in Section 5.4. Conditions on t_{lower} and t_{upper} are given in Section 3.4, where theoretical properties of the estimator are discussed.

In order to construct the approximate likelihood, we temporarily assume that the errors $\epsilon_i = Y_i - g(X_i)$ are *Normally* distributed. Then we can interpret the quantity $S(x, \theta)$ at (2.2.2) as a kernel-weighted local log-likelihood that $\mathcal{C}$ passes through the point x, and has its tangent oriented in the direction of the vector θ. The maximiser of the log-likelihood for the alternative case, that is when $\text{ball}(x, h) \cap \mathcal{C} = \emptyset$, is simply the locally constant estimator known as the Nadaraya-Watson kernel estimator (e.g. Korostelev and Tsybakov, 1993). Thus we obtain the following expression for the local log-likelihood

ratio, up to additive and proportionality constants:

$$\begin{aligned}\ell(x,\theta) &= \Bigg[\sum_i \{Y_i - \bar{Y}(x)\}^2 K_i(x) - \sum_{i \in \mathcal{I}_+(x,\theta)} \{Y_i - \bar{Y}_+(x,\theta)\}^2 K_i(x) \\ &\qquad - \sum_{i \in \mathcal{I}_-(x,\theta)} \{Y_i - \bar{Y}_-(x,\theta)\}^2 K_i(x) \Bigg] N(x)^{-1} \\ &= \frac{N_+(x,\theta)\, N_-(x,\theta)}{N(x)^2} \left\{ \bar{Y}_+(x,\theta) - \bar{Y}_-(x,\theta) \right\}^2 , \end{aligned} \tag{3.2.1}$$

where we adopt the convention $0/0 = 0$; see also Remark 3.1 below. In (3.2.1), the definitions of $\mathcal{I}_\pm$ and the quantities $N_\pm, \bar{Y}_\pm$ and K_i are the same as in Section 2.2 and at (2.4.1), but taking $x' = x$ there, so that no double bandwidth is used. Also, $N(x) = \sum_i K_i(x)$, and the quantity $\bar{Y}(x) = N(x)^{-1} \sum_i Y_i\, K_i(x)$ is the aforementioned Nadaraya-Watson kernel estimator. The reason for not using double bandwidths, in contrast to Chapter 2, is that here we do not employ a grid for the potential estimates of $\mathcal{C}$ as was done to facilitate the proof of consistency. In this regard, we also note that the 'continuous' view of the edge estimator, mentioned in the paragraph below (2.4.10), harmonises better with the differential-geometric stance of this chapter and Chapter 4. Figure 2.1 again serves to illustrate the foregoing definitions.

Alternatively to reading the definition of ℓ at (3.2.1) as a local log-likelihood ratio, it can also be considered in terms of local least squares, as is evident from the supremum operation over θ in (3.2.2). This alternative viewpoint motivates the abandonment of the assumption that the error variables are Normally distributed.

In order to construct $\mathcal{C}$ from a grid of points obtained in the way described in the previous paragraph, and in the finite-sample context, $\widehat{\mathcal{C}}$ may be thought of as the limit, as $\delta \to 0$, of the polygonal estimates $\widehat{\mathcal{C}}_\delta$ whose vertices are obtained, for a given realisation of (X_i, Y_i), as follows. Cover Π with a finite family $\{\Upsilon_k = \Upsilon_{k,\delta}\}$ of dissecting lines such that

$$\max_k \min_{k' \neq k} d_H\big(\Upsilon_k, \Upsilon_{k'}\big) \leq \delta \,,$$

so that their interspacing is at most δ. (The Hausdorff distance d_H was defined at (A.8.2).) Problems with parts of the fault line that are (nearly) tangential to the transects Υ_k are clearly more acute than in the discussion in the last paragraph of Remark 2.8, and there is no simple (and nonrandom) rule to overcome it. Along each of the Υ_k, search for the (local or global) maximisers $\hat{x}_0 = \hat{x}_{0,k}$ as at (3.2.2), requiring as at the beginning of this section that $m(\hat{x}_{0,k}) > t_{\text{upper}}$. The ridge estimate may then then obtained by a relatively standard rule which connects close neighbours, similar to techniques used in Section 5.4.

For regression surfaces with edges that have a high signal-to-noise ratio (see the definition at (4.2.5)), an already appealing result will usually be obtained by connecting the points $\hat{x}_{0,k}$ in such a way that progression through Π is maintained, for example by stipulating that the $x^{(1)}$ coordinate of the point sequence is monotonically increasing.

Remark 3.1. *Data sparseness.* To avoid difficulties caused by data sparseness one would consider only values of $\ell(x, \theta)$ for which the minimum of $\inf_\theta N_\pm(x, \theta)$ over values of the $\pm$ signs exceeded a predetermined threshold, ν say. In practice this might be achieved by choosing h to depend on x, and taking it sufficiently large in each case. What 'sufficiently large' means is primarily dependent on the geometry of $\mathcal{C}$. Note that because of the 'search cone condition' from Section 2.2 that will again be stipulated here, it is straightforward to verify that if $\theta_* \in [0, 2\pi)$ denotes the previous tracking direction, then

$$\frac{\inf_\theta N_+(x, \theta)}{N_+(x, \theta_*)} \to 1 \quad \text{a.s.}, \quad n \to \infty \,,$$

and hence, by the observations leading to (2.4.17),

$$\begin{aligned} P\big(\inf_\theta N_+(x, \theta) \ge \nu, \inf_\theta N_-(x, \theta) \ge \nu\big) &\approx P\big(N_+(x, \theta_*) \ge \nu, N_-(x, \theta_*) \ge \nu\big) \\ &= \left\{ \sum_{r=\langle \nu \rangle + 1}^{\infty} \frac{\big(\frac{1}{2}\, nh^2\big)^r}{r!} \exp\big(-\tfrac{1}{2} nh^2\big) \right\}^2 . \end{aligned}$$

If the sparseness probability is to be bounded along longer stretches of $\mathcal{C}$, a Bonferroni-type argument similar to that in Remark 4.3 may be used.

Note that (3.2.1) is only approximately equal to the true likelihood, for several reasons. Apart from those already pointed out in Chapter 2, note that in the definition at (3.2.1) we have standardised by dividing by the 'effective local sample size,' $N(x)$, which is also an approximation to the number of degrees of freedom of the local likelihood-ratio test. It is pertinent to notice that $2\, N_\pm(x, \theta)/(nh^2) \to C(x)$, a constant only depending on x, and that $\bar{Y}_+(x, \theta) - \bar{Y}_-(x, \theta)$ is approximately Normally distributed with variance $2\big\{\frac{1}{2}\, nh^2 C(x)\big\}^{-1}\sigma^2$. Somewhat surprisingly, it turns out, however, that a version of Wilks' theorem does not hold in the present context, in that the limiting distribution derived from the likelihood surface $\big\{m(x), x \in \Pi\big\}$ defined at (3.2.2) below is not of a chi-squared type; see Theorem 4.2 in Chapter 4. The adjustment by dividing by the local sample size can be important in practice if we are to achieve reasonable performance, since it makes it feasible to compare likelihoods at different x values. However, it has no influence on theoretical convergence rates discussed in Section 3.4.

Note that $\ell(x,\theta)$ is nonnegative, and that if $\text{ball}(x,h) \cap \mathcal{C} \neq \emptyset$ then $\ell(x,\theta)$ indeed tends to assume larger values. Therefore, we estimate $\mathcal{C}$ as a ridge of $\ell(x,\theta)$, adopting the notion of Definition A.1. Strictly speaking, the ridge is well-defined only in relation to the surface derived from g instead of the empirical data (X_i, Y_i), using definitions that are analogous to those around (3.2.1) but with replacing sums by expressions involving integrals; for example, $\bar{Y}_\pm(x,\theta)$ is replaced by the convolution $2g(\cdot) * (H\,K)\{(\cdot - x)/h\}$, where $H \equiv \{y : \pm\theta^\perp \cdot (y - x) \geq 0\}$. However, ridges will be referred to as belonging to the estimates in the indicated way without further mention.

For the purpose of ridge finding we need an initial approximation $\hat{x}_0$ to a point on $\mathcal{C}$. For example, we might take

$$\hat{x}_0 = \operatorname*{argsup}_x m(x)\,, \qquad \text{where } m(x) = \sup_\theta \ell(x,\theta)\,, \tag{3.2.2}$$

and we calculate the argsup within a relatively small region through which we expect $\mathcal{C}$ to pass. The validity of this definition of $\hat{x}_0$ is discussed in Remark 3.2. Formula (3.2.2) assumes that $t_{\text{upper}} > \|m\|_\infty$, otherwise we would conclude that there is no significant evidence for a fault line. Alternatively, we might draw a line transect across the plane of the explanatory dataset, so that the line traverses a region through which we expect $\mathcal{C}$ to pass, and project onto the transect those values of X_i in its neighbourhood; and compute $\hat{x}_0$ by solving a relatively simple change-point problem in univariate regression, for the new dataset, as described in Remark 2.8. The measurability of $\hat{x}_0$ is ensured by the theory given in Section A.3.

As a corollary to the definition of the likelihood ridge, if $\hat{x}$ denotes a point on $\widehat{\mathcal{C}}$ then the value of $m(x)$ decreases if x moves to either side of $\hat{x}$ in a direction perpendicular to the tangent to $\widehat{\mathcal{C}}$ at $\hat{x}$. The orientation of this tangent is that of the unit vector $\hat{\theta} = \text{argmax}_\theta\, \ell(\hat{x},\theta)$, or of its negative value.

Define $\hat{\theta}(x) = \text{argsup}_\theta\, \ell(x,\theta)$. With probability 1, $\hat{\theta}(x)$ is uniquely determined by this definition, except for sign. To trace out the fault line estimate numerically we may use the algorithm described in Section 2.2, employing the threshold checking rules given earlier (see p. 56). As evinced by the example in Section 3.1, the tracking should also employ location correction. Alternatively, one could select the point $\hat{x}$ that maximises $m(x)$ on the semicircle

$$\mathcal{A} = \{y : y = x + \delta e\,, \|e\| = 1\,, |\arg(e) - \phi_0| < \pi/2\}\,, \tag{3.2.3}$$

centred at the previous tracking direction ϕ_0 and with radius $\pi/2$. The movement is now

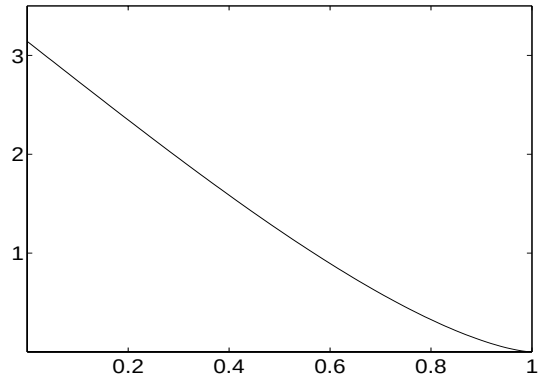

Figure 3.2: The lens function $B(x)$, defined at (3.2.4).

to the point (in the continuum) which is located in direction $\hat{\theta}(x)$, taking care of parity, and distant δ from the previous point. Unless prior information is available, one would in practice usually choose a compromise of the two alternatives mentioned, and select the maximiser of $\ell(x, \cdot)$ from the arc $\mathcal{A}$ defined as at (3.2.3) but with $\pi/2$ replaced by an angle $\Delta\phi \in (0, \pi/2)$. (Note that this is in line with the 'search cone condition' from Section 2.2.) As was done in the introductory example of this chapter, we may also trace out $\mathcal{C}$ in the opposite direction, starting from $\hat{x}_0$.

In a finite-sample context, the choice of δ would depend on assumptions on maximum curvature of $\mathcal{C}$. For example, if the bandwidth h was chosen to vary with the location on the curve estimate, this could be done by fixing the ratio δ/h at a constant value $C \in (0, 1)$. If curvature is large then we would prefer a larger overlap of the kernel smoothing discs in two subsequent steps. Note that the lens of intersection between two unit discs whose centres are $2x$ apart has the area (see e.g. Hall, 1988, p. 67)

$$B(x) = \pi - 2\,x\sqrt{1 - x^2} - 2\arcsin x\,, \quad 0 \le x \le 1\,, \tag{3.2.4}$$

and in the present context $x = \delta/2$. Function $B(x)$ is illustrated in Figure 3.2. If the left-hand side of expression (2.3.2) was specified to have the value 0.5, for example, this would require choosing δ/h as the solution of $B(x)/\pi = 0.5$, or $\delta = 0.808\,h$.

Remark 3.2. *Measurability.* If the design points X_i are gridded and the noise distribution has atoms (for example, in the case of salt-and-pepper noise defined in the paragraph containing (1.2.3)), then it is possible for the maximiser of $\ell(x, \theta)$ to be non-unique. Moreover, as noted by Kim and Pollard (1990, p. 194) in a very similar context, "Arbitrary tie-breaking rules [...] raise other questions regarding measurability of the argmax." However, complete rigour is achievable by using the theoretical framework from Section A.3.

This framework is especially important in the case where a discontinuous (e.g., uniform) kernel function is used, in which case the likelihood of this and the next section are not continuous functions of their arguments. It should be noted that numerical optimisation is an awkward task in this case. For the purposes of the present chapter, we shall adhere to the simpler definition of the estimator via the argmax functional, and defer a more rigorous discussion of the problems with the argmax functional to the proof of the main theorem (Theorem 4.2) of Chapter 4; see Section 4.3.2.

The theoretical formula for our estimator $\widehat{\mathcal{C}}$ of $\mathcal{C}$ has $\delta = 0$ in either of the definitions in the paragraph containing (3.2.3). More precisely, assuming that tracking in one of the variants described in the previous paragraph has taken us up to the point with coordinates $P = (\hat{x}^{(1)}, \hat{x}^{(2)})$, say, the next infinitesimal step in our construction of $\widehat{\mathcal{C}}$ is taken in the direction of $\hat{\theta}(s + ds)$, away from P and in the direction opposite from the estimated starting point Q. The curve $\widehat{\mathcal{C}}$ of the theoretical formula may be thought of as the limit, as $\delta \to 0$, of the sequence of steps defined by $(\hat{x}^{(1)}_{j+1}, \hat{x}^{(2)}_{j+1}) = (\hat{x}^{(1)}_j, \hat{x}^{(2)}_j) + \delta\hat{\theta}_j$ and $(\hat{x}^{(1)}_0, \hat{x}^{(2)}_0) = (\hat{x}^{(1)}(0), \hat{x}^{(2)}(0))$, where $\hat{\theta}_j$ denotes the vector that minimises $S(\theta)$ when $\hat{x}_j = (\hat{x}^{(1)}_j, \hat{x}^{(2)}_j)$ and is chosen to point away from the direction from which we have come. Specifically, if P is distant s, along $\widehat{\mathcal{C}}$, from Q; if $\hat{\theta}$ equals the gradient computed at P; and if $\hat{x}^{(1)}_1(s)$ and $\hat{x}^{(2)}_1(s)$ denote the first and second components, respectively, of $\hat{\theta}$; then

$$\hat{x}^{(1)}(s) = \hat{x}^{(1)}(0) + \int_0^s \hat{x}^{(1)}_1(t)\, dt\,, \quad \hat{x}^{(2)}(s) = \hat{x}^{(2)}(0) + \int_0^s \hat{x}^{(2)}_1(t)\, dt\,. \tag{3.2.5}$$

In this notation, $\widehat{\mathcal{C}}$ is defined as the plot of $(\hat{x}^{(1)}(s), \hat{x}^{(2)}(s))$ as a function of s. Our estimator of a general point on $\mathcal{C}$ is $\hat{x}$, say (one of the $\hat{x}_j$s), and our estimator of the direction of the normal to $\mathcal{C}$ at $\hat{x}$ is $\operatorname{argsup}_\theta w(\hat{x}, \theta)$.

If there exist several separate fault lines, the number of which may not be known, the algorithm described in the paragraph below the one which contains (3.2.2) can still be applied. Once the tracking of a ridge line is completed, the subsequent search region in that equation is chosen to exclude a sufficiently small closed neighbourhood of that tracking estimate; in the ith step ($i \in \mathbb{N}_0$), let this neighbourhood be denoted F_i. The estimation process is complete at step M when $\cup_{i=0}^{M} F_i = \Pi$, with $(M + 1)$ being the number of fault lines detected in Π. In case that the starting points are searched along line transects passing through Π, as suggested earlier, invariance of the estimator with respect to rotations of the design points does no longer hold true.

In the presence of intersecting fault lines, we suggest two strategies that can to some extent be combined with each other. The first of these, which is more directly data-

driven, draws on ideas from Hall, Qiu and Rau (2002). An excerpt of their methodology has been given in Remark 2.9. In the notation of that remark, we have $\theta^{(i+2)} = \theta^{(i)} + \pi$ for $i = 1, 2$. Hall, Qiu and Rau (2002) used a second local (log)likelihood in order to estimate such knots; the associated likelihood surface has its peaks at the loci of knots. This likelihood is thresholded separately, obeying the same conditions as for t_{lower} and t_{upper} given in Section 3.4. This of course compounds the difficulty associated with the choice of threshold. An alternative, and more graphical, approach consists in examination of the ridge surface $\mathcal{M}$ based on the differential-geometric concepts from Section A.2. It was pointed out there that a point on a ridge line may be found by traversing a contour, using for example the algorithm $x_j \mapsto x_{j+1} = x_j \pm \delta\, v_\perp(x_j)$ (for a small positive δ, and with either the + or the − sign used throughout), starting at any given point in Π; and working out the value of $\|m\|$ at each step. Whatever approach is being adopted, the assumption, in our theory in Section 3.4, that fault lines do not intersect is mainly for technical convenience.

As mentioned in Remark 2.9, the likelihood can be designed in a straightforward manner to accommodate local surface models that are even more complex than the pattern created by two or more essentially superimposed fault lines. The changes necessitated thereby reduce, rather than add to, the computational (and theoretical) burden.

3.3 Fault Lines in Intensity Surfaces

Here we observe, as stated in condition 2 of Section 1.2 (see p. 19), i.i.d. bivariate data X_i, for $1 \le i \le n$, from a distribution with density g. The density is smooth except for a fault line running through its support, $\mathcal{S}$. We wish to estimate this line. To this end we write down the likelihood ratio for a test of the null hypothesis that g is locally constant at x, versus the alternative that g has a jump discontinuity along a line that passes through x with its normal in direction θ, and g is locally constant on either side of the jump. The kernel-weighted local (log)likelihood ratio is proportional to

$$\begin{aligned}\ell(x,\theta) &= \log\left\{\left(\frac{N_+}{N}\right)^{N_+}\left(\frac{N_-}{N}\right)^{N_-}\right\} N(x)^{-1} \\ &= \Big\{N_+(x,\theta)\,\log N_+(x,\theta) + N_-(x,\theta)\,\log N_-(x,\theta) \\ &\qquad - N(x)\,\log N(x)\Big\}\, N(x)^{-1}\,, \end{aligned} \tag{3.3.1}$$

where $N_\pm$ is as defined in the paragraph containing (3.2.1). Where the above expression is undefined, we set $\ell(x,\theta) = -\infty$. However, as in Section 3.2, we would restrict attention

to xs for which $\inf_\theta \min_{+,-} N_\pm(x,\theta) \geq \nu$, a threshold. Also as in Section 3.2, we have divided by $N(x)$ so as to make local likelihoods at different locations comparable.

Note that $\ell(x,\theta)$ takes values in $[-\log 2, 0] \cup \{-\infty\}$, and, if the disc centred at x and with radius h lies across $\mathcal{C}$, $\ell(x,\theta)$ tends to assume larger values when the alternative hypothesis is true. The tracking algorithm suggested in Section 3.2 applies without change in the present setting. There are obvious analogues of the methods there for computing a starting point for the algorithm. Our local-likelihood approach is rather simple to apply relative to, for example, "maximin" methods (e.g. Müller and Song, 1994), since it is based on ridges of a smooth surface rather than heights of jumps of a discontinuous, erratic surface.

3.4 Theoretical Properties

Because of the parallels with the more extensive proof in Section 4.3.2, in the exposition in this and the next section we limit ourselves to the salient points. For the sake of simplicity we shall limit ourselves here to i.i.d. data; this allows us to consider the regression and density cases together. However, in the setting of response surfaces our methods may be used to derive convergence rates when the Xs are generated differently, for example on a grid. This configuration will be addressed at the end of this section. In both cases we shall assume that the (continuous version of the) 'search cone condition' of Section 2.2 holds, and that:

(a) the distribution of X has a density that is bounded away from 0 in an open neighbourhood of $\mathcal{C}$;

(b) in neighbourhoods on either side (but not both sides simultaneously) of $\mathcal{C}$, the response or intensity surface has a bounded derivative;

(c) the size of the jump at the fault line is bounded away from 0 along $\mathcal{C}$;

(d) the kernel K is a nonnegative, radially symmetric function with support equal to ball$(0,1)$, and is Hölder continuous in $\mathbb{R}^2$;

(e) the bandwidth $h = h(n)$ satisfies $h \to 0$ and $n^{1-\epsilon}h^2 \to \infty$, for some $\epsilon > 0$, as $n \to \infty$;

(f) the local likelihood estimator is started either at a point $\hat{x}_0$ that is within $O(h^2)$ of $\mathcal{C}$, or at $\hat{x}_0 = \text{argsup}_x\, m(x)$, where the argsup is computed within a sufficiently small open neighbourhood of $\mathcal{C}$; and

(g) for some $c > 0$ and with $t_+ \in \{t_{\text{upper}}, t_{\text{lower}}\}$ (cf. p. 56), it holds that $n^c = O(t_+)$ and $t_+ = o(nh^2)$ as $n \to \infty$.

In the context of response surfaces, and as stated in Section 1.2, we ask in addition that

(h) the distribution of $Y - g(X)$ does not depend on X and has finite moment generating function in a neighbourhood of the origin.

In condition (d), the assumption of Hölder continuity is not essential and in fact the results below may be derived using uniform weighting schemes. As stated in Remark 3.2, difficulties with defining the maximisers of ℓ when dispensing with the assumption that K satisfies (d) arise primarily for numerical reasons. Condition (e) ensures that there exist thresholds satisfying (g).

Theorem 3.1. *In the cases of response and intensity surfaces, assume conditions (a)–(g), and for response surfaces assume in addition (h). Then with probability 1, and provided we construct local-likelihood estimators within a sufficiently small but fixed neighbourhood of $\mathcal{C}$, the estimators of points on $\mathcal{C}$, and of the angles made by tangents to $\mathcal{C}$ at those points, are accurate to within $O\{(nh)^{-1} \log n + h^2\}$ and $O\{(nh^2)^{-1} \log n + h\}$, respectively, uniformly on any subset of $\mathcal{C}$ that starts at least ϵ from one end of $\mathcal{C}$ and finishes at least ϵ from the other end, for any $\epsilon > 0$.*

Taking h equal to a constant multiple of $(n^{-1} \log n)^{1/3}$ we see that the uniform rate of convergence of our estimator to $\mathcal{C}$ is $O\{(n^{-1} \log n)^{2/3}\}$. As pointed out before, this is within a logarithmic factor of the minimax-optimal pointwise convergence rate in both response and intensity cases.

For both response and intensity surfaces, when the Xs are arranged in a regular lattice (triangular, square or hexagonal) with n points per unit area of the x plane, an argument similar to that used to derive Theorem 3.1 may be employed to prove that, taking h equal to a constant multiple of $(n^{-1} \log n)^{1/4}$, the uniform rate of convergence to $\mathcal{C}$ is $O\{(n^{-1} \log n)^{1/2}\}$.

3.5 Outline Proof of Theorem 3.1

We employ the definitions of $\mathcal{I}_\pm$, $N_\pm$, $\bar{\epsilon}_\pm$ and $\mu_\pm$ as at (2.4.1), using the support radius of the kernel, h, as indicated in the paragraph containing (3.2.1). In addition, put

$\nu_\pm(x,\theta) = E\{N_\pm(x,\theta)\}$ and

$$S_\pm(x,\theta) = N_\pm(x,\theta) - \nu_\pm(x,\theta)\,, \quad T_\pm(x,\theta) = \sum_{i\in\mathcal{I}_\pm} g(X_i)\,K_i(x) - \mu_\pm(x,\theta) \tag{3.5.1}$$

and $\gamma_\pm(x,\theta) = \nu_\pm(x,\theta)^{-1}\,E\{\mu_\pm(x,\theta)\}$. Let $\delta = \delta(n) \to 0$ in such a manner that $h^2 \le \delta \le (1-\eta)h$ for some $\eta > 0$. (This is the version of the "search cone condition" from Section 2.2 that is appropriate in the present context.) Then for all sufficiently large $B_1 > 0$, for some $B_2, B_3, B_4 > 0$, and with $R_1(x,\theta)$ denoting any one of $S_\pm(x,\theta)/(nh^2)$, $T_\pm(x,\theta)/(nh^2)$ or $\bar\epsilon_\pm(x,\theta)$, we have, using Bernstein's inequality similarly as in the proof of Theorem 2.1,

$$\sup_{x,\theta,x',\theta'} P\Big[|R_1(x,\theta) - R_1(x',\theta')| > B_1\big\{\delta\,(nh^3)^{-1}\log n\big\}^{1/2}\Big] = O\big(n^{-B_1B_2}\big)\,, \tag{3.5.2}$$

$$\sup_{x,\theta} P\Big[|R_1(x,\theta)| > B_1\big\{(nh^2)^{-1}\log n\big\}^{1/2}\Big] = O\big(n^{-B_1B_2}\big)\,, \tag{3.5.3}$$

$$\sup_{x,\theta} P\big\{B_3nh^2 < N_\pm(x,\theta) < B_4nh^2\big\} = 1 - O\big(n^{-B_1}\big)\,, \tag{3.5.4}$$

where the suprema are taken over all unit vectors θ, θ' and vectors x, x' such that x is in an open neighbourhood of $\mathcal{C}$ and $\|x - x'\| \le \delta$, $\|\theta - \theta'\| \le \delta/h$. (Note that $(nh\delta)^{1/2}/(nh^2) = \{\delta(nh^3)^{-1}\}^{1/2}$.) Hence, defining

$$\lambda(x,\theta) = \frac{\nu_+(x,\theta)\,\nu_-(x,\theta)}{\nu(x,\theta)^2}\,\big\{\gamma_+(x,\theta) - \gamma_-(x,\theta)\big\}^2\,,$$

we have

$$\ell(x,\theta) - \ell(x',\theta') = \lambda(x,\theta) - \lambda(x',\theta') + R_2(x,\theta,x',\theta')\,,$$

where, in view of (3.5.2), (3.5.3) and (3.5.4),

$$\sup_{x,\theta,x',\theta'} P\Big[|R_2(x,\theta,x',\theta')| > B_1\big\{\delta\,(nh^3)^{-1}\log n\big\}^{1/2}\Big] = O\big(n^{-B_1B_2}\big)\,, \tag{3.5.5}$$

and the supremum is interpreted in the same manner as that at (3.5.2).

From (3.5.5) and the Borel-Cantelli lemma, and noting the Hölder continuity of K, we conclude that if $B_5 > 0$ is sufficiently large then with probability 1, for all sufficiently large n, $|R_2(x,\theta,x',\theta')| \le B_5\{\delta\,(nh^3)^{-1}\log n\}^{1/2}$ uniformly in the values of (x,θ,x',θ') over which the supremum at (3.5.2) is taken. Arguments based on the calculus now show that by maximising $\ell(x,\theta)$ over x and θ (for all values x within a neighbourhood of radius no more than $(1-\eta)h$ of $\mathcal{C}$, and for all unit vectors θ), the obtained approximations to

points x on $\mathcal{C}$, and angles made by normals θ to $\mathcal{C}$ at those points, are accurate to within $\pm B_6\{(nh)^{-1}\log n + h^2\}$ and $\pm B_6\{(nh^2)^{-1}\log n + h\}$, respectively.

To appreciate why, let (x', θ') denote respectively the point on $\mathcal{C}$ nearest to x and the unit vector in the direction of the tangent to $\mathcal{C}$ at x'. Note that if x is distant δ from x', and if θ is distant δ/h from θ', then $\lambda(x, \theta)$ is distant δ/h from $\lambda(x', \theta')$. In view of (3.5.5), this is detected (above stochastic error and bias terms) for all sufficiently large n, with probability 1, through reduction in the value of $\ell(x, \theta)$ relative to that of $\ell(x', \theta')$, if δ/h is larger than a sufficiently large constant multiple of $\{\delta\,(nh^3)^{-1}\log n\}^{1/2} + h$. (The term in h arises through departure of $\mathcal{C}$ from a straight line, due to curvature. Note that the function λ does not have two bounded derivatives at $\mathcal{C}$, since the fault discontinuity in g is bounded away from 0.) Equivalently, the reduction in the value of $\ell(x, \theta)$ is detected if δ is larger than a sufficiently large constant multiple of $(nh)^{-1}\log n + h^2$. The case $\delta > (1 - \eta)h$ may be treated using a subsidiary argument. This completes the outline proof of Theorem 3.1 in the context of fault lines in response surfaces. The case of fault lines in intensity or density surfaces may be treated similarly.

Chapter 4

Likelihood-Based Confidence Bands

4.1 Introduction

In Chapter 3 we introduced a class of fault line estimators based on local likelihoods, and proved that, like the estimator from Chapter 2, they are nearly optimal in the minimax sense. Here we derive explicitly, under appropriate regularity conditions, the asymptotic behaviour of the perpendicular distance between the true fault line and its local-likelihood estimator as design intensity diverges. The analysis of asymptotic behaviour consists of two main parts, which are given as Theorems 4.1 and 4.2: the bias, which is introduced through the kernel smoothing, and the joint limiting distribution of the stochastic component of the perpendicular distance and slope. That distribution will be seen to coincide with the location of the maximum of a spatially-indexed Gaussian process with quadratic drift, introduced earlier in Section A.4.2. Critical values, which determine the width of the confidence bands, have to be approximated numerically. However, once scale, geometric parameters and noise variance have been estimated, asymptotic confidence bounds for fault lines in regression surfaces can be computed without the need to simulate the quantiles anew. Moreover, it turns out that in the density and Poisson intensity surface cases, the same limit distribution of stochastic error applies (see Theorem 4.4).

Practical benefits of the explicit expressions which we derive for the bias, and the distribution of the perpendicular distance of the fault line estimator at a given point, are immediate. First, a bias adjustment can be made so to produce an improved estima-

tor. Secondly, and of at least equal importance, the results enable us to establish an asymptotic confidence band for the true fault line, taking account of bias.

Section 4.2 describes the model assumptions and the local-likelihood method. As the exposition largely parallels that of Section 3.2, we keep rather brief. Details of the technical arguments, in the case of Theorem 4.2 preceded by an intuitive outline of the proof, are given in Section 4.3.

4.2 Methodology and Main Theorems

As in Chapter 3, we develop the methodology around the case of fault lines in response surfaces, the density and Poisson intensity surface cases being similar but less complex. We continue to work with the general assumptions on the model as stated in Section 1.2, using the notation from around equation (3.2.1). For the purposes of the present chapter, however, we do not standardise the likelihood by dividing through an additional factor $N(x)$. Hence the definition now reads

$$\ell(x,\theta) = \frac{N_+(x,\theta)\, N_-(x,\theta)}{N(x)} \left\{ \bar{Y}_+(x,\theta) - \bar{Y}_-(x,\theta) \right\}^2 . \tag{4.2.1}$$

We estimate $\mathcal{C}$ as a ridge line, $\widehat{\mathcal{C}}$ say, of $m(x)$, in the same way as in Section 3.2. The comments made there on spurious ridge points apply here as well. We shall assume that the ridge points have been searched out in a sufficiently narrow open neighbourhood of $\mathcal{C}$, so that these issues do not arise.

Next we give regularity conditions for an expansion of the dominant terms in bias of the estimator $\widehat{\mathcal{C}}$ of $\mathcal{C}$. As in Chapters 2 and 3, let K be a radially symmetric, bivariate probability density, with a uniformly bounded derivative. The radius M of $\mathrm{supp}(K) = \mathrm{ball}(0, M)$ may now be any number in $\mathbb{R}_+^*$. Our assumption of radial symmetry is not essential to obtain the convergence rates evinced by our theorems. Anisotropic choices would be advantageous in some situations; for example, if the function defining the response surface was known to vanish on one side of $\mathcal{C}$, a one-sided kernel would yield smaller multiplying constants in the (otherwise unchanged) convergence rates. Generally however, the expense in additional complexity lessens the attractiveness of this alternative. As in Chapter 3, smoothness of K could principally be dispensed with, at the expense of the continuity of ℓ.

While the theory to be developed covers both cases of Poisson-distributed and gridded design, we shall generally assume that $\mathcal{X} = \{X_i\}$ represents a (homogeneous) Poisson

process. Adjustments for the case when the design points are located on a deterministic lattice are addressed in Remark 4.4. Versions of our results may be formulated in the case where $\mathcal{X}$ represents an inhomogeneous Poisson process, or in fact a much wider class of point processes which satisfy the requirements of the infill asymptotic. The conditions that are then to be imposed are straightforward, but to avoid being encumbered by added technicalities and notation we shall only consider the case where the local probability density function is constant.

Assume further that the origin $O = (0,0) \in \mathcal{C}$, and that in a neighbourhood of O the curve is the locus of points $x = (x^{(1)}, x^{(2)})$, where $x^{(2)} = f(x^{(1)})$. In accordance with the general assumptions from Section 1.2, we ask that the function f have two continuous derivatives in a neighbourhood of 0 and satisfies $f(0) = f'(0) = 0$, $f''(0) = 2a$, for $a \in \mathbb{R}$. Then, the gradient of $\mathcal{C}$ at O may be represented by $\theta_* = (1,0)$, which corresponds to the first principal component direction e_1 in the paragraph preceding Definition A.1. The direction of θ is determined by the rule given presently. In a neighbourhood of O the two parts into which $\mathcal{C}$ divides the plane will be denoted by $\mathcal{R}_+$ and $\mathcal{R}_-$. In contrast to Chapters 2 and 3, we now require a parity in that the '+' and '−' subscripts correspond in that order to the locally-convex and concave sides of $\mathcal{C}$. If $\mathcal{C}$ is given by a straight line at O, or if f has a point of inflection at 0, then we allow any assignment of $\mathcal{R}_\pm$. Suppose too that the regression mean g is given by

$$g(x) = g_+(x)\, I(x \in \mathcal{R}_+) + g_-(x)\, I(x \in \mathcal{R}_-)\,, \tag{4.2.2}$$

where g_+ and g_- are real-valued functions defined in the plane, with bounded, continuous derivatives in an open neighbourhood of O; and that, similar to condition (2.3.3), the limit of $|g_+(x) - g_-(x)|$, as x converges to a point on $\mathcal{C}$, is bounded away from zero for those parts of $\mathcal{C}$ that lie in a neighbourhood of O. Call these conditions (C_1).

On occasion we shall strengthen the smoothness assumptions in (C_1) by asking that in a neighbourhood of O, $\mathcal{C}$ be given by $x^{(2)} = f(x^{(1)})$, where f has three bounded derivatives in a neighbourhood of 0; and that g_+ and g_- have two bounded derivatives in a neighbourhood of O. Call these conditions (C_2).

Under conditions (C_1), and with probability 1, $\ell(x,\theta)/n$ has a proper, well-defined limit as $n \to \infty$, for h held fixed. Call this limit $\ell_0(x,\theta)$, and let (x_0, θ_0) denote the value of (x,θ) that maximises $\ell_0(x,\theta)$ subject to $x^{(1)} = 0$ and $\theta \cdot \theta_* \geq 0$. (The latter constraint serves to specify the orientation of θ_0, but may also be seen as a convention which

determines the direction in which $\mathcal{C}$ is traced.) Let $t = t_0 \in (0, M)$ be the solution of

$$\int_0^t K(u,0)\,du = \int_t^M K(u,0)\,du\,, \tag{4.2.3}$$

which, under conditions (C_1), is uniquely defined. Note that, due to the assumption of rotational invariance, $K(x) = K^{(0)}(\|x\|)$ is a function of $\|x\|$ only (where $\|\cdot\|$ denotes the Euclidean norm), and that $K^{(0)}(u)$ may replace the integrands in (4.2.3).

Theorem 4.1. *If conditions* (C_1) *hold then as* $h \to 0$, $x_0 = (0, h^2\xi_0 + o(h^2))$ *where* $\xi_0 = at_0^2$, *and* $\|\theta_0 - \theta_*\| = o(h)$. *If both* (C_1) *and* (C_2) *hold then* $x_0 = (0, h^2\xi_0 + O(h^3))$ *and* $\|\theta_0 - \theta_*\| = O(h^2)$.

We may view Theorem 4.1 as an assertion about the size of the bias of $\widehat{\mathcal{C}}$. It states that if $\mathcal{C}$ is smooth then the 'non-stochastic part' of $\widehat{\mathcal{C}}$, interpreted as the limit of $\widehat{\mathcal{C}}$ as $n \to \infty$ for fixed h, lies perpendicularly distant $at_0^2h^2$, up to terms of order h^3, from any given point P on $\mathcal{C}$, where $2a$ denotes the curvature of $\mathcal{C}$ at P and t_0 is an absolute positive constant. It may be verified that this estimator is essentially the same as the one mentioned in the proof of Theorem 2.4 (see p. 45). Indeed, and as was done in Hall and Rau (2002), it is straightforward to adjust that proof to the present situation, and hence to obtain the results below (where $h = h(n) \to 0$) by means of conditional expectations instead of the strong law for large numbers. For later purposes, let $\mathcal{C}_h$ denote the curve obtained by moving the appropriate distance, representing bias, away from each point P on $\mathcal{C}$; for example, moving a distance $x_0^{(2)}$ in the vertical direction when P is the origin, with the tangent to $\mathcal{C}$ equal to the $x^{(1)}$ axis.

In Theorem 4.2 we shall see that to a first approximation, the stochastic component of the perpendicular distance from $\mathcal{C}$ to $\widehat{\mathcal{C}}$ has zero mean, and in fact has a symmetric distribution. A first-order asymptotic approximation to the distance from $\mathcal{C}$ to $\widehat{\mathcal{C}}$, measured perpendicularly from $\mathcal{C}$, is thus given by the bias term suggested by Theorem 4.1 plus the error-about-the-mean component described in Theorem 4.2.

It should be appreciated that there is a degree of ambiguity in how one defines the pointwise distance between $\mathcal{C}$ and $\widehat{\mathcal{C}}$. Our approach, expressed in terms of the perpendicular distance (relative to $\mathcal{C}$) from the nearest point on $\widehat{\mathcal{C}}$ to a given point on $\mathcal{C}$, has the closest connections with the directed Hausdorff distance $\vec{d}_H(\mathcal{C}, \widehat{\mathcal{C}})$, defined at (A.8.1). This notion of distance is both conceptually and theoretically simpler than the perpendicular distance (relative to $\widehat{\mathcal{C}}$) from the nearest point on $\mathcal{C}$ to a given point on $\widehat{\mathcal{C}}$. The difference consists in the generally differing tangent orientations for the respective points on $\mathcal{C}$ and $\widehat{\mathcal{C}}$. In order to operationalise the second notion of distance, and in consequence

also the Hausdorff distance d_H, a tangent estimator for $\widehat{\mathcal{C}}$ is needed, which is supplied by the S^* component appearing in Theorem 4.2; see Remark 4.6. In view of these difficulties, the Fréchet distance $d_{\mathcal{F}}$, though suffering from drawbacks in numerical respects (see Section A.8.1), seems an attractive alternative.

Next we state regularity conditions for Theorem 4.2. Recall that the r.v. s $\epsilon_i = Y_i - g(X_i)$ are supposed to be (jointly) independent of $\mathcal{X}$, and conditional on $\mathcal{X}$, have all moments finite, zero mean and variance $\sigma^2 > 0$. Suppose too that $a \neq 0$ (that is, $\mathcal{C}$ has nonzero curvature at O) and for some $\epsilon > 0$ and all sufficiently large n, $n^{-(1/3)+\epsilon} \leq h \leq n^{-(1/6)-\epsilon}$. Call these conditions (C_3).

Under conditions (C_1)–(C_3) there exists a closed disc centred at the origin such that $\mathcal{C}$ is locally convex or concave there. Let $\hat{x} = (0, \hat{x}^{(2)})$ denote the point in this disc at which $m(x)$ is maximised, subject to $x^{(1)} = 0$. Let $W(s,t)$ denote the polygonal Gaussian process from Subsection A.6.2. Define further

$$(S^*, T^*) = \text{argmax}_{s,t} \left\{ W(s,t) - \left(s^2 + t^2\right) \right\}, \tag{4.2.4}$$

and let $d = g_+(0) - g_-(0)$ and $\tau^2 = a^2 \rho\, t_0^{-2} t_1$, where $\rho = \rho(d, \sigma) = 1 + 4(\text{SNR})^2$, where the *signal-to-noise ratio* (SNR) is defined as

$$\text{SNR} = \frac{|d|}{\sigma}, \tag{4.2.5}$$

and

$$t_1 = \frac{2\, t_0^2\, M^3}{K(t_0, 0)^2} \int_0^M K(u, 0)^2 \, du\,. \tag{4.2.6}$$

Notice that t_1 is invariant under a rescaling of K; that is, t_1 is unaffected if K is replaced by $K_\zeta(\cdot) \equiv \zeta^2 K(\zeta \cdot)$ for any $\zeta > 0$. Table 4.1 displays the values of t_0 and t_1 for some commonly-used kernels supported on the interval $[-1, 1]$ (that is, these are the kernels denoted $K^{(0)}$ in the paragraph preceding the statement of Theorem 4.1). In the case of the Epanechnikov kernel, t_0 is the solution of the equation

$$-\frac{4}{5}t^5 + \frac{8}{3}t^3 - 4t + \frac{16}{15} = 0$$

in the interval $(0, 1)$.

Theorem 4.2. *Assume conditions* (C_1)–(C_3)*, and let* ξ_0 *be as in Theorem 4.1. Then* $\hat{x}^{(2)} = \xi_0\, h^2 + \tau^{2/3}\, T_n\, h n^{-1/3}$*, where the random variable* T_n *converges in distribution to* T^* *as* $n \to \infty$*.*

Kernel	Definition	t_0	t_1
Epanechnikov kernel	$f_1(u) \propto (1-u^2)$	0.3473	0.1664
Biweight (quartic) kernel	$f_2(u) \propto (1-u^2)^2$	0.2811	0.0893
Triweight kernel	$f_3(u) \propto (1-u^2)^3$	0.2423	0.0576

Table 4.1: Values of t_0 and t_1 for commonly-used kernel functions.

By utilising parts of the proof in Section 4.3, specifically with the exclusion of the results related to Lemma 4.2, we obtain a version of Theorem 4.2 for a 'locally-constant' estimator. With that expression we mean that the likelihood ℓ is maximised over x only, with the slope being fixed at the true value (equal to zero). The family of r.v.s $\{T_c^{*,0}\}$, indexed by $c \in \mathbb{R}_+^*$, was introduced before the statement of Theorem A.3.

Theorem 4.3. *Assume the same conditions as in Theorem 4.2, and let $\hat{x}_0^{(2)}$ denote the previously defined profiled maximiser. Then $\hat{x}_0^{(2)} = \xi_0\, h^2 + \tau^{2/3}\, T_n'\, hn^{-1/3}$, where the random variable T_n' converges in distribution to $T_{1/2}^{*,0}$ as $n \to \infty$.*

As an immediate consequence of Theorems 4.2 and A.8, we see that the estimator T_n is capable of achieving 'near' optimality in the minimax sense. Because the estimator T_n' uses extra knowledge on the geometry of the edge, it does not fit in the framework of Subsection A.8.2.

In the subsequent remarks we briefly discuss some aspects and implications of the result of Theorem 4.2, which will transfer to the situation of Theorem 4.3 in an obvious way where applicable.

Remark 4.1. *Well-definedness of T^*.* It follows directly from Lemma A.1 that the process $W(s,t)$ is indeed continuous. Theorem A.2 then establishes that T^* is well defined.

Remark 4.2. *Pointwise confidence bands.* Theorem 4.2 suggests the following procedure for constructing a conservative, pointwise confidence region for either $\mathcal{C}$ or its deterministic, biased approximation $\mathcal{C}_h$. Let a_0 denote an upper bound to the absolute value of a; compute consistent estimates $\hat{d}$ and $\hat{\sigma}^2$ of d and σ^2, respectively, and thereby obtain an estimate $\hat{\rho} = \rho(\hat{d}, \hat{\sigma})$ of $\rho(d, \sigma)$; and, from the Monte Carlo results in Section 5.3, obtain a value $t(\alpha)$ such that $P\{|T^*| \le t(\alpha)\} = 1 - \alpha$. Put

$$r_1 = \big(a_0^2\, t_1\, \hat{\rho}\big)^{1/3}\, t(\alpha)\, hn^{-1/3} \quad \text{and} \quad r_2 = a_0\, t_0^2\, h^2 + r_1\,. \tag{4.2.7}$$

Then, the band $\mathcal{B}_1$ [respectively, $\mathcal{B}_2$] formed as the 'tube'

$$\widehat{\mathcal{C}}_h^{\,r_1} = \widehat{\mathcal{C}}_h \,\oplus\, \mathrm{ball}\big(0, r_1\big) \qquad \big[\widehat{\mathcal{C}}^{\,r_2} = \widehat{\mathcal{C}} \,\oplus\, \mathrm{ball}\big(0, r_2\big)\big] \tag{4.2.8}$$

is an asymptotically conservative, pointwise, nominal $(1-\alpha)$-level confidence band for $\mathcal{C}_h$ [for $\mathcal{C}$], in the sense that any given point on $\mathcal{C}_h$ [on $\mathcal{C}$] which is not at an end of the curve is contained within the band with probability at least $1-\alpha-\epsilon$, for any $\epsilon > 0$ and all sufficiently large n.

Indeed, suppose P is a point on $\mathcal{C}$. Then by Theorem 4.2, the probability that there is a point of $\widehat{\mathcal{C}}$ within distance r_2 of P, and along the normal to $\mathcal{C}$ at P, is at least $1-\alpha-\epsilon$ for any given $\epsilon > 0$ and for all sufficiently large n. The point P certainly lies in $\mathcal{B}_2$. In this argument it is not necessary for P to remain fixed as n increases, as long as it is moved around on $\mathcal{C}$ in a manner that does not depend on the data, and as long as its location is bounded away from any ends of $\mathcal{C}$. When applying the argument to $\mathcal{C}_h$, this 'moving P' interpretation is important as the point P is then traced along the trajectory which is unambiguously defined by its endpoint obtained as the limit as $h \to 0$. We may allow a_0, h, $\hat{d}$ and $\hat{\sigma}$ to vary with location on the curve without appreciably affecting the argument. It should be noted that the estimator $\mathcal{C}_h$ is essentially the one already introduced in Section 2.4 (see p. 45) and is more of theoretical rather than practical relevance.

An estimator $\hat{d}$ of the local jump height d is directly obtainable from the algorithm that produces our estimator $\widehat{\mathcal{C}}$. Indeed, $\bar{Y}_+(\hat{x},\hat{\theta}) - \bar{Y}_-(\hat{x},\hat{\theta})$ estimates $g_+(x) - g_-(x)$ at a point x on $\mathcal{C}$ that is near to $\hat{x}$ on $\widehat{\mathcal{C}}$; there is a degree of ambiguity in the sign of that difference, which is a consequence of $\hat{\theta}$ being specified only modulo π, but this does not affect ρ. The disadvantage of this approach in the case of high curvature, as well as a method for remedying this effect, are as discussed in Remark 2.1.

Remark 4.3. *Simultaneous confidence bands.* The results on confidence bands given in Remark 4.2 may be extended to allow for construction of conservative simultaneous confidence bands for either $\mathcal{C}$ or $\mathcal{C}_h$. With the same notation as there, put

$$r_{1,\mathrm{sim}} = \left(\tfrac{2}{3}\log n\right)^{1/2} r_1 \quad \text{and} \quad r_{2,\mathrm{sim}} = a_0\, t_0^2\, h^2 + r_{1,\mathrm{sim}}\,. \tag{4.2.9}$$

Then, the band $\mathcal{B}_{1,\mathrm{sim}}$ [respectively, $\mathcal{B}_{2,\mathrm{sim}}$] formed as the tube produced by the union of the set of all discs centred on $\widehat{\mathcal{C}}$ and with radius $r_{1,\mathrm{sim}}$ [radius $r_{2,\mathrm{sim}}$], is an asymptotically conservative nominal $(1-\alpha)$-level confidence band for $\mathcal{C}_h$ [for $\mathcal{C}$], minus its endpoints. Call this property (P). For an outline of the argument, note first that the proof of Theorem 4.1 in fact establishes uniform $O(h^2)$ convergence of $\mathcal{C}_h$ to $\mathcal{C}$. (To bound the supremum distance between these two curves, we may use arguments similar to those employed to derive (P).) In order to verify the assertion about $r_{1,\mathrm{sim}}$ in (4.2.9), split $\widehat{\mathcal{C}}$ into small segments, each of approximate length $2h$. Due to its locally-linear nature,

the properties of $\widehat{\mathcal{C}}$ are controlled, with probability tending to 1 as $n \to \infty$, by a linearly increasing collection of independent copies of T^*. As in the case of Chernoff's distribution, the distribution of T^* has lighter tails than those of a standard Normal variate. Hence, we conclude from Galambos (1978, Theorem 4.4.4, p. 230) or Embrechts, Klüppelberg and Mikosch (1997, Example 3.5.4, pp. 174ff) that the ratio of the maximum of, say, q independent copies of T^* and $(2 \log q)^{1/2}$ tends a.s. to 1. In our argument, $2hq \approx$ (length of $\mathcal{C}$). In view of conditions (C_3) and result (4.2.7), property (P) follows. It should be remarked that the rate of convergence here is rather slow, although the finite-sample approximation is not so poor that it is impractical; see Section 5.3. This is analogous to the well-known property of sequences of independent and Normally distributed r.v.s; see Hall (1979) and Leadbetter, Lindgren and Rootzén (1983, pp. 39–40).

In view of this and the remarks preceding the statement of Theorem 4.2, it seems to be more natural to adopt a global error measure. Assuming only for the purposes of this paragraph that $\mathcal{C}$ (and hence also $\widehat{\mathcal{C}}$) is a closed curve, and that $\mathcal{C}$ and $\widehat{\mathcal{C}}$ enclose the compact sets $\mathcal{G}$ and $\widehat{\mathcal{G}} = \widehat{\mathcal{G}}_n$ respectively, it may be deduced by methods used in the proof of Theorem 4.1 in Subsection 4.3.1 that

$$d_1(\mathcal{G}, \widehat{\mathcal{G}}_n) = O\Big(n^{-(2/3)+\epsilon}\Big)\,,$$

where the semimetric d_1 was defined in Subsection A.8, and $\epsilon > 0$ may be made arbitrarily small by the choice of h. (The comments made in Remark 4.5 below on further improvement of the convergence rate apply also in the present context.) Edge effects that occur in adapting the definition to non-closed curves, pose few problems. The metric d_1 seems to be the best suited one in the description of the global properties of $\widehat{\mathcal{C}}$. The task of bounding the implied constant in (4.3) requires knowledge about the overall geometry of $\mathcal{C}$, which poses the greatest problem in construction of confidence regions. Nevertheless, this metric will be shown in Section 5.3 to be practical in assessing performance of the estimator.

Remark 4.4. *Design points on a lattice.* The case where the points X_i are on a regular lattice (triangular, square or hexagonal), and so are nonrandom, is similar. The only necessary changes are to replace the condition $n^{-(1/3)+\epsilon} \le h \le n^{-(1/6)-\epsilon}$ in (C_3) by $Cn^{(-1/4)+\epsilon} \le h \le n^{-(1/6)-\epsilon}$, for $C > 0$ depending on the edge width of the lattice, and replace $\rho(d, \sigma)$ by $4\sigma^2/d^2$. That is, we drop the component '1' from ρ; it is contributed by the stochastic variability of points in the Poisson process, and so is not relevant in the case of a fixed grid. Taking $h \sim n^{-(1/4)+\epsilon}$ shows that the radius of the confidence band is within a polynomial factor of the minimax-optimal order. This can be compared with

Theorems 3.1–3.3 in Gayraud and Tsybakov (2002) who considered a testing problem in a related 'macroscopic' setting; minimax rates of testing were defined in Ingster (1993). The relations between our work and that of Gayraud and Tsybakov (2002) have been delineated earlier in this thesis (see Section 1.3, p. 27).

Remark 4.5. *Different orders of bandwidth.* For bandwidths h of order $n^{-1/6}$ or larger, it remains true that $\hat{x}^{(2)} - x_0^{(2)} = \tau^{2/3}\, T_n\, h n^{-1/3}$ where $T_n \Longrightarrow T^*$. However, an expansion of $x_0^{(2)}$ that is valid up to terms of order $hn^{-1/3}$ now includes terms of order at least h^3. Therefore, the simple manner in which Theorem 4.2 is expressed, leading to elementary confidence bands for $\mathcal{C}$, is no longer valid.

Sufficient regularity conditions for both a Taylor approximation to $x_0^{(2)}$, of the form $x_0^{(2)} = \xi_0\, h^2 + \ldots + \xi_{r-2}\, h^r + O(h^{r+1})$, and for the limit theorem $\hat{x}^{(2)} = \xi_0\, h^2 + \ldots + \xi_{r-2}\, h^r + \tau^{2/3}\, T_n\, h n^{-1/3}$ where $T_n \Longrightarrow T^*$, are (C_1), (C_3) with the restriction $h \le n^{-(1/6)-\epsilon}$ replaced by $h \le n^{-1/(3r)-\epsilon}$ for an integer $r \ge 2$, and in place of (C_2): f has $r+1$ bounded derivatives in a neighbourhood of 0, and g_+ and g_- have r bounded derivatives in a neighbourhood of O. Furthermore, we ask that in either case, K have a uniformly bounded derivative of order $(r-1)$.

If $h \sim n^{-1/3}$ then the limiting distribution of $\hat{x}^{(2)} - x_0^{(2)}$ is not a Gaussian functional, and depends intimately on the error distribution. If the condition imposed on that distribution in (C_3) is strengthened merely to the extent that a finite moment generating function in the neighbourhood of 0 is required, bandwidths of strictly larger order than $n^{-1/3}\,(\log n)^{2/3}$ are admissible, and thus a convergence rate within a logarithmic factor of optimal is achievable. To appreciate why, note that under these altered circumstances, exponential bounds may replace Markov-inequality bounds in the proof in Subsection 4.3.2. This enables estimation of various intermediate quantities up to remainders of logarithmic (instead of polynomial) size in n. Hence, deterioration of the least-squares fit in (2.2.2) can be detected with smaller point counts in the local kernel disc of interest.

Remark 4.6. *The role of S^*.* From the proof of Theorem 4.2, it can be seen that under the same conditions as stated there, the variation of the estimated angle θ_0 around the true value zero satisfies $\arg(\theta_0) = O(h^2) + \tau^{2/3}\, S_n\, n^{-1/3}$, where the random variable S_n converges in distribution to S^* as $n \to \infty$. Apart from the assertion of near-minimax optimality, this result may be used in practice to monitor the behaviour of the estimator. For example, by varying the bandwidth h at a given location and recomputing S_n as a function of h, evidence for the locally-linear character of $\mathcal{C}$ may be gathered. The case of zero curvature is addressed in more detail in Remark 4.8.

Remark 4.7. *Correlated errors.* In practice it is often more realistic to model the errors as dependent random variables. The purpose of this subsection is to show that if the errors instead exhibit a 'mild' or *short-range* dependency (relative to the bandwidth), the results obtained previously may be retained with only the asymptotic variance having to be corrected by a scaling factor. Theoretical framework for this situation has been laid down in work by Opsomer (1995, 1997) and Opsomer, Yang and Wang (1999), in continuation of work by other authors (e.g. Altman (1990, 1993)) in the one-dimensional (time-series) context. That framework has aimed at generalisation of time-series methods, both with regard to the domain of the covariates (which is spatial) and the non-gridded character of the design data.

Below we state conditions and the adjustments of the results obtained by Opsomer (1995) and Opsomer, Yang and Wang (1999) to cater for our setup. These papers consider locally-linear regression estimators rather than kernel estimators. It seems, however, that this restriction was solely motivated by the known asymptotic inadmissibility of kernel-based estimators in the case of random design; see Wand and Jones (1996).

Conditions on the correlation structure:

The error variables have the representation $\{\epsilon_i = \epsilon(X_i),\, i \in I\}$, where $\{\epsilon(z),\, z \in \Pi\}$ is a stationary second-order process with mean zero and covariance function

$$E(\epsilon(x)\epsilon(y)) = \sigma^2 \psi(\|x - y\|)\,. \tag{4.2.10}$$

Of the sequence of correlation functions $\psi = \psi_n$, we ask that it have the following properties, which essentially agree with the properties (P.1) and (P.2) from Opsomer (1995) and Opsomer, Yang and Wang (1999).

(D.1) ψ_n is differentiable, $n^{-1/3}h \int |\psi_n(t-x)|\,dt = O(1)$ and $n^{-1/3}h \int |\psi_n(t-x)|^2\,dt = O(1)$;

(D.2) There exists $\bar{\omega} > 0$ such that

$$\int |\psi_n(t)|\, I\{\|t| > \bar{\omega}h\}\, dt = o\left(\int |\psi_n|(t)\, dt \right).$$

Under conditions (D.1) and (D.2), the results in the previously cited papers continue to hold, with a correction factor multiplied by σ^2 in Theorem 4.1, in analogy to the factor of Opsomer (1999, Theorem 2.1). An admissible, and often plausible, choice for the correlation function is $\psi_n(t) = \{1 + (n^{1/2}t)^\alpha\}^{-1}$, where $\alpha > 1$ denotes a (globally or

locally estimated) parameter. Estimation of α poses a difficult problem in practice, on which further research is needed.

If the exploratory data are gridded, dependence structures which are, in a sense, cross-products of ordinary ARMA(p, q) time series models have been studied by various authors; see Etchison, Patula and Brownie (1994) and the references cited therein. Note that these models do not encompass dependence structures as at (4.2.10).

Remark 4.8. *Case of zero curvature.* The case $a = 0$, which is not covered by Theorem 4.2, leads to a different and non-symmetric distribution type when smoothness assumptions are strengthened. Under the old assumptions, consistency of $\hat{x}^{(2)}$ still holds.

The practical relevance of this case stems from the situation where curvature vanishes on a whole subinterval (rather than at a single point) of the parameter interval indexing $\mathcal{C}$. In the typical situation when no knowledge about $\mathcal{C}$ is available *a priori*, the method presented in this paper should be seen as sequential to a 'macroscopic,' or low-level, procedure for locating straight segments of $\mathcal{C}$. Marcel and Cattoen (1997) have described a method on which a construction of a pilot estimator of the fault line, gathering evidence of possible straight line segments, may be based. In Section 5.4 we shall suggest another macroscopic approach that also deals with corner points, that is, discontinuities in the second derivative. Note that in many circumstances where curved and linear parts of $\mathcal{C}$ join, corner points arise at the ends of the linear parts, as is the case with the example function in Section 5.4. For any separated straight line segment, and through use of its slope, confidence bands may be obtained by using Theorem 4.3. In order for this approach to be theoretically valid, the slope of the linear part of the edge must have an estimation error no larger then $O(h)$. Also, the bands will usually not connect with the bands yielded by the procedure in the vicinity of the curved parts of $\mathcal{C}$.

If curvature vanishes at isolated points on $\mathcal{C}$ only, good practical performance of confidence bands can be expected without any alteration to the procedure, as evinced by the simulation study in Section 5.3.

Remark 4.9. *Case of corners.* The case where $\mathcal{C}$ exhibits corners may be addressed by using a low-level method as in Remark 4.8, or by resorting to the devices mentioned in Remark 2.9. While a version of Theorem 4.1, modified so to accommodate kinks, may be readily devised, the limiting distribution will differ from that in Theorem 4.2 mainly insofar as the drift term subtracted from the pertaining Gaussian process is no longer quadratic.

In the remaining part of this section we address the cases of edges in density and Poisson intensity surfaces. In the density setting, and as in Chapter 3, we assume that i.i.d. data

$\mathcal{X} = \{X_i, 1 \le i \le n\}$ are sampled from a distribution with density g, where n is nonrandom, and g is smooth except for a fault line running through its support. The relevant kernel-weighted log-likelihood ratio is defined as

$$\ell^{(2)}(x,\theta) = \Big\{ N_+(x,\theta)\, \log N_+(x,\theta) + N_-(x,\theta)\, \log N_-(x,\theta) \\ -N(x)\, \log N(x) \Big\}\, nh^2\, N(x)^{-1}\,. \tag{4.2.11}$$

This differs from the likelihood proposed at (3.3.1) (see also the conventions there) only by the normalisation factor nh^2, which is included only to stabilise the variance of $\ell^{(2)}$, in a similar fashion as for 4.2.1; it is of course irrelevant for finding the maximiser of $\ell^{(2)}$. The fault line within the density surface defined by φ, or a part thereof, is then estimated as a ridge in the likelihood surface defined in analogy to the regression setting.

As a second variation of the regression problem, consider the situation when a realisation of an (inhomogeneous) planar Poisson process $\mathcal{X} = \mathcal{X}_n$ with intensity ng, where the function g is bounded away from zero and infinity throughout Π; that g is represented in the interior of Π by two sufficiently smooth functions g_+, g_- corresponding to the respective sides of a fault line. We also assume that $|g_+ - g_-|$ is bounded from below by a positive constant; and that the fault line has regularity properties as specified in the regression setting. To estimate (a section of) this line, we may again use the statistic at (4.2.11). The effective sample size, that is $N_n = \mathcal{X}_n(\Pi)$, clearly satisfies the condition stated on p. 19.

Theorem 4.4 below is an analogue of Theorems 4.1–4.2 for the case of a fault line in a density or intensity surface. Let (C_1') and (C_2') denote conditions (C_1) and (C_2), with in addition the assumption that g is bounded away from 0 in an open neighbourhood of the fault line. Define $\tilde{t}_0 \in (0, M)$ as the solution of (4.2.3), where an extra multiplier of $\varsigma \equiv g_+(0)/g_-(0)$ is taken on the left-hand side, and note that the bias decreases (increases) in the respective cases $\pm \log \varsigma > 0$. The quantities $\tilde{t}_1$ and $\widetilde{\tau}_1$ are defined as at (4.2.6), but with $\tilde{t}_0$ now replacing t_0. Take $\rho = 1$ in the definitions preceding Theorem 4.2 and put $n_1 = \frac{1}{2} n\{g_+(0) + g_-(0)\}$. When presented with an intensity surface, we would adopt entirely similar definitions.

Theorem 4.4. *Assume the previous notation, and consider either of the density or intensity surface problems described above.*
(i) *If conditions* (C_1') *hold then as* $h \to 0$, $x_0 = (0, h^2\xi_0 + o(h^2))$ *where* $\xi_0 = a\tilde{t}_0^2$, *and* $\|\theta_0 - \theta_*\| = o(h)$. *If both* (C_1') *and* (C_2') *holds then* $x_0 = (0, h^2\xi_0 + O(h^3))$ *and* $\|\theta_0 - \theta_*\| = O(h^2)$.
(ii) *Assume conditions* (C_1')–(C_2') *and* (C_3). *Then* $\hat{x}^{(2)} = \xi_0\, h^2 + \widetilde{\tau}_1^{2/3}\, T_n\, h n_1^{-1/3}$, *where*

the random variable T_n converges in distribution to T^* as $n \to \infty$.

4.3 Proofs

4.3.1 Proof of Theorem 4.1

Let $d = g_+(0) - g_-(0)$, and for $x = (x^{(1)}, x^{(2)})$ and $0 \le t < M$, define $g_{\pm,i}(x) = (\partial/\partial x^{(i)})\, g_\pm(x)$, $\kappa = \int_{x^{(2)}>0} x^{(2)}\, K(x)\, dx$,

$$\kappa_1(t) = \int_t^M \bigl(u^2 - t^2\bigr)\, K(u,0)\, du\,, \quad \kappa_2(t) = \int_0^t \bigl(t^2 - u^2\bigr)\, K(u,0)\, du\,. \tag{4.3.1}$$

Recall the definitions around (3.2.1), (3.5.1) and those referenced there, and in addition put

$$G_\pm(x,\theta) = \sum_{i \in \mathcal{I}_\pm(x,\theta)} g(X_i)\, K_i(x)\,,$$

and $\nu = EN(x) = nh^2$ (not depending on x). As in the proof of Theorem 2.1 in Section 2.4 (see also Hall and Rau, 2002), it may be shown that maximisation of the original likelihood ℓ is asymptotically equivalent to finding the value $(x,\theta) = (x_0,\theta_0)$ that maximises $\psi(x,\theta) \equiv |E\{G_+(x,\theta) - G_-(x,\theta)\}|$ subject to $x^{(1)} = 0$ and $\theta_0 \cdot \theta_* \ge 0$, where we recall that $\theta_* = (1,0)$.

We need several properties of $\psi(x,\theta)$, which are summarised in the next lemma. In this discussion it is sufficient to confine attention to vectors θ such that $\theta \cdot \theta_* \ge 0$. Assume for the time being that the curvature, $2a$, of $\mathcal{C}$ at O is nonzero, and let x and θ depend on h.

Lemma 4.1. *Let $\psi(x,\theta)$ be defined as above.*
(i) *If $x^{(1)} = 0$ and $x^{(2)} = O(h^2)$, the value of $\psi(x,\theta_*)$ is strictly smaller, for all sufficiently small h, if $x^{(2)}$ has the same sign as $2a$, the curvature, than it is if the sign is opposite.*
(ii) *If $x^{(1)} = 0$ and $x^{(2)}$ is of strictly larger order than h^2 then $\psi(x,\theta)$ is strictly larger, for all sufficiently small h, than $\psi(0,\theta_*)$.*
(iii) *If $x^{(1)} = 0$, if $x^{(2)} = O(h^2)$, and if $\|\theta - \theta_*\| \ge Ch$ (where $C > 0$), for all sufficiently small h, then there exists a bounded function $\xi = \xi(h)$, with the same sign as a, such that $\psi((0,\xi h^2),\theta_*)$ is strictly greater than $\psi(x,\theta)$ for all sufficiently small h.*

Proof of Lemma 4.1. Part (i) is clear from geometric considerations. In order to prove (ii) and (iii), we may use the calculations in Sections 4.3.2 and 4.3.3, specifically for $EG_\pm(x,\theta)$

and $\chi(x,\theta)$ in the notation introduced there. (In the case of $\chi(x,\theta)$, we dispense with the assumption $|\beta| \le n^{-\epsilon}$.) In either case, it is easy to verify the asserted reduction of the value of $\psi(x,\theta)$, which is the coefficient of the leading term of size $O(nh^3)$. □

Combining properties (i)–(iii) of Lemma 4.1, and using a subsequence argument, we deduce that, provided $a \neq 0$, $\theta_0 = \theta_* + o(h)$ and, for a bounded function ξ of h having the same sign as a, $x_0 = (0, h^2\xi)$. The case where $a = 0$ is straightforward; there it is clear from considerations of symmetry that $x_0 = (0, o(h^2))$ and $\theta_0 = \theta_* + o(h)$. To derive the first part of the theorem it remains to find the asymptotic value of ξ in the case $a \neq 0$.

To this end we shall prove that if $x = (0, h^2\xi)$ and $\theta = \theta_* + o(h)$ then

$$E\{G_+(x,\theta) - G_-(x,\theta)\} = nh^2\Big[\tfrac{1}{2}\,d + h\kappa\,\{g_{+,2}(0) - g_{-,2}(0)\} - 2\,h|a|d\,\{\kappa_1(w) + \kappa_2(w)\} + o(h)\Big]\,, \tag{4.3.2}$$

where $w = w(\xi) = \min\{(\xi/a)^{1/2}, M\}$. It follows that $|E\{G_+(x,\theta) - G_-(x,\theta)\}|$ equals $-2nh^3|ad|\,\{\kappa_1(w) + \kappa_2(w)\}$ plus terms, with a positive sum, that either do not depend on ξ or are of smaller order than h^3 as $h \to 0$. Therefore, $\xi = \xi_0 + o(1)$, where ξ_0 is the quantity that minimises $\kappa_1(w_0) + \kappa_2(w_0)$ when $w_0 = w(\xi_0)$. Elementary calculus shows that $\kappa_1(t) + \kappa_2(t)$ is minimised at $t = t_0$, and so $\xi_0 = at_0^2$. This gives the first part of the theorem.

To derive the second part of the theorem, note that under the stronger conditions there the arguments two paragraphs above imply that $\theta_0 = \theta_* + O(h^2)$; cf. Lemma 4.2 below. Furthermore,

$$\kappa_1(w) + \kappa_2(w) = \kappa_1(w_0) + \kappa_2(w_0) + b_1\,(w - w_0)^2 + o\big(|w - w_0|^2\big)\,,$$

where $b_1 = 2t_0K(t_0, 0)$. We may now derive the following form of (4.3.2): writing $x_0 = (0, h^2\xi_0 + h^2\eta)$ where $\eta = \eta(h) \to 0$, we have, provided $\theta = \theta_* + O(h^2)$,

$$E\{G_+(x,\theta) - G_-(x,\theta)\} = nh^2\Big[\tfrac{1}{2}\,d + h\,Q_1(h) - 2\,h|a|d\,\Big\{\kappa_1(w_0) + \kappa_2(w_0) + b_2\,\eta^2 + o\big(\eta^2\big)\Big\} + O\big(h^3\big)\Big]\,, \tag{4.3.3}$$

where $b_2 = (2a^2t_0)^{-1}K(t_0, 0)$ and $Q_1(h)$ is bounded and does not depend on η. It follows that $\eta = O(h)$, which implies that $x_0 = (0, h^2\xi_0 + O(h^3))$.

Finally we prove (4.3.2). Without loss of generality, $a > 0$. Then for any bounded

function $\xi = \xi(h)$ we have, with $x = (0, \xi h^2)$ and $\theta = \theta_* + o(h)$,

$$\begin{aligned}
E\{G_+(x,\theta)\} &= nh^2 \int_{y^{(2)}>0} g_+(x+hy)\, K(y)\, dy \\
&\qquad + 2nh^2 \int_w^M dy^{(1)} \int_0^{h\{a(y^{(1)})^2-\xi\}} \big\{g_-(x+hy) \\
&\qquad - g_+(x+hy)\big\}\, K(y)\, dy^{(2)} + o\big(nh^3\big) \\
&= \tfrac{1}{2}\, nh^2\, g_+(0) + nh^3\, g_{+,2}(0)\, \kappa - 2nh^3 ad\, \kappa_1(w) + o\big(nh^3\big)\,, \\
E\{G_-(x,\theta)\} &= nh^2 \int_{y^{(2)}>0} g_-(x+hy)\, K(y)\, dy \\
&\qquad + 2nh^2 \int_0^w dy^{(1)} \int_{h\{a(y^{(1)})^2-\xi\}}^0 \big\{g_+(x+hy) \\
&\qquad - g_-(x+hy)\big\}\, K(y)\, dy^{(2)} + o\big(nh^3\big) \\
&= \tfrac{1}{2}\, nh^2\, g_-(0) + nh^3\, g_{-,2}(0)\, \kappa + 2nh^3 ad\, \kappa_2(w) + o\big(nh^3\big)\,,
\end{aligned}$$

implying (4.3.2).

4.3.2 Proof of Theorem 4.2

As we did for Theorem 2.1, we shall start with an intuitive outline of the proof before giving the stringent arguments.

Intuitive Outline of Proof

Under the conditions of Theorem 4.2 we may, by Theorem 4.1, write $x_0 = (0, x_0^{(2)}) = (0, h^2\xi_0 + O(h^3))$, where ξ_0 is as in Theorem 2.1; and $\theta_0 = (\cos\omega_0, \sin\omega_0)$, where $\omega_0 = O(h^2)$. Consider a semidisc of radius h with its centre at the point $x_1 = (0, h^2\xi_0 + \alpha t)$, where $\alpha = hn^{-1/3}$ and $-\infty < t < \infty$. Assume that the diameter, $\mathcal{D}$, of the semidisc makes an angle $\omega_1 = (\alpha/h)s$ to the horizontal, where $-\infty < s < \infty$. Note that α is of strictly larger order than h^3, and so we have shifted the value of $x^{(2)}$ by an amount αt that is of larger order than the remainder in our approximation to $x_0^{(2)}$.

Tilting the disc through α/h radians raises or lowers the ends of its diameter by the same amount, in order of magnitude terms, as moving the disc's centre vertically through distance α. This is the reason for the choosing the particular relative values of the amounts of translation and tilting.

It may be deduced from (4.3.3) that the effect on $\lambda(x,\theta)$, defined below at (4.3.5), of

shifting x a distance αt vertically away from x_0, as suggested by our definition of $x_1^{(2)}$, is to reduce its value by an amount $\nu h\beta^2\tau_1 t_0^{-1}t^2$, where $\nu = nh^2$, $\beta = \alpha/h^2 = (nh^3)^{-1/3}$ and $\tau_1 > 0$. Lemma 4.2 will show that the effect of tilting $\mathcal{D}$ through $(\alpha/h)s$ radians is to produce a further reduction, by $\nu h\beta^2\tau_1 t_0 s^2$, in $\lambda(x,\theta)$; and that the 'interaction' effect of doing both these things together is negligibly small. Therefore, up to terms that are either negligibly small or do not depend on s or t, translation of x_0 to x_1 and rotation of θ_0 to θ_1 changes λ from $\lambda(x_0,\theta_0)$ to $\lambda(x_1,\theta_1) = \lambda(x_0,\theta_0) - \nu h\beta^2\tau_1\,(t_0 s^2 + t_0^{-1}t^2)$.

This takes care of the deterministic effects of translating and rotating the diameter of the semidisc. It may be proved that to first order, the stochastic effects equal $\nu h\beta^2\tau_2\,\widehat{W}_1(s,t)$, where $\tau_2 > 0$ and $\widehat{W}_1$ and $\widehat{W}_2$ will denote bivariate stochastic processes that are asymptotically distributed as W. We chose the formula for β, i.e. $\beta = (nh^3)^{-1/3}$, so that the deterministic and stochastic terms are both of size $\nu h\beta^2$. Their net contribution is their sum,

$$\nu h\beta^2 \left\{\tau_2\,\widehat{W}_1(s,t) - \tau_1\left(t_0 s^2 + t_0^{-1}t^2\right)\right\}. \tag{4.3.4}$$

Note that in (4.3.4), the quantity t_0 appears once in the numerator and once in the denominator. This feature is discussed in Remark 4.10. To first order, $\ell(x_1,\theta_1)$ equals the quantity at (4.3.4) plus terms that do not depend on s or t. Changing scale appropriately, so that (s,t) transforms to (u,v), renders the quantity at (4.3.4) equal to a constant multiple of $\widehat{W}_2(u,v) - (u^2+v^2)$. It follows that the value of t at the maximum of $\ell(x_1,\theta_1)$, over s and t, equals a constant multiple of a random variable whose limiting distribution is that of T^*, defined in Section 4.2. Theorem 4.2 follows from this result.

Details of Proof

We now turn to the details of the proof of Theorem 4.2. First we state and prove a lemma, describing the effects of small rotations on $E\{G_+(x,\theta) - G_-(x,\theta)\}$.

Lemma 4.2. *Assume the conditions of Theorem 4.2, with $a > 0$. Let $x = (0,\xi h^2)$, where $\xi = \xi(h) \geq 0$ is bounded, and let $\theta = (\cos h\beta, \sin h\beta)$. Then,*

$$\begin{aligned}\chi(x,\theta) &\equiv E\{G_+(x,\theta) - G_-(x,\theta)\} - E\{G_+(x,\theta_*) - G_-(x,\theta_*)\}\\ &= -d\,K\left\{(\xi/a)^{1/2},0\right\}\left(\xi/a^3\right)^{1/2}\nu h\beta^2 + o\big(\nu h\beta^2\big) + O\big(\nu h^3\big)\,,\end{aligned}$$

uniformly in values of β satisfying $|\beta| \leq n^{-\epsilon}$ for any given $\epsilon > 0$.

From the lemma and (4.3.3) we conclude that if $x(t) = (0, h^2\xi_0 + h^2\beta t)$, $\theta(s) = (\cos(h\beta s),$

$\sin(h\beta s)$) and

$$\lambda(x,\theta) = \nu^{-1}\big\{EG_+(x,\theta) - EG_-(x,\theta)\big\}^2\,, \tag{4.3.5}$$

then

$$\begin{aligned}\lambda(x(t),\theta(s)) = -\nu h\beta^2\tau_1\left(t_0 s^2 + t_0^{-1}t^2\right) + o\big(\nu h\beta^2 s^2\big) + o\big(\nu h\beta^2 t^2\big) + O\big(\nu h^3\big)\\ + \text{ terms that do not depend on } s \text{ or } t\,,\end{aligned} \tag{4.3.6}$$

where $\tau_1 = |a|^{-1}d^2K(t_0,0)$. If we take $\beta = (nh^3)^{-1/3}$ then (4.3.6) holds uniformly in $|s|, |t| \le n^\epsilon$ for some $\epsilon > 0$.

Proof of Lemma 4.2. Put $c_\pm = g_\pm(0)$, in which notation $d = c_+ - c_-$. Write $x = (0, \xi h^2)$ where $\xi = \xi(h)$ is bounded. Without loss of generality, $\beta \ge 0$. For simplicity we shall consider only the case where $g_\pm \equiv c_\pm$ and $\mathcal{C}$ has exactly the equation $y^{(2)} = a(y^{(1)})^2$; in a somewhat more complicated manner than in Subsection 4.3.1, it may be shown that this induces an error term of size $O\big(\nu h^3\big)$. Put $u_1 = (\xi/a)^{1/2}$, and let $y^{(1)} = u_2 < 0, u_3 > 0$ denote the two solutions of the pair of equations $y^{(2)} = ha(y^{(1)})^2$ and $y^{(2)} = h(\xi + \beta y^{(1)})$. Then we have $u_2 < u_1 < u_3$, $|u_2| < |u_3|$ and

$$\begin{aligned}&(2\nu)^{-1}\chi(x,\theta)\\ &\quad= c_-\int_{-M}^{-u_1} dy^{(1)} \int_{h\beta y^{(1)}}^{0} K(y)\,dy^{(2)} + c_-\int_{-u_1}^{u_2} dy^{(1)} \int_{h\beta y^{(1)}}^{h\{a(y^{(1)})^2-\xi\}} K(y)\,dy^{(2)}\\ &\qquad+ c_+\int_{-u_1}^{u_2} dy^{(1)} \int_{h\{a(y^{(1)})^2-\xi\}}^{0} K(y)\,dy^{(2)} + c_+\int_{u_2}^{0} dy^{(1)} \int_{h\beta y^{(1)}}^{0} K(y)\,dy^{(2)}\\ &\qquad- c_+\int_{0}^{u_1} dy^{(1)} \int_{0}^{h\beta y^{(1)}} K(y)\,dy^{(2)} - c_-\int_{u_1}^{u_3} dy^{(1)} \int_{0}^{h\{a(y^{(1)})^2-\xi\}} K(y)\,dy^{(2)}\\ &\qquad- c_+\int_{u_1}^{u_3} dy^{(1)} \int_{h\{a(y^{(1)})^2-\xi\}}^{h\beta y^{(1)}} K(y)\,dy^{(2)} - c_-\int_{u_3}^{M} dy^{(1)} \int_{0}^{h\beta y^{(1)}} K(y)\,dy^{(2)}\\ &\qquad= \text{(I)}+\text{(II)}+\ldots+\text{(VIII)}\,.\end{aligned}$$

(The 2 in the factor $(2\nu)^{-1}$ on the left-hand side derives from the fact that rotation results in counts being simultaneously added to $E\{G_+(x,\theta) - G_-(x,\theta)\}$ and removed from $E\{G_+(x,\theta_*) - G_-(x,\theta_*)\}$, or vice versa; so each change in counts needs to be included twice.) In the inner integral of term (II), we perform a substitution $\tilde{y}^{(2)} \mapsto -y^{(2)}$, and combine terms (I),(VI) and (VIII). The terms (III),(IV) and (V) can also be combined. Hence, retaining the name of the variable $y^{(2)}$, it follows that $(2\nu)^{-1}\chi(x,\theta)$

equals $(c_- - c_+)\,A$, where

$$A = \int_{-u_2}^{u_1} dy^{(1)} \int_{-h\beta y^{(1)}}^{h\{a(y^{(1)})^2-\xi\}} K(y)\,dy^{(2)} + \int_{u_1}^{u_3} dy^{(1)} \int_{h\{a(y^{(1)})^2-\xi\}}^{h\beta y^{(1)}} K(y)\,dy^{(2)}\,. \quad (4.3.1)$$

Now, $u_1 = (\xi/a)^{1/2}$ and, by Taylor expansion, $-u_2 < u_3$ are given by $-u_2 = u_1 - (\beta/2a) + O(\beta^2)$ and $u_3 = u_1 + (\beta/2a) + O(\beta^2)$. Therefore, approximating the parabola $y^{(2)} = ha(y^{(1)})^2$ by a straight line in the vicinity of (u_1, hau_1^2), we see that

$$\begin{aligned} A &= \int_{u_1-(\beta/2a)}^{u_1} dy^{(1)} \int_{-h\beta y^{(1)}}^{2hau_1(y^{(1)}-u_1)} K(u_1,0)\,dy^{(2)} \\ &\qquad + \int_{u_1}^{u_1+(\beta/2a)} dy^{(1)} \int_{2hau_1(y^{(1)}-u_1)}^{h\beta y^{(1)}} K(u_1,0)\,dy^{(2)} + o\big(h\beta^2\big) \\ &= 2\,K(u_1,0) \int_0^{\beta/2a} dy^{(1)} \int_{2hau_1 y^{(1)}}^{h\beta u_1} dy^{(2)} + o\big(h\beta^2\big) \\ &= \tfrac{1}{2}\,K(u_1,0)\,h\beta^2 u_1 a^{-1} + o\big(h\beta^2\big)\,. \end{aligned}$$

This concludes the proof of the lemma. □

Remark 4.10. *The role of the kernel-related quantities.* As remarked in the paragraph containing (4.3.4), the quantity t_0 appears once in the numerator as well as in the denominator of that equation. This can be intuitively explained as follows. In the proofs of Theorem 4.1 and Lemma 4.2, and disregarding the kernel weights for the moment, we have approximated the modulus of continuity of the 'area function' $(s,t) \mapsto \|\mathcal{R}_+ \triangle \mathcal{A}(s,t)\|$, where $\mathcal{R}_+$ and $\mathcal{A}(\cdot,\cdot)$ were defined in the paragraphs containing (4.2.2) and at (A.6.3) respectively. It is not surprising that the order of growth of the coordinate-wise moduli of the area function are equal. However, the areas transferred from one semi-disc to the other as one moves away from the minimum of the area function are far from the centre of the kernel disc in the case of translating (that is, for varying t), but close to that centre in the case of tilting or changes in the s variable. Being a measure of concentration of the mass of K about the origin, the quantity t_0 naturally reflects these properties.

Now we return to the proof of Theorem 4.2, and take $\beta = (nh^3)^{-1/3}$ rather than the more general definition assumed in the lemma. Recall that $\epsilon_i = Y_i - g(X_i)$, and (redefining some of the notation from Chapters 2 and 3 for convenience) put $\mu_\pm(x,\theta) = E\{G_\pm(x,\theta)\}$,

$$S_\pm(x,\theta) = \sum_{i \in \mathcal{I}_\pm(x,\theta)} \epsilon_i\,K_i(x)\,, \quad (4.3.7)$$

$T_\pm(x,\theta) = G_\pm(x,\theta) - \mu_\pm(x,\theta)$, $U_\pm(x,\theta) = N_\pm(x,\theta) - \frac{1}{2}\,\nu$, $S = S_+ + S_-$, $T = T_+ + T_-$, $U = U_+ + U_- = N - \nu$ and $\gamma_\pm(x,\theta) = 2\nu^{-1}\mu_\pm(x,\theta)$. For all $B, \epsilon > 0$, and with R_1 denoting any one of $S_\pm/\nu$, $T_\pm/\nu$ and $U_\pm/\nu$, we may obtain, using Markov-inequality bounds,

$$\sup_{x,\theta} P\Big\{|R_1(x,\theta)| > \nu^{-1/2}\,n^\epsilon\Big\} = O\big(n^{-B}\big)\,, \tag{4.3.8}$$

where the supremum is taken over all unit vectors θ and vectors x such that x is in a sufficiently small open neighbourhood of $\mathcal{C}$.

In the formula for ℓ, express $N_\pm$ as $\frac{1}{2}\nu + U_\pm$ and $G_\pm$ as $\mu_\pm + T_\pm$, and then Taylor expand ℓ as a power series in $S_\pm$, $T_\pm$ and $U_\pm$, to obtain:

$$\ell(x,\theta) = \lambda(x,\theta) + L(x,\theta) + Q(x,\theta) + R_2(x,\theta)\,, \tag{4.3.9}$$

where λ, defined at (4.3.5), represents terms that do not involve $S_\pm$, $T_\pm$ and $U_\pm$,

$$L \equiv \tfrac{1}{4}\,U\,(\gamma_+ - \gamma_-)^2 + (\gamma_+ - \gamma_-)\,\big\{(S_+ + T_+ - \gamma_+\,U_+) - (S_- + T_- - \gamma_-\,U_-)\big\} \tag{4.3.10}$$

represents terms that are linear in $S_\pm$, $T_\pm$ and $U_\pm$, Q represents terms that are quadratic in those quantities, and R_2 represents terms that are cubic or of higher order.

In view of (4.3.8) we have, for all $B, \epsilon > 0$,

$$\sup_{x,\theta} P\Big\{|R_2(x,\theta)| > \nu^{-1/2}\,n^\epsilon\Big\} = O\big(n^{-B}\big)\,. \tag{4.3.11}$$

A similar argument shows that, if $\delta = h^2 n^\eta$ for $\eta > 0$ sufficiently small,

$$\sup_{x,\theta,x',\theta'} P\Big\{|Q(x,\theta) - Q(x',\theta')| > \big(\delta/nh^3\big)^{1/2}\,n^\epsilon\Big\} = O\big(n^{-B}\big)\,, \tag{4.3.12}$$

where the supremum is taken over all unit vectors θ, θ' and vectors x, x' such that x is in an open neighbourhood of $\mathcal{C}$ and $\|x - x'\| \le \delta$, $\|\theta - \theta'\| \le \delta/h$. (Note that $(nh\delta)^{1/2}/(nh^2) = (\delta/nh^3)^{1/2}$.)

Put

$$L_1 = \tfrac{1}{4}\,d^2\,U + d\big\{(S_+ + T_+ - c_+U_+) - (S_- + T_- - c_-U_-)\big\}\,.$$

Recall the definition of (x_0, θ_0) as the maximiser of the asymptotic likelihood, as given in the paragraph containing (4.2.3). Observe that $\gamma_+ - \gamma_-$ differs from d by $O(\delta/h)$, uniformly in x, θ such that $\|x - x_0\| \le \delta$ and $\|\theta - \theta_0\| \le \delta/h$, and that in the same sense, $\gamma_\pm$ differs from $c_\pm$ by $O(\delta/h)$, with $c_\pm$ as in the proof of Lemma 4.2. Therefore, using

Markov-inequality bounds we may show that, for $B > 0$ arbitrarily large,

$$\sup_{x,\theta}{}^{0} P\Big\{|L(x,\theta) - L_1(x,\theta)| > n^{(1/2)+\epsilon}\,\delta\Big\} = O\big(n^{-B}\big)\,, \tag{4.3.13}$$

where $\sup_{x,\theta}^{0}$ denotes the supremum over x, θ such that $\|x - x_0\| \le \delta$ and $\|\theta - \theta_0\| \le \delta/h$. Combining (4.3.9)–(4.3.13) we conclude that

$$\ell(x,\theta) = \lambda(x,\theta) + L_1(x,\theta) + R_3(x,\theta) + R_4(x_0,\theta_0)\,, \tag{4.3.14}$$

where $R_4(x_0,\theta_0)$ does not depend on (x,θ) and, since $\delta = h^2 n^{\eta}$ and $n^{-(1/3)+\epsilon} \le h \le n^{-(1/6)-\epsilon}$ for some $\epsilon, \eta > 0$, we have for $j = 3$ and $\eta = \eta(\epsilon) > 0$ sufficiently small,

$$\sup_{x,\theta}{}^{0}\ P\Big\{|R_j(x,\theta)| > n^{(1/2)+\epsilon}\,\delta\Big\} = O\big(n^{-B}\big)\,. \tag{4.3.15}$$

Defining $c = c_+ + c_-$,

$$L_2 = d\left\{2\,(S_+ + T_+) - c\,U_+\right\} \quad \text{and} \quad L_3 = \left(\frac{d}{4} + c_-\right) d\,U - d\,(S+T)\,,$$

we have

$$L_1 = L_2 + L_3\,. \tag{4.3.16}$$

Using the inequality $|K_i(x) - K_i(x_0)| \le C\|x - x_0\|/h$, where the constant C depends only on K, we may prove using Markov-inequality bounds that, with V denoting any of S, T and U, we have for all $B, \epsilon > 0$,

$$\sup_{x,\theta}{}^{0} P\Big\{|V(x,\theta) - V(x_0,\theta_0)| > n^{(1/2)+\epsilon}\,\delta\Big\} = O\big(n^{-B}\big)\,. \tag{4.3.17}$$

Therefore,

$$\sup_{x,\theta}{}^{0} P\Big\{|L_3(x,\theta) - L_3(x_0,\theta_0)| > n^{(1/2)+\epsilon}\,\delta\Big\} = O\big(n^{-B}\big)\,. \tag{4.3.18}$$

Put $\mathcal{J}_1(x,\theta) = \mathcal{I}_+(x,\theta)\backslash\mathcal{I}_+(x_0,\theta_0)$, $\mathcal{J}_2(x,\theta) = \mathcal{I}_+(x_0,\theta_0)\backslash\mathcal{I}_+(x,\theta)$, $\mathcal{J}_3(x,\theta) = \mathcal{I}_+(x,\theta) \cap \mathcal{I}_+(x_0,\theta_0)$,

$$\begin{aligned}
D_1(x,\theta) &= \sum_{i\in\mathcal{J}_1(x,\theta)} v_i\,K_i(x)\,, \quad D_2(x,\theta) = \sum_{i\in\mathcal{J}_1(x,\theta)} v_i\,K_i(x_0)\,,\\
D_3(x,\theta) &= \sum_{i\in\mathcal{J}_2(x,\theta)} v_i\,K_i(x_0)\,, \quad D_4(x,\theta) = \sum_{i\in\mathcal{J}_3(x,\theta)} v_i\,\Big\{K_i(x) - K_i(x_0)\Big\}\,,\\
\Delta_j &= D_j - E(D_j)\,,
\end{aligned}$$

where $v_i = \epsilon_i, g(X_i), 1$ in the cases $V = S, T, U$ respectively. Then,

$$V_+(x,\theta) - V_+(x_0,\theta_0) = \Delta_1(x,\theta) - \Delta_3(x,\theta) + \Delta_4(x,\theta)\,, \tag{4.3.19}$$

and by an argument similar to that leading to (4.3.17),

$$\begin{aligned} \sup_{x,\theta}{}^0\, P\Big\{|\Delta_1(x,\theta) - \Delta_2(x,\theta)| > n^{(1/2)+\epsilon}\,\delta\Big\} &= O\big(n^{-B}\big)\,, \\ \sup_{x,\theta}{}^0\, P\Big\{|\Delta_4(x,\theta)| > n^{(1/2)+\epsilon}\,\delta\Big\} &= O\big(n^{-B}\big)\,. \end{aligned} \tag{4.3.20}$$

Combining (4.3.1)–(4.3.20) we see that we may write

$$\ell(x,\theta) = \lambda(x,\theta) + L_2(x,\theta) + R_5(x,\theta) + R_6(x_0,\theta_0)\,,$$

where $R_6(x_0,\theta_0)$ does not depend on (x,θ), and $R_5(x,\theta)$ satisfies (4.3.15). Let $N_{+,\pm}(x,\theta)$ denote the sum of $K_i(x)$ over indices $i \in \mathcal{I}_+(x,\theta)$ such that $X_i \in \mathcal{R}_\pm$, with the $\pm$ signs taken respectively. Then, $T_+ \approx c_+N_{+,+} + c_-N_{+,-} - \mu_+$, and in fact (4.3.15) holds if we take R_j there equal to $T_+ - (c_+N_{+,+} + c_-N_{+,-} - \mu_+)$. Therefore we may write

$$\ell(x,\theta) = \lambda(x,\theta) + L_0(x,\theta) + R_7(x,\theta) + R_6(x_0,\theta_0)\,, \tag{4.3.21}$$

where R_7 satisfies (4.3.15) and

$$\begin{aligned} L_0 &= d\left\{2\left(c_+N_{+,+} + c_-N_{+,-} - \mu_+ + S_+\right) - cU_+\right\} \\ &= d\left\{d\left(N_{+,+} - N_{+,-}\right) + 2\left(S_+ - \mu_+\right) + \frac{c}{2}\,\nu\right\} \\ &= d\left\{d\left(N_{+,+} - N_{+,-} - E\big(N_{+,+} - N_{+,-}\big)\right) + 2\,S_+\right\}. \end{aligned} \tag{4.3.22}$$

Define $x(t) = (0, \xi_0h^2 + h^2\beta t)$ and $\theta(s) = (\cos\omega(s), \sin\omega(s))$, where $\omega = h\beta s$, $-\infty < s,t < \infty$ and $\beta = (nh^3)^{-1/3}$. Put $S[s,t] = S_+(x(t),\theta(s))$,

$$\tau_2^2 = 2d^2\left(d^2 + 4\sigma^2\right)M^3\int_0^M K(u,0)^2\,du \tag{4.3.23}$$

and

$$\begin{aligned} N[s,t] = {} & N_{+,+}(x(t),\theta(s)) - N_{+,-}(x(t),\theta(s)) \\ & - E\big\{N_{+,+}(x(t),\theta(s)) - N_{+,-}(x(t),\theta(s))\big\}\,. \end{aligned} \tag{4.3.24}$$

In conclusion we establish the following assertions, which will enable us to conclude the proof of Theorem 4.2: (a) the stochastic process $\widehat{W}_1$, defined by

$$\widehat{W}_1(s,t) \equiv \left(\nu h \beta^2\right)^{-1} \tau_2^{-1} d\left(d\,N[s,t] + 2\,S[s,t]\right), \tag{4.3.25}$$

converges weakly to a Gaussian process $W^{(K)}$ in the Skorokhod space of functions to which we referred in Subsection A.6.3; and (b) the 'argmax' functional, which defines the sequence of estimators of perpendicular distance and slope, transfers to the Gaussian limit in (a). To appreciate the implications of (a) for $\ell(x,\theta)$, note that by (4.3.6) and (4.3.21), and by using an almost sure representation on a complete probability space $(\Omega, \mathcal{E}, P')$ (the existence of which is ensured by Theorem A.1), there exists for all $\delta > 0$ a set $\mathcal{U} \in \mathcal{E}$ with $P'(\mathcal{U}) \geq 1 - \delta$, on which

$$\begin{aligned}\ell(x(t), \theta(s)) = \nu h \beta^2 \left\{\tau_2\, W^{(K)}(s,t) - \tau_1 \left(t_0 s^2 + t_0^{-1} t^2\right)\right\} + o\{\nu h \beta^2 (t_0 s^2 + t_0^{-1} t^2)\} \\ + O\left(\nu h^3\right) + \text{ terms that do not depend on } s \text{ or } t\,,\end{aligned} \tag{4.3.26}$$

uniformly in $(s,t) \in \mathcal{A}_{n,\epsilon}$, where

$$\mathcal{A}_{n,\epsilon} \equiv \{(u,v) : |u| \leq c_1 n^{\epsilon},\ |v| \leq c_2 n^{\epsilon}\} \tag{4.3.27}$$

for some $c_1, c_2 > 0$ and $\epsilon > 0$ sufficiently small.

First we prove (a). Within this part of the proof, it may be assumed that attention is restricted to a compact subset $\mathcal{S}_1 \subseteq \mathcal{S}$, where $\mathcal{S} \equiv \{|x^{(1)}| \leq M\}$. Let $\mathcal{F}_0$ denote the collection of subsets of $\mathbb{R}^2$ defined in the paragraph containing (A.6.1). By Lemma A.3, the class $\mathcal{F}_0$ is an indexing collection in the sense of Definition A.5. Lemma A.2 implies that the process W is continuous on $\mathcal{F}_0$ with respect to the metric d_1.

We now proceed to verify the conditions of Theorem A.5 and Remark A.1. Define the counting process

$$\widetilde{N}(A) = \sum_{i: X_i \in A} K_i(x)\,, \qquad A \in \mathcal{A}\,, A \subset \mathcal{S}_1\,, \tag{4.3.28}$$

let $\widetilde{N}[s,t] = \widetilde{N}(\mathcal{A}(s,t) \triangle \mathcal{A}(0,0))$, and define $\widetilde{S}$ in a similar way by summing in (4.3.28) over $\epsilon_i K_i$ instead of K_i. The process that is defined as $\widehat{W}_1$ at (4.3.25), with replacing N, S by $\widetilde{N}$ and $\widetilde{S}$ respectively, is a process somewhat similar to that studied in (IM, Section 9.2, pp. 181ff). It is the *continuity* property of the process which allows us, by Remark A.1, to establish weak convergence. Conditions 2 on 3 from Remark A.1 in Section A.6.3 may be verified by similar Markov-inequality arguments as were used to derive formulae (4.3.8) and (4.3.17), respectively, in conjunction with the Borel-Cantelli

lemma. The variance measure of this process is asymptotically equal to Lebesgue measure with kernel weights depending on the distance from the axis $x^{(1)} = 0$, as at (4.3.23). Using the conditional uniformity property of the Poisson process, it may further be deduced that the remaining conditions 1–4 of Theorem A.5 hold; the calculations are similar to those in (IM, pp. 181ff), and note that condition 2 is trivially satisfied. The limiting Gaussian process depends on K, with the variance measure having the previously mentioned kernel-weighted form. In view of the observations made above, we can apply the result stated in Remark A.1, to obtain the convergence in distribution on $\mathcal{S}_1$ of the process considered, and hence also of $\widehat{W}_1$, to W. Since $\mathcal{S}_1 \subseteq \mathcal{S}$ was arbitrary, the proof of (a) is complete.

Next we derive (b). Let (S_n, T_n) be any measurable sequence such that

$$\ell(S_n, T_n) \geq \sup_{(s,t)\in\mathcal{A}_{n,\epsilon}} \ell(x(t), \theta(s)) - o_p(1)\,,$$

where $\mathcal{A}_{n,\epsilon}$ was defined at (4.3.27). To obtain (b) we will apply Theorems A.6 and A.7 on the argmax functional. It follows from Theorem A.7 in Subsection A.7.2, using the statement from its last sentence, that with probability tending to 1, uniformly in n, as $B \to \infty$, the supremum of $\tau_2\widehat{W}_1(s,t) - \tau_1(t_0 s^2 + t_0^{-1}t^2)$ over $(s,t) \in \mathcal{A}_{n,\epsilon}$, for $\epsilon > 0$ sufficiently small, occurs for $(s,t) \in (-B, B)^2$, whence the sequence (S_n, T_n) is uniformly tight. A bound on the continuity modulus of $N[s,t]$ is readily available, and due to the moment condition on the errors imposed in (C_3), it may be shown that that bound also applies to the process $S[s,t]$. To obtain these results we use the fact that K has a bounded derivative. Hence, assertion (b) follows.

In the final step of the proof we separate out those quantities on the right-hand side of (4.3.26) which, via t_0, depend on K. We display this dependency explicitly, so that $t_0(L)$ denotes the value of t_0 calculated for a kernel L. First, rescale K by letting $K_{t_0}(x) = t_0^2\,K(t_0 x)$, so that the disc on which K_{t_0} is supported has radius $t_0^{-1}M$. Then it is readily seen that $t_0(K_{t_0}) = 1$, whence we may consider (4.3.26) with t_0 replaced by unity. In equation (4.3.21) we change variable from (s,t) to (u,v), where $s = \tau_3 u$, $t = \tau_3 v$ and $\tau_3 = (\tau_2/\tau_1)^{2/3} = \tau^{2/3}$, τ being as in the paragraph containing (4.2.4). Put $\widehat{W}_2(u,v) = \tau_3^{-1/2}\widehat{W}_1(\tau_3 u, \tau_3 v)$. By assertion (a), $\widehat{W}_2$ converges weakly, on bounded rectangles, to $W^{(K)}$. Furthermore, the previous assertions and (4.3.21) imply that

$$\begin{aligned}\operatorname{argmax}_{s,t}\big\{\ell(x(t),\theta(s))\big\} &\Longrightarrow \operatorname{argmax}_{s,t}\big\{\tau_2\,\widehat{W}_1(s,t) - \tau_1\left(s^2+t^2\right)\big\} \\ &= \operatorname{argmax}_{u,v}\big\{\tau_2\,\tau_3^{1/2}\,\widehat{W}_2(u,v) - \tau_1\,\tau_3^2\left(u^2+v^2\right)\big\}\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{argmax}_{u,v}\big\{\widehat{W}_2(u,v) - \big(u^2+v^2\big)\big\} \\
&\stackrel{\mathrm{d}}{=} \tau^{2/3}\big(S^*, T^*\big)\,.
\end{aligned}
$$

In view of the scaling of $x(t)$, this proves Theorem 4.2 if, in computing the supremum that defines $\hat{x} = (0, \hat{x}^{(2)})$, we restrict attention to values of $x = (0, x^{(2)})$ that satisfy $|x^{(2)}| \leq hn^{-(1/3)+\epsilon}$ for $\epsilon > 0$ sufficiently small, and also note that W may replace $W^{(K)}$ in the definition of the distribution of the maximiser. A subsidiary argument may be used to extend this result to the range $|x^{(2)}| \leq r$, where $r > 0$ is such that $\mathcal{C}$ is convex or concave throughout the disc of radius r centred at the origin. By analogous reasoning for $\theta(s)$, we may verify the statement in Remark 4.6.

4.3.3 Outline Proof of Theorem 4.4

The arguments employed in the present instance are similar to those in Sections 4.3.1 and 4.3.2, and hence we shall keep rather brief. By expanding the likelihood at (4.2.11), we obtain

$$
\ell^{(2)}(x,\theta) = \lambda^{(2)}(x,\theta) + L^{(2)}(x,\theta) + Q^{(2)}(x,\theta) + R_2^{(2)}(x,\theta)\,,
$$

where

$$
\begin{aligned}
\lambda^{(2)}(x,\theta) &= \Big\{\nu_+(x,\theta)\,\log\nu_+(x,\theta) + \nu_-(x,\theta)\,\log\nu_-(x,\theta) \\
&\qquad - \nu(x)\,\log\nu(x)\Big\}\, nh^2\,\nu(x)^{-1}\,, \\
L^{(2)}(x,\theta) &= \Big(nh^2 \log\left(\frac{\nu_+}{\nu_-}\right)(U_+\nu_- - U_-\nu_+)\Big)/\nu^2 \\
&= \Big(nh^2 \log\left(\frac{\nu_+}{\nu_-}\right)(N_+\nu_- - N_-\nu_+)\Big)/\nu^2\,, \\
\nu_\pm = EN_\pm\,, &\qquad U_\pm = N_\pm - EN_\pm\,,
\end{aligned}
\tag{4.3.29}
$$

$Q^{(2)}$ represents terms that are quadratic in $U_\pm$, and $R_2^{(2)}$ represents terms that are cubic or of higher order. In several instances on the right-hand-side of (4.3.29) and in some of the displays below, arguments (x,θ) will be suppressed for notational convenience. In order to prove the bias assertion (part (i) of Theorem 4.4), we rewrite $\ell^{(2)}$ as follows:

$$
\begin{aligned}
\ell^{(2)}(x,\theta) &= \frac{\nu_+}{\nu_+ + \nu_-}\log\left\{\frac{\nu_+}{\nu_+ + \nu_-}\right\} + \frac{\nu_-}{\nu_+ + \nu_-}\log\left\{\frac{\nu_-}{\nu_+ + \nu_-}\right\} \\
&= (\tfrac{1}{2} - r)\log(\tfrac{1}{2} - r) + (\tfrac{1}{2} + r)\log(\tfrac{1}{2} + r)
\end{aligned}
$$

$$= -\log 2 + 2\,r^2 + \frac{4}{3}\,r^4 + O(r^6)\,,$$

where $r = \frac{1}{2} - \nu_+/\big(\nu_+ + \nu_-\big)$. We therefore seek the value $\big(x_0^{(2)}, \theta_0^{(2)}\big)$ of (x, θ) that maximises

$$\left|\frac{1}{2} - \frac{\nu_+}{\nu_+ + \nu_-}\right| = \left|\frac{1}{2} - \frac{\nu_-}{\nu_+ + \nu_-}\right| = \frac{1}{2}\left|\frac{\nu_+ - \nu_-}{\nu_+ + \nu_-}\right|\,. \tag{4.3.30}$$

The treatment of the expression $|\nu_+ - \nu_-|$ on the right-hand side of (4.3.30) is the same as in Section 4.3.1. It can be shown that the weighting by expected effective sample size $\nu_+ + \nu_-$ leads to the incorporation of the factors $g_\pm(0)$ in the form asserted in part (i) of the theorem.

As for part (ii), the appearance of n_1 as an arithmetic mean of two asymptotic local sample sizes arises due to the asymptotically linear nature of $\mathcal{C}$ when considered at the scale of shifts that are considered in calculation of the stochastic error. This is similar to the proof of Theorem 4.2. A rigorous proof of this property is expected to proceed by further examination of (4.3.29), however we are unable to provide the full details here.

In order to deal with the case where $\mathcal{X}$ represents a Poisson process, we use the fact that conditional on the r.v. $\mathcal{X}(\Pi)$, the points in $\mathcal{X}$ are i.i.d. with common distribution Y, where

$$P(Y \in A) = \frac{\int\limits_A g(x)\,dx}{\int\limits_\Pi g(x)dx}\,,$$

whence the foregoing results yield the conclusion of the theorem.

Chapter 5

Numerical Properties

5.1 Introduction

This chapter is concerned with the finite-sample properties of the tracking estimator of Chapter 2, as well as the likelihood-based estimators of the edge and concomitant confidence bands of Chapters 3 and 4. These two topics are respectively addressed in Sections 5.2 and 5.3. Though the edge curves studied here have an appreciably higher degree of complexity than the one of the example function of Section 3.1, they are arguably still not versatile enough for many practical situations. This applies especially to *corner points* (discontinuities in the second derivative of the edge curve) and *straight line segments*. Both of these cases are not covered by Theorem 4.2. In order to address these issues, in Section 5.4 we augment the local-likelihood estimator by a 'macroscopic' component. This methodology is exemplified by discussion of an example curve with the above features.

As noted in Remark 4.8 (see Section 4.2), confidence bands for straight line segments may be obtained by using Theorem 4.3. This, however, presupposes that the linear character of the edge has been established where it is present, and slope has been estimated with sufficient accuracy. At least the first of these tasks, if not the second, poses appreciable problems. Remark 4.8 pointed to work by other authors, which integrated a step for identification of straight line segments. The procedure we shall develop in Section 5.4 does not present novel methodology in addition to the previous chapters, nor does it aim to extend the theoretical optimality properties proved for the estimators there. Instead, our goal is to demonstrate that the local-likelihood estimator has its place in more 'robust'

image processing schemes that use methodologies that are familiar from the computer vision literature. Specifically, we will use the Canny edge detector which, as noted in Section 1.3, has comparatively solid theoretical foundations among the non-statistical edge detectors. In the corner detection step we shall also draw on scale-space theory as described in Section A.2.

In a coarse-to-fine approach to detecting edges, and also for more complex image processing tasks, it is plausible for the local-likelihood estimator to be used in the final stage. Therefore, its potential to be integrated into other and relatively standard schemes seems large indeed. To give one more example, which will not be further developed here, consider the estimation of splitting or terminating fault lines (already addressed in Remark 2.9) by a 'divide and conquer' approach. Start by dividing the observation window Π into a number of subsets, whose diameter converges to zero in successive stages of the model; in the simplest case, these subsets would consist of an increasing collection of squares, or their parts which intersect the observation area. Within each subset in turn, check for evidence for a fault line by using a pre-determined threshold, as in the paragraph containing (3.2.2). In order to construct the estimator globally, rules for connecting neighbouring edges and discarding outliers are needed, similarly to those in the procedure of Section 5.4. It is this versatility in appending the tracking and local-likelihood estimators to other image processing modules that motivates our use of the term *edge detection architecture*.

For the purposes of Sections 5.2 and 5.3, we use the notation from Chapters 2 and 4 respectively, other than that introduced here.

5.2 Numerical Properties of Tracking Method

To explore numerical performance of the tracking method of Chapter 2 we investigated several levels of noise degradation of the regression surface defined at (5.2.1) below, taking $d = 1$ there. We used the explicit formula for the quantity $S(x, \theta)$ defined at (2.2.2); see (2.4.2). As in the context of Section 2.3, the index $j \in \mathbb{N}_0$ counts how many steps the tracking sequence has advanced, with $\hat{x}_0 = \widehat{Q}$ being the starting point. To minimise S as a function of θ, for fixed $x = \hat{x}_j$, we used an algorithm based on golden-section search, described e.g. in Press, Teukolsky, Vetterling and Flannery (1992). In the context of the paragraph containing (3.2.3), we chose the arc radius for the search angle around the previous tracking direction ϕ_0 as $\Delta\phi = \pi/10$.

We experimented with a number of variants of the methodology suggested in Chap-

ter 2, which approximated the continuum version of the estimator; see the discussion at the close of Section 2.2, and the paragraph containing (3.2.5). One was to take $x = \hat{x}_j$ in (2.2.2), and move an amount δ in the direction of the fitted value of θ, rather than move directly to a neighbouring point of the grid $\mathcal{G}$. (We took $\delta = 0.005$.) This amounted to using an *unsieved* M-estimator (cf. the comment in the second paragraph of Section 2.2). Unsieved M-estimators are known to perform well in many examples, their larger 'variance' (in a somewhat loose sense) notwithstanding (Van der Vaart and Wellner, 1996, p. 323).

As in the example in Section 3.1, in the finite-sample context it is important to employ the suggestion of Remark 2.3, thereby refitting location. We found that it was adequate, and saved time, to refit location at points distant δ and $\delta/2$ on either side of x on the line that passed through x and was perpendicular to the estimated orientation of the curve at x. In low-curvature parts of $\mathcal{C}$ the cycle length between successive location refittings could be varied within a wide range, from 1 to about 10 steps, without having any noticeable effect on performance. However, refitting at *each* step was noticeably beneficial in places of high curvature. A related matter is that of estimating θ, which represents a derivative and so is inherently more prone to error than an estimate of $\mathcal{C}$. Nevertheless, these difficulties could be overcome in part by using a smaller value of δ in places where $\mathcal{C}$ has higher curvature. In the computer vision context, and because δ and h should at least approximately have a constant ratio (see the paragraph containing (3.2.4)), this issue is seen to be connected with the scale-space of an image, defined in Section A.2 (see p. 142). It might even be appropriate to use cycle length and δ as spatially-varying tuning parameters. In principle, curvature could be estimated from values of $\widehat{\mathcal{C}}$ computed in earlier steps, and used to select these quantities at the current location. This is the approach taken in Section 5.3.

For definiteness the curves were drawn using an algorithm that refitted location at each step. Also, to smooth out sharp turns to some extent we derived $\hat{\theta}$ from a moving average of the present step and two previous ones, with

$$\hat{\theta}_{j+1} = w_1 \hat{\theta}_j + w_2 \hat{\theta}_{j-1} + w_3 \hat{\theta}_{j-2}$$

and the weight vector $(w_1, w_2, w_3) = (0.6, 0.3, 0.1)$.

Inspired by similar ideas previously employed by Qiu (1997, 2002a), we smoothed over an ellipse with half-axes $h_{\min}$ and h_{maj}, rather than a disc, with its shorter axis $h_{\min}$ in the direction of motion. (The efficacy of this idea becomes much more apparent in the case of fault lines with knots; see Hall, Qiu and Rau (2002).) This meant using the kernel

$K_{i,\theta}^{\text{asymm}}(x)$ (where 'asymm' stands for 'asymmetric') instead of $K_i(x)$, defined by:

$$K_{i,\theta}^{\text{asymm}}(x) = K\big(U_i^{(1)}/h_{\min}, U_i^{(2)}/h_{\text{maj}}\big)\,,$$

with the respective (normalised) projections on the tangent vector $\theta = \theta(x)$ and its normal $\theta^{\perp}$,

$$\begin{aligned} U_i^{(1)} &= \big(X_i^{(1)} - x^{(1)}\big)\cos\theta + \big(X_i^{(2)} - x^{(2)}\big)\sin\theta\,, \\ U_i^{(2)} &= -\big(X_i^{(1)} - x^{(1)}\big)\sin\theta + \big(X_i^{(2)} - x^{(2)}\big)\cos\theta\,. \end{aligned}$$

In our experiments we took $(h_{\min}, h_{\text{maj}}) = (0.08, 0.12)$.

For the experiments in this and the following Section 5.3, we shall consider the family $\big\{g_d(\cdot) : d > 0\big\}$ of example functions, conveniently described in polar coordinates:

$$\begin{aligned} g_d(r,\phi) &= d \cdot I\big\{f(\phi) \le r\big\}\,, \\ f(\phi) &= \big[\exp\{\cos(3\phi)\} + \sin(\phi/2)\big]/8\,, \quad 0 \le \phi \le 2\pi\,. \end{aligned} \tag{5.2.1}$$

Hence the parameter d represents the (constant) jump height of the fault line in the regression surface. For the example in this section, we will limit ourselves to the case $d = 1$, deferring the case of varying d to Section 5.3. This example illustrates how the algorithm copes with 'meandering' functions that do not admit simple representations in Cartesian coordinates. Specifically, the fault line described by f exhibits some challenging geometric properties, since absolute curvature reaches a value of 33.95 at angles (in radians) $\phi_1 = 0.302$ and $\phi_2 = 2\pi - \phi_1 = 5.981$ respectively, corresponding to the points $(0.239, \pm 0.075)$; further peaks are located at $\phi_3 = 1.791$, $\phi_4 = 2.398$, $\phi_5 = 2\pi - \phi_4 = 3.885$, and $\phi_6 = 2\pi - \phi_3 = 4.493$. In contrast, curvature vanishes at points on $\mathcal{C}$ located at angles $\phi_7 = 0.574$, $\phi_8 = 1.342$, $\phi_9 = 2.662$, $\phi_{10} = 2\pi - \phi_8 = 4.941$ and $\phi_{11} = 2\pi - \phi_7 = 5.709$. With the notation from the paragraph preceding the statement of Theorem 4.1, we used the biweight kernel, $K^{(0)}(t) = (15/16)\,(1 - t^2)^2$ for $|t| \le 1$, so that

$$K(x) = K\big(x^{(1)}, x^{(2)}\big) = \frac{3}{\pi}\big(1 - x^{(1)} - x^{(2)}\big)^2\,, \quad \|x\| \le 1\,,$$

and $\epsilon_i \sim \text{N}(0, \sigma^2)$, where $\sigma = 0.5$ or 0.75. Design points were either on a regular 330×330 lattice within the square $[-0.55, 0.55] \times [-0.55, 0.55]$, or in the form of a homogeneous Poisson process with the same expected number of points per unit area. An illustration of the data appeared in Figure 1.2 (see p. 13).

Our first objective was to estimate $\mathcal{C}$ over its full length (note that length$(\mathcal{C}) = 4.272$).

For that purpose, and for definiteness, we chose the single starting point $\widehat{Q} = (0.2, 0.08)$ (which is located close to $\mathcal{C}$) and the approximate tangent direction $\theta(0) = \pi$. If $\widehat{Q}$ is selected by eye, and if the design points form a regular lattice, then one or more of the design points may lie on the line that divides the kernel domain into two half-discs, causing ambiguity. While in theory this causes only negligible problems, in practice the possibility of these ties should be recognised. Note, however, that this problem only affects the starting point. The possibility to move principally in any direction between the vertices of the grid comprising the design, which we construe here as pixels, implies what is often called *sub-pixel accuracy*. In order to overcome the ambiguity problem with the starting point, the method of *dithering* can be used, which means a random jiggling of the points along the disc diameter. For a discussion of sub-pixel accuracy and dithering, see for example Huertas and Medioni (1986), Liu and Ehrich (1995), and the references cited therein.

Panels (a)–(d) of Figure 5.1 are to be read in pairs, the left-hand side showing the case of a lattice design and the right-hand side, the case of Poisson-distributed design. Each pair of panels corresponds to a different value of σ. Six typical realisations are superimposed in each panel, and show that performance deteriorates only slightly from $\sigma = 0.5$ to $\sigma = 0.75$. From the displays, the conclusion can be drawn that even those parts of $\mathcal{C}$ where absolute curvature briefly exceeds an absolute value of 30, cause few problems. However, we found that erratic behaviour of the estimates begins to occur regularly for $\sigma = 0.75$ if the previously mentioned auxiliary devices were not employed. Moreover, we found generally that higher noise levels, say $\sigma = 1.0$, most often led to unsatisfactory results in the non-straight parts of $\mathcal{C}$. It would be possible, however, to improve performance here by further increasing the intensity of the point process.

Next we investigated the sensitivity of the algorithm with respect to the starting point $\widehat{Q}$. We continued to use the previous regression surface, focusing now on the higher noise level $\sigma = 0.75$ and the case of Poisson-distributed design. Figure 5.2 displays our simulation results. Panels (a)–(d) depict behaviour of three typical estimates in the cases $\widehat{Q} = (q + 0.08\,k, 0)$, with $q = \mathrm{e}/8 \approx 0.34$ and $k \in \{-2, \ldots, 1\}$. In this part of our example we smoothed over discs (instead of ellipses) with radius $h = 0.08$, and for illustration we have added a disc of that radius, centred at $\widehat{Q}$, to our displays. Also, on this occasion we tracked the fault line on both sides of the starting point. The results indicate that the algorithm is sensitive to choice of $\widehat{Q}$, although it readily tracks the fault line once the latter has been found. In practice it is a good idea to experiment with different, neighbouring starting points.

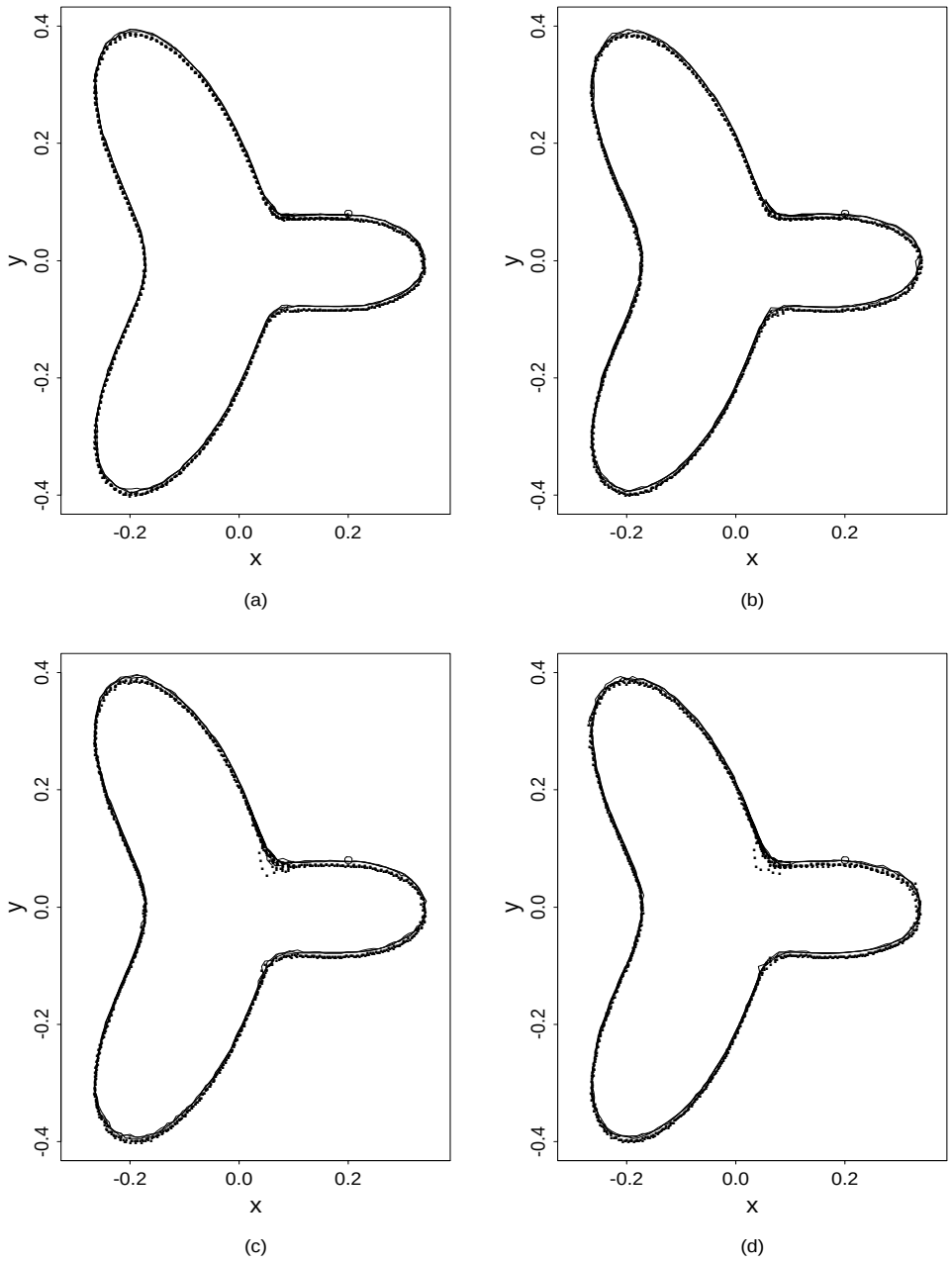

Figure 5.1: Performance of the algorithm is illustrated for Normal $N(0, \sigma^2)$ noise, with $\sigma = 0.5$ in panels (a) and (b), and $\sigma = 0.75$ in panels (c) and (d). Panels (a) and (c) depict the case of gridded design, while panels (b) and (d) are for the case of Poisson-distributed design. The starting point $\bar{Q} = (0.2, 0.08)$ is encircled.

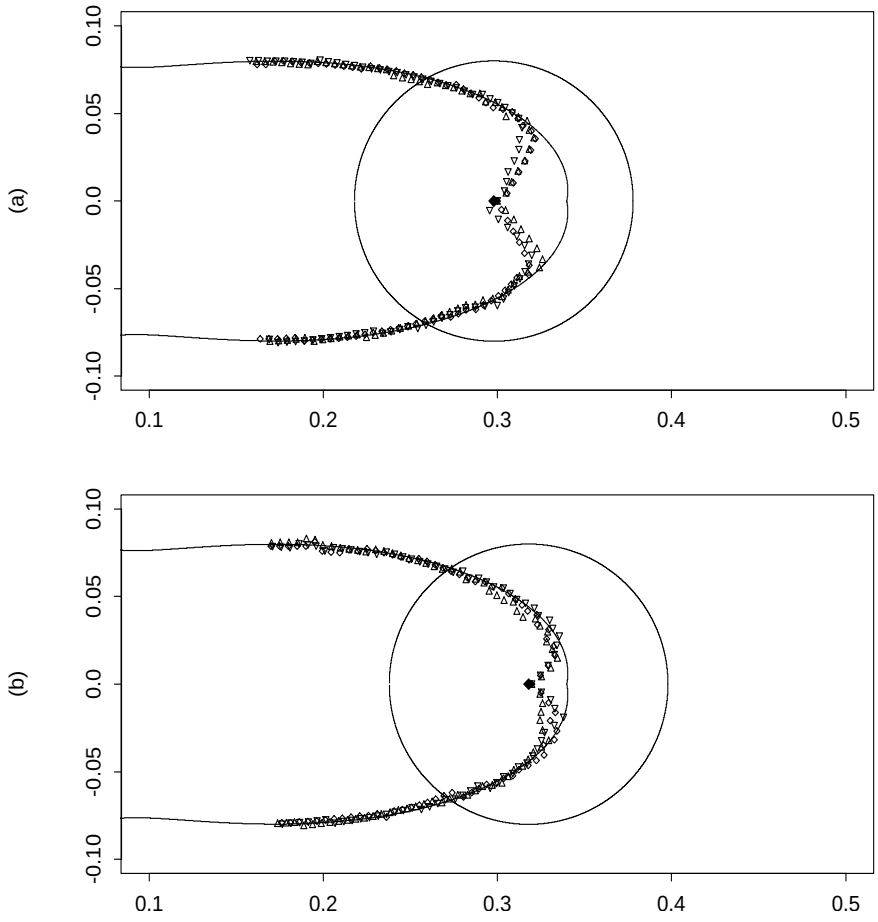

Figure 5.2: Panels show the effects of choice of $\hat{Q}$ on behaviour of the tracking estimate. Three realisations are depicted on each panel, represented respectively by empty diamonds, by empty triangles pointing upwards, and by empty triangles pointing downwards. The starting point $\hat{Q}$ is represented by a filled diamond. It moves steadily from the left-hand side to the right-hand side of the curve as we progress through panels (a)–(d).

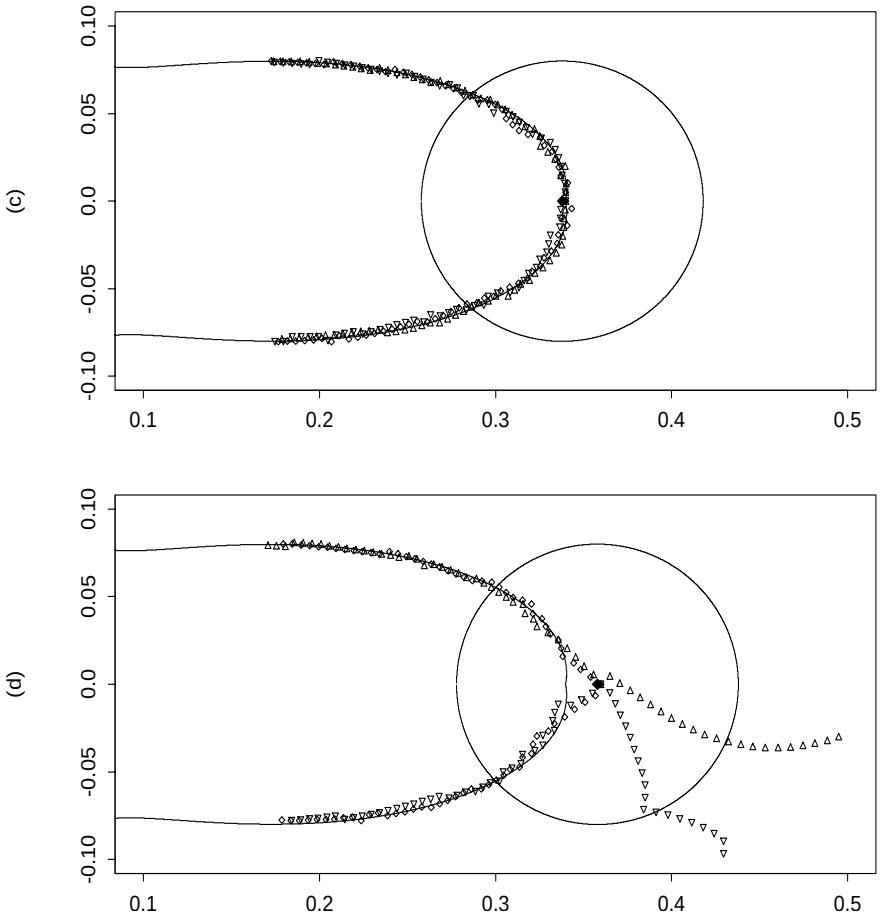

Figure 5.2: Continued

5.3 Numerical Properties of Likelihood Method

In this section we report the results of simulation studies of finite-sample performance of the asymptotic confidence band r_{sim} suggested in Remark 4.3 of Section 4.2. We again consider, in the context of nonparametric regression, the family $\{g_d\}$ of surfaces defined in the paragraph containing (5.2.1), this time with varying $d > 0$. The design process $\mathcal{X}$ was homogeneous Poisson, and observed within the square $\Pi = [-0.55, 0.55] \times [-0.55, 0.55]$. For the intensity of $\mathcal{X}$ we considered the two values $n_1 = 375^2$ and $n_2 = 750^2$, and the error distribution was chosen to be Normal with standard deviation $\sigma = 0.75$. We further took $h = 0.08 \approx 4.16\, n_1^{-1/3}$. As in Section 5.2, K was chosen as the biweight kernel function.

In order to estimate $\rho = \rho(d, \sigma) = 1 + d^2/\sigma^2$ as the natural plug-in estimator $\hat{\rho} = \rho(\hat{d}, \hat{\sigma})$, we employed the empirical procedures suggested in Remark 4.2. These were found to perform well when the estimate $\hat{\sigma}$ was obtained from response values within the sub-window $[0.35, 0.55] \times [0.35, 0.55] \subset \{g = 0\} \subset \Pi$, and provided that the estimate $|\hat{d}|$ was obtained from stretches of $\widehat{\mathcal{C}}$ which were sufficiently far from the points $(f(\phi_i)\cos\phi_i, f(\phi_i)\sin\phi_i)$ for $i = 1, \ldots, 6$, thereby ensuring that the bias due to curvature was only moderate. (The maximisers $\phi_1, \ldots, \phi_6$ of absolute curvature are as defined in Section 5.2). Because of the constancy of the jump height in this example, and for the sake of simplicity, we took $\hat{\rho}$ equal to its true value in the simulations of this section. If $\hat{\rho}$ is instead estimated from data which suffers from curvature bias, the confidence bands will be wider, as is evinced by formula (4.2.7).

Estimation of a_0 has significant impact on estimation of the local band radii r_{ptw} and r_{sim}. In fact, it is the sparseness difficulties arising in this context that necessitates the choice of a relatively large value for n_1. To compute the estimate we searched along the rays represented by the equation $\phi = c$, with c running over the points on the uniform grid on the interval $[-3\,h, 2\pi + 3\,h]$, with edge size 0.01, and including the left end-point. This served to overcome the problems by which the derivative estimators would otherwise have been affected in the boundary regions. For non-closed fault lines, of course, edge effects would be more difficult to remedy. In construction of the band itself, the overlap region of total length $6h$ was discarded. Note that the above search method is viable whenever the fault line encloses a strictly star-shaped region (as defined in Section 1.1) that is known to contain O. However, in this example, constructing the estimate using the definition given in Section 4.2 (see p. 57) to obtain points on the ridge $\widehat{\mathcal{C}}$ requires only moderate extra effort.

Points on $\widehat{\mathcal{C}}$ were computed by the procedure described shortly. Based on these points,

local curvature was estimated using Gasser-Müller type kernel estimators for each of the functions f, $f' = df/d\phi$, $f'' = d^2f/d\phi^2$. For an account of the underlying theory, see Gasser, Kneip and Köhler (1991), Herrmann (1997), and the references cited therein. (Usage of notation f', f'' is limited to this section only. A similar remark applies to the interpretation of the function κ introduced shortly.) Gasser *et al.* (1991) employed an iterative algorithm to estimate the bandwidth which is asymptotically optimal in the sense of integrated mean square error. We used an implementation in Fortran code, available from the URL address

http://www.unizh.ch/biostat/Software/.

The estimators for f' and f'' are also displayed in Ramsay and Silverman (1997, p. 53). Estimates were computed at 130 points equidistantly spaced on $[0, 2\pi]$, and bandwidths for the three functions were selected independently for each realisation. As expected, these tended to be larger for $n = n_1$ than for $n = n_2$, and increased with decreasing jump height d, for which we considered the three values 1.0, 0.3, and 0.08. For $n = n_2$ and $d = 1.0$, the average bandwidths over 40 realisations were $h_f = 0.024$, $h_{f'} = 0.052$, and $h_{f''} = 0.108$. The kernel estimates were plugged into the exact formula for the true curvature at a point $(f(\phi)\cos\phi, f(\phi)\sin\phi) \in \mathcal{C}$, as displayed in equation (A.2.5). (We used the individual estimates of $h_f, h_{f'}$ and $h_{f''}$ in these computations, rather than the averages.) Since the true regression surface is a black-and-white image, the area-based curvature formula from Section A.2 does not apply in this example. We note that in other cases, the use of that formula can be coupled with the tracking approaches from Chapters 2 and 3. The ensuing procedure may be construed as the tracking estimator for the *tube* $\widehat{\mathcal{C}}^{r_2}$, defined at (4.2.8). If the curvature estimate at a given point on the estimate is based on tracked points that are further advanced, the tube will have a 'lag' of the size of approximately one bandwidth. An algorithm that incorporates the area-based curvature formula is presented at the end of this section.

Approximations to critical values $t(\alpha)$ are given in Table 5.1. We used two different methods to obtain the displayed results. The first utilised a central limit theorem for partial-sum processes, based on a minor modification of Example 2.12.9, p. 230 of Van der Vaart and Wellner (1996); the second employed a bivariate Karhunen-Loève expansion as described in Subsection A.5.2.

Panels (a)–(d) of Figure 5.3 illustrate, for different input parameters in each case, three realisations of point sequences on $\widehat{\mathcal{C}}$, calculated by the search method described above, and represented by squares, diamonds and 'x' symbols respectively. Also shown are the

α	0.250	0.100	0.050	0.025	0.010
Critical value $t(\alpha)$	0.73	1.04	1.24	1.41	1.58

Table 5.1: Approximate critical values $t(\alpha)$ such that $P\{|T^*| \leq t(\alpha)\} = 1 - \alpha$.

respective simultaneous confidence bands, represented by the solid, dash-dotted, and dotted lines. In order to construct the radius of the band locally, we used the vector $\hat{\kappa} \in \mathbb{R}^{130}$ from above, and obtained an estimate of a_0, of the same length, by taking the maximum modulus of estimated curvature at five sequential points, thus obtaining a vector $\hat{a}_0$ which had the same value within each of the $130/5 = 26$ stretches. The extra points at the boundaries were again used to obtain the estimates near $\phi = 0$ and $\phi = 2\pi$. In order to lessen the bias inherent in this procedure, we used a random offset to shift the location of the entries within $\hat{a}_0$. The points on the 'inner' and 'outer' bands were computed relative to the kernel estimate of f, rather than the ridge line $\widehat{\mathcal{C}}$ itself. Finally, as suggested in Section 3.2 (see p. 57), the bands were obtained by joining these calculated points by straight lines, to yield polygons. Alternative or post-processing of the bands, for example by using B-splines, is clearly desirable, but we shall not further discuss such ramifications.

We used L^1 distance to measure deviation of the estimate; see Remark 4.3. This distance can be computed in MATLAB with the function `polyarea.m`, which computes the area enclosed by a polygonal curve. For this purpose, the edge was approximated by the polygon with vertices at the points $(f(\gamma_i)\cos(\gamma_i), f(\gamma_i)\sin(\gamma_i))$ with $\gamma_i = 0.02\,i$ for $i = 0, \ldots, 314$. Table 5.2 shows the median and upper and lower quartile behaviour of the estimator according to these measurements for $n = n_1$ and $n = n_2$, and the three jump heights d mentioned earlier on. In order to obtain an average of a local performance measure, these results should be related to the arc length of $\mathcal{C}$ (equal to 4.272). The pointwise coverage was computed using the MATLAB function `inpolygon.m`, which determines whether a given point lies inside a polygonal region specified by the vertex coordinates. Pointwise coverage was below the nominal levels, with a median of about 65% in the two respective cases, and a variability of roughly five to ten percent. It is not too surprising that the coverage does not improve with increasing d. Indeed, as the signal-to-noise ratio is lowered, the estimate of the second derivative has a higher variance, and by construction of $\hat{a}_0$ this implies that estimated radii of the confidence bands increases. This effect counteracts the increasing deviation of the ridge from the true fault line, which is only proportional to $\rho^{1/3}$. The centring convention from the previous paragraph also caused coverage to generally improve. was already lessened.

In conclusion to this section we give the definition of a purely statistical estimation proce-

	$n = n_1$			$n = n_2$		
Quantile	25%	50%	75%	25%	50%	75%
$d = 1$	0.0160	0.0176	0.0198	0.0132	0.0144	0.0152
$d = 0.3$	0.0184	0.0211	0.0230	0.0156	0.0176	0.0201
$d = 0.08$	0.0208	0.0263	0.0337	0.0172	0.0191	0.0221

Table 5.2: Empirical median and upper and lower quantile behaviour, obtained using in each case 40 realisations of the fault line estimator with an L^1-criterion, as described in Section 4.2.

dure for both $\mathcal{C}$ and its confidence envelope, under the additional assumption that $\mathcal{C}$ has the interpretation of a level curve as in the paragraph containing (A.2.7). The tracking algorithm which yields the points $\{\hat{x}_j\} \subset \widehat{\mathcal{C}}$ is the same as described in Chapters 2 and 4. Assume that the sequence of points $\{\hat{x}_0, \ldots, \hat{x}_M\}$, for $M \geq 0$, has already been tracked. The local tangent estimate $\hat{\theta} = \hat{\theta}(\hat{x}_M)$ divides the kernel disc of radius h, centred at $\hat{x}_M$, into two halves. Choose the one for which the mean-square error is the smallest, and use it to estimate the derivatives required in (A.2.8), for example by cubic spline fitting. This enables estimation of the local radius $r_2 = r_2(\hat{x}_j)$ and hence augmentation of the confidence envelope to the point $\hat{x}_{M+1}$. Note that the mean-square criterion in the foregoing definition is used in a similar way as in one of the surface estimation procedures proposed by Qiu (2002b).

The foregoing procedure has the disadvantage of being applicable only under more severe restrictions than those required by the conditions of Theorem 4.2. In the next section we devise an extension of the edge estimator which is specifically aimed at wider applicability.

5.4 An Edge Detection Architecture Using Canny's Method

In this section we examine how the edge estimators proposed in Chapters 3,4 and particularly 5 may assume a role in the primary processing of images even if edge has features that are not covered by our theorems. (Our use of the term 'features' here differs from the abstract notion from pattern recognition, introduced in Section 1.2.) Such features, which are commonly encountered in practice, prominently include splitting or terminating edges, straight line segments (for the theory of Chapter 4), and corners. The first of these cases has been addressed in Remark 2.9, whence we shall focus on the remaining two. The corner-sensitive edge estimator to be developed will accommodate the detection of straight line segments, which is the simpler of the two tasks from an image analysis (and non-statistical) viewpoint. As pointed out in Section 1.2 (see p. 15), a corner may

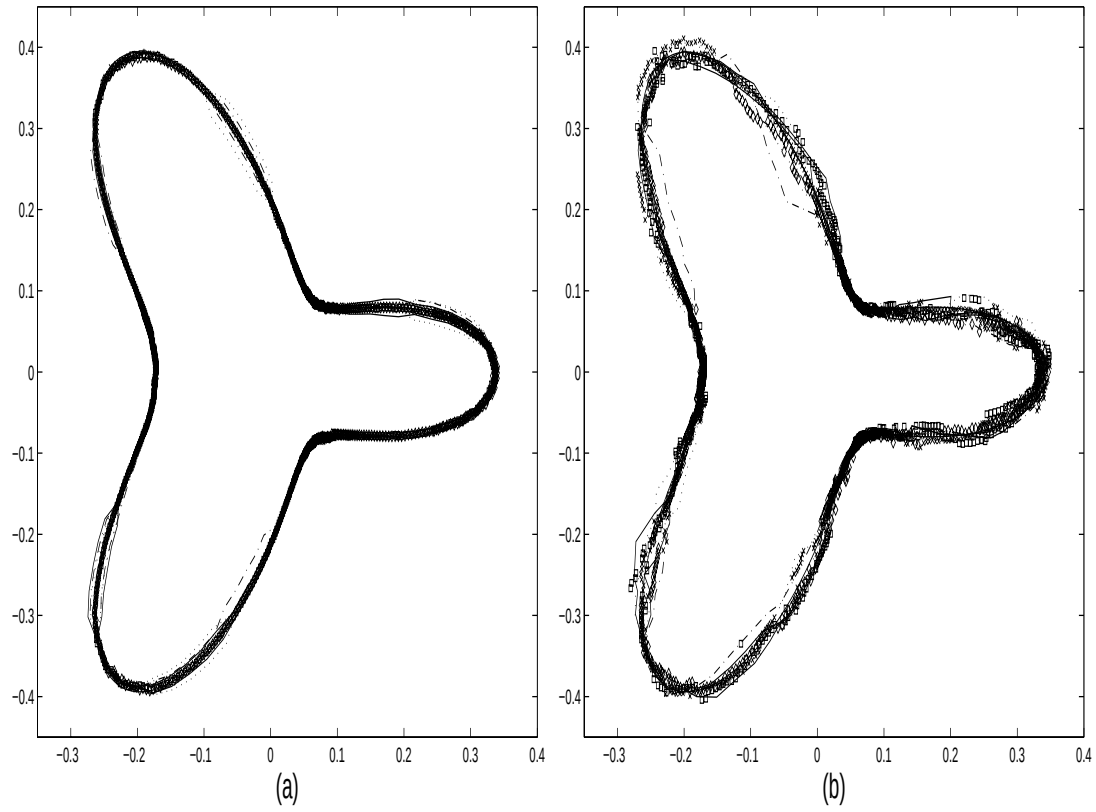

Figure 5.3: Illustration of three superimposed realisations of simultaneous confidence bands, constructed using the procedure described in Section 5.3. Panel (a): $n = n_1$ and $d = 1.0$. Panel (b): $n = n_1$ and $d = 0.08$. Panel (c): $n = n_2$ and $d = 1.0$. Panel (d): $n = n_2$ and $d = 0.08$.

be construed as a feature of one level of complexity above an edge.

Our goal here is not to claim any optimality property of the edge estimator, which would be complex to analyse for the (non-smooth) class of edges considered. Instead, we want to back our assertion that it can be used as a building block in a more sophisticated architecture, whose 'macroscopic' (i.e., working on scales above the bandwidth of the local-likelihood estimator) parts are readily exchangeable.

In order to achieve greater robustness for the edge detector to be developed, we utilise concepts from computer vision, and especially scale-space theory as described in Section A.2. Upon detection of these features, extraction of the fault line may proceed separately for each of the smooth segments of $\mathcal{C}$. A preliminary analysis showed that within those segments, the standard edge detectors do not perform well enough to be competitive to the ridge estimator from Chapter 4. On the other hand, their computation consumes only a fraction of the computing time for the ridge estimator, while yielding

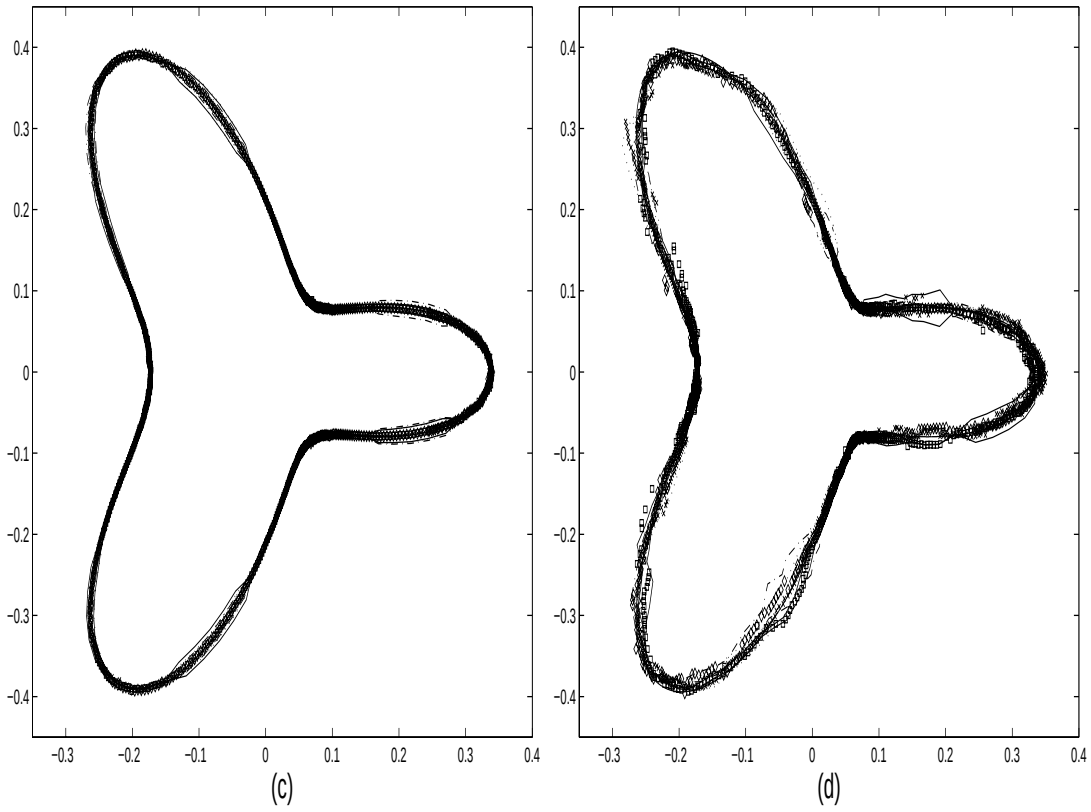

Figure 5.3: Continued.

results that can be utilised in an edge detection architecture. We shall therefore not dwell on an examination of performance of the edge detectors considered in this section relative to each other, but instead propose a method that, as we argue, makes good use of their relative advantages, with a special focus on computational ease. We do not claim that the approach will perform well with the diversity of real-world data, but want to highlight instead that local-likelihood methods are capable of being integrated into such architectures.

For the purposes of this section, we relax the assumptions on the fault line $\mathcal{C}$ by allowing for a finite number of exceptions. Specifically, we assume that there exists $N \in \mathbb{N}$ and $0 = x_0 \leq x_1 \leq x_2 \leq \ldots \leq x_N = T$, where $[0, T]$ is the parameter set in the natural parametrisation of $\mathcal{C}$ ($T = \text{length}(\mathcal{C})$), such that $\mathcal{C}$ is twice differentiable on $[x_i, x_{i+1}]$, for $0 \leq i \leq N - 1$. As before, if $\mathcal{C}$ is closed then an arbitrary point on $\mathcal{C}$ may be used as starting and end point. The above mentioned features are then characterised by the behaviour of the second derivative: the interest is then in its zero-crossings and infinities, respectively. In analogy to derivative-based edge detectors (see Section 1.3), many corner

detection algorithms operate by examining sharp extrema of the second derivative of the curve.

The discontinuous regression function that we shall consider in this section is defined as follows, where the sampling window is again equal to $\Pi = [-0.55, 0.55]^2$:

$$\begin{aligned} g(r,\phi) &= I\big\{f(\phi) \le r \text{ or } \big(x^{(1)}, x^{(2)}\big) \in \mathcal{P}\big\}\,, \\ f(\phi) &= \big[\exp\{\cos(3\phi)\} + \sin(\phi/2)\big]/8\,, \quad 0 \le \phi \le 2\pi\,, \end{aligned} \tag{5.4.1}$$

where the polygon $\mathcal{P}$ has the following definition in Cartesian coordinates:

$$\mathcal{P} = \big\{\big(x^{(1)}, x^{(2)}\big) :\, x^{(1)} \ge -0.08,\, x^{(1)} \le 0.3 - 0.5\,x^{(2)}, x^{(2)} \ge 0.3\,x^{(1)} - 0.5, x^{(2)} \le 0.1\big\}\,.$$

We will restrict ourselves to the case where the explanatory data are on a grid with $n = 175^2$ points per unit area. As has been mentioned already in Section 1.2, gridded design is a standard requirement of edge detectors like the Canny method. The function at (5.4.1) is a modification of the example used in Sections 5.2 and 5.3, with the intention to examine the features of special interest in this section. Panel (a) of Figure 5.4 shows the image with the previous specifications, subjected to Normally distributed noise with standard deviation $\sigma_{\text{Noise}} = 0.3$. Panel (b) shows the Euclidean norm of the image gradient, approximated by differencing.

In the first step of rough isolation of edges, various (robust) edge detectors could be used, such as the Sobel, the Laplacian of Gaussian, and other user-specified filters. For the purposes of this section, we shall limit ourselves to the Canny edge detection algorithm essentially as proposed by Canny in his original paper (Canny, 1986), to which we refer for more details. A major reason for this choice is that the Canny algorithm is considered to be superior theoretically, as well as more successful in practice, than other derivative-based methods. The Canny edge detector defines edges as the loci of points where the second derivative changes sign from + to − while following the direction of the image gradient. Estimates of these derivatives can be computed using (A.2.10).

With regard to the edge pixels which are found in the vicinity of the true edge by the Canny algorithm (and in fact many others such as the one proposed by Qiu and Yandell, 1997), two tasks emerge, which are described in the next two paragraphs.

Because the peak of the estimated gradient function, from which the JDC of the Canny algorithm derives, may extend over several pixels perpendicular to the edge direction, the first problem consists in the need to 'thin' the edges. (Recall from Section 1.3 (see p. 20) that JDC stands for 'jump detection criterion.') The common remedy of *non-*

maximum suppression takes the original values of the JDC, and discards those pixels whose gradient values on either side, obtained by interpolation from neighbouring pixels, are not both smaller than the JDC value for the original pixel. For their own JDC, Qiu and Yandell (1997) used a procedure (called MP1 there) which is similar to non-maximum suppression, but using only the set of previously-computed candidate points, rather than the JDCs of these points and their neighbourhood. Importantly, in both of the above cases, $\mathcal{C}$ is assumed to be differentiable: the theory does not apply to corners in images. It is important to know how strongly this restriction is felt in practical situations, where kinks are almost invariably occurring as isolated points, and whether that problem can be dealt with by means of a corner detection algorithm. We will return to these issues later in this section.

The second of the aforementioned tasks consists in making the boundaries contiguous, as the value of the JDC will generally fluctuate across the edge, a phenomenon commonly referred to as *streaking*. To overcome streaking, one commonly uses *hysteresis thresholding*, which operates by setting a lower as well as an upper threshold limit. The concept of hysteresis thresholding already appeared in Section 3.2 (see p. 56). Denoting these thresholds again by t_{lower} and t_{upper}, then all points with JDC $\geq t_{\text{upper}}$ are immediately accepted, points with JDC $\leq t_{\text{lower}}$ are immediately rejected, and points in the interval $(t_{\text{upper}}, t_{\text{lower}})$ are accepted if they are, for a given choice of neighbourhood (for example, 3×3), connected to a pixel for which JDC $\geq t_{\text{upper}}$. As can be seen in panel (c), the value of the pair $(t_{\text{lower}}, t_{\text{upper}})$ that was selected by the 'magic' number conventions of MATLAB, which is based on an assumption of the proportion of edge pixels in the entire image, did not yield a satisfactory outcome. Panel (d) shows the result after using the modified lower threshold t^*_{lower} with

$$t^*_{\text{lower}} = 0.25\, t_{\text{lower}} + 0.75\, t_{\text{upper}}\,. \tag{5.4.2}$$

Another, although somewhat subordinate, issue in post-processing the edge candidate points is the deletion of isolated spurious edge pixels. In their modification procedure MP2, Qiu and Yandell (1997) reject a candidate pixel if less than half of its neighbours are flagged as candidates. In our post-processing method, to be described next, such spurious pixels are detected through a somewhat different notion of connectivity.

We now proceed to describe our edge detection scheme in full. The first step after obtaining the set of edge pixels through the Canny algorithm is the linking of the edge points by line segments. In a second phase, the ensuing segments are linked together to produce a more connected estimate. (Although the Canny method produces *ab initio* a

connected edge because of its zero-crossing definition, the subsequent processing destroys this property.) Both steps require the specification of tolerances. The first applies to the locations of the points in a sequence: points are considered collinear if the maximum deviation from a line is not too large. For algorithms that are instrumental here, see for example Chapter 5 in Preparata and Shamos (1985, pp. 185ff). The second tolerance value relates to the maximum turning angle between neighbouring segments, and the distance of one such segment to the next. The implementations we used are authored by Peter Kovesi, and were obtained from the URL address

http://www.cs.uwa.edu.au/~pk/Research/MatlabFns/index.html,

as the functions `edgelink.m`, `lineseg.m`, `mergeseg.m` and `maxlinedev.m`. The output of the pre-processing steps is shown in panel (e) of Figure 5.4. The spurious cusp is clearly undesirable, but will be seen to result in a false positive in this case. A small number of such points is not really detrimental to the procedure, and may thus be tolerated.

In the next step, the list of line segments is ordered according to two criteria. These are, in order of priority, of the same nature as in the construction of each individual line segment as outlined in the previous paragraph, namely proximity and slope, but with parameter relaxation. By way of this procedure, the problem of spurious edge pixels is also largely overcome.

With the first stage of edge detection thus completed, attention is now turned to the detection of corners. As mentioned previously, corner detection relies on edge detection, either implicitly or explicitly, in most cases. An approach in which the latter step is of less significance appeared in Kohlmann (1996), and we shall use this approach in getting a pilot estimate of the number of corners. It is based on the discrete Hilbert transform. For a definition of this transform, also in two dimensions as needed here, we refer to Kohlmann (1996, pp. 226–228). Several examples of the effects of the Hilbert transform on noise-free images may also be found in that paper. The key point is that discontinuities appear as singularities in the Hilbert transform, whereas the smooth parts are suppressed, leading to a highlighting of corners in the output image to an extent depending on its orientation. In order to select the corner candidates, we used the following 'histogram-type' rule. First, we took the 100 pixels whose brightness values were the largest. This gives a cluster of marked pixels, most of which are expected to be near the locus of a corner of $\mathcal{C}$. Then, we used morphological operations to reduce these clusters to single points. The resulting number was taken to calibrate the value of σ for the curvature scale space parameter for the procedure, which is used to yield the estimated set of corner points $\{\hat{x}_i,\, i = 1 \ldots, N'\}$. For reasonably 'well-behaved' edges, the number of

false positives is not too large and that of false negatives is small; ideally, after suitable enumeration, $N = N'$ and $x_i = x'_i$ for every i.

The definition of the evolution of $\mathcal{C}_\sigma$ was given around equation (A.2.11). We start with a relatively large value of σ and successively lower its value. The number $K = K(\sigma)$ of corner candidates as obtained by the curvature scale space (CSS) procedure described by Mokhtarian and Suomela (1998), which built on the work of Mokhtarian and Mackworth (1992) mentioned at the close of Section A.2, is then recorded. Due to the observation made in Section A.2 on the postulates of the scale space (see p. 142), K can be expected to be a monotonically decreasing function of σ. We take the target number of corners from the procedure described in the previous paragraph as the benchmark, in that we successively lower the value of σ until the number of corners computed from $\mathcal{C}_\sigma$ exceeds this value. The sequence of values for σ need not to be spaced very densely, as the function K will usually have abrupt jumps and be constant over reasonably large intervals.

Panel (f) of Figure 5.4 shows the output of the CSS procedure after 20 iterations, based on one simulation; further replica showed comparable or better behaviour. The starting point for either the tracking method of Chapter 2, or the likelihood methods of Chapters 3 and 4 are highlighted as the red + symbols. For reference, the output of the two-dimensional Hilbert transform and the corner detector of Kohlmann (1996) are displayed in panels (g) and (h) respectively. From panels (f) and (h) it can be seen that although some false positives as well as some false negatives are created, performance of the estimator is overall quite satisfying.

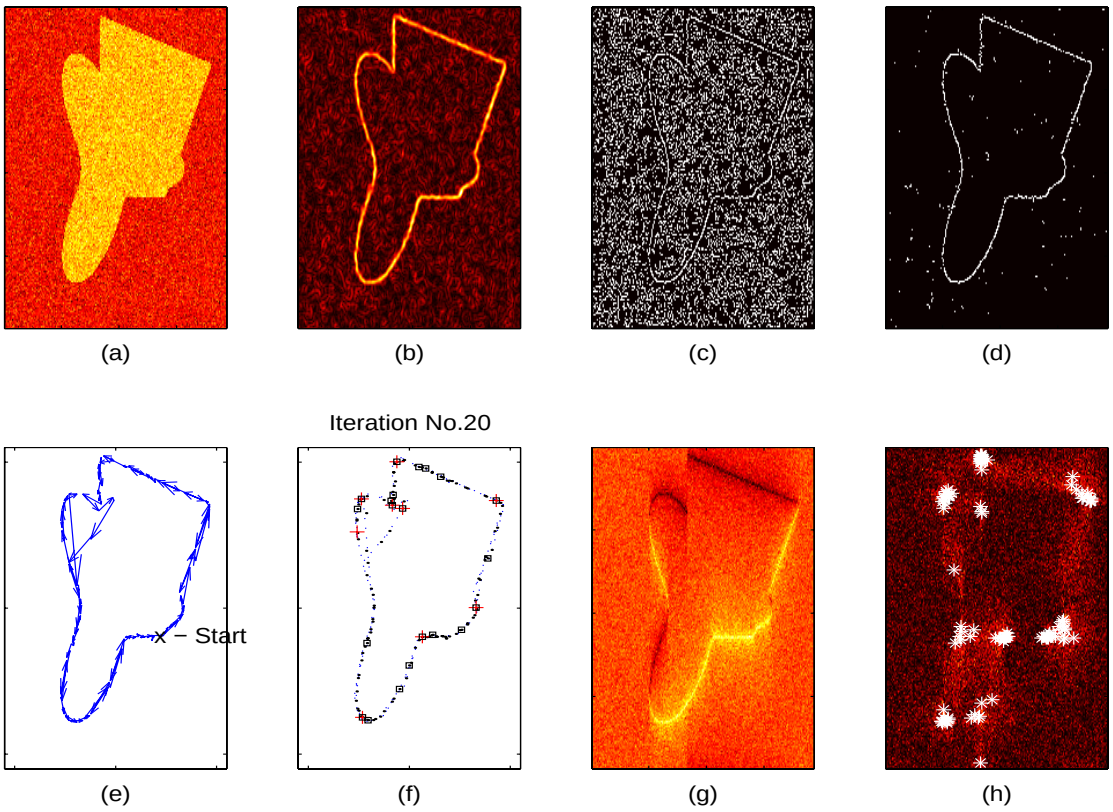

Figure 5.4: Edge detector performance for the image defined by equations (5.4.1) in Section 5.4.

Chapter 6

Functional Signal Discrimination

6.1 Introduction

In this chapter, a discrimination method for spatial functional data, in the sense of Section 1.2, is considered. Following the general theme of the thesis, dimension reduction assumes a key role in the analysis. Practical issues related to the methodology of the classifier, which includes data pre-processing steps, are explored by means of a real dataset of navigational radar targeting ships. The two cornerstones of our approach in the above respects are nowadays commonly referred to as *wavelet shrinkage* in the sense developed in the seminal paper by Donoho and Johnstone (1994), and the already mentioned area of *functional data analysis*, which was first extensively treated in the monograph by Ramsay and Silverman (1997).

As described in Subsection A.5.1, we interpret the data as realisations of a stochastic process defined over a continuous space. Earlier in this thesis (see Section A.5) we gave an exposition of the background of the continuous Karhunen-Loève transform in terms of (non-statistical) functional analysis. We also argued that it is an attractive candidate for the task of condensing data that has a probabilistic aspect, into a low-dimensional subspace. Among the bonuses that the Karhunen-Loève approach offers is, as noted earlier (see the corresponding remarks on p. 10), that the underlying probability model requires few specifications, and is instead mainly data-driven. The most important point is the need to specify the set of (Bayesian) prior probabilities, denoted $\{\omega_l, 1 \leq l \leq L\}$ in the paragraph containing (1.3.5), where L is the number of classes. In the absence of prior information on observation occurrence we shall assume a uniform prior, that is,

$\omega_l = 1/L$ for $1 \leq l \leq L$.

In Subsection A.5.3 we pointed out that it is generally necessary to exercise care in reducing dimension for the purpose of discrimination, so as to preserve salient features. In this context, saliency is not to be equated with visual evidence from a graphical plot of the data. The latter task can be fairly challenging, especially when the variability within each class is large compared to variability across classes, such as is the case for the dataset studied in Section 6.3.

Similar to the use of principal components in classical multivariate analysis, a central role in FDA is played by the approximation of the input data by linear combinations of basis elements. This basis then comprises *functions* (one-dimensional curves in Hall, Poskitt and Presnell (2001) and two-dimensional surfaces in this thesis), rather than just vector ensembles as in multivariate analysis. These functions may be subjected to smoothness and boundary constraints, although only the former turn out to be of importance here. In the FDA setting that we pursue, it is natural to require at least the following two conditions on the functions with respect to which the observations, modulo statistical error, are represented:

1. The basis functions (surfaces) should have qualitative properties that are closely comparable to, or at least not inconsistent with, the 'true' functions (surfaces) from each type.

2. The information from the original domain of the data (temporal in the case of curves, spatial in the case of images) should be used to a greater extent than frequency (Fourier) domain information, if the latter is not entirely disregarded.

The point stated in item 1 is essentially the same as in Ramsay and Silverman (1997, p. 46). Item 2 is included because artifacts created by the imaging device or subsequent pre-processing steps are most likely to occur in the Fourier domain. The caution towards Fourier domain information expressed in item 2 was motivated by the statement in Hall, Poskitt and Presnell (2001, p. 1) on the "problems associated with the highly nonlinear preliminary filtering [to which signals are subjected.]" As mentioned there, this makes it less attractive to employ tools from time series analysis. The problem of artifacts due to pre-processing is also prevalent in the analysis of fault lines with noisy surface data, as has been remarked in Section 1.2. In the present context, however, these issues are particularly serious.

As at the close of Subsection A.5.1, the coefficients of the low-dimensional linear combinations of basis functions which approximate the observations will be called *scores*.

Because our interest is in discrimination, which is based on the score vectors, the goal as well as the method to be described put a much stronger emphasis on data-analytic rather than visual-perceptional properties. Nevertheless, the dimension reduction technique employed here plays an important role in a visually-oriented explanatory data analysis. It may also serve as a valuable guide in choosing model parameters, for example by complementing the method suggested in Subsection 6.3.3 for selecting the basis functions.

We proceed with the formal definition of the discrimination or classification problem. As noted in Section 1.3, we shall use the two terms synonymously.

Definition 6.1. (Mardia, Kent and Bibby, 1979, p. 300) *Let* $\Pi_1, \ldots, \Pi_L$*, for* $L \geq 2$*, denote* L *populations or groups, and suppose that for each population* Π_l*, there is a probability density* $f_{(l)}$ *on* $\mathbb{R}^m$ *such that an individual from population* Π_l *has probability density function* $f_{(l)}$*. A* discriminant rule *d corresponds to a division of* $\mathbb{R}^m$ *into disjoint subregions* $R_1, \ldots, R_L$ $(\cup_l R_l = \mathbb{R}^m)$*. The rule is defined by*

$$\text{allocate } x \text{ to } \Pi_l \quad \text{if} \quad x \in R_l \,.$$

We shall apply the Karhunen-Loève expansion to the case where the underlying stochastic model is that of a mixture distribution of $L \in \mathbb{N}$ classes or types, as introduced in Subsection A.5.1. Recall equation (A.5.7):

$$P_X = \sum_{l=1}^{L} \omega_l P_{\tilde{X}_l} \,,$$

where the prior probability of drawing a sample from the lth population equals ω_l, and, under the assumption of a uniform prior, may be replaced by the relative frequency n_l/n of drawing a sample from class l.

From Subsection A.5.3 we recall that the singular value decomposition (SVD) of the empirical covariance matrix of observations of the process X leads to natural estimators of the quantities defining the Karhunen-Loève expansion, particularly the score vectors. The relationships between the functional and vectorial, that is the classical multivariate, aspects of the analysis in this chapter are illustrated in Figure 6.1. Importantly however, for the continuous form of the Karhunen-Loève transform, smoothing has an interpretation and is used in Subsection 6.3.2. This can help these functions meet the stipulation from item 1 in the display before Definition 6.1. For our dataset, however, classification error was only marginally affected by this device.

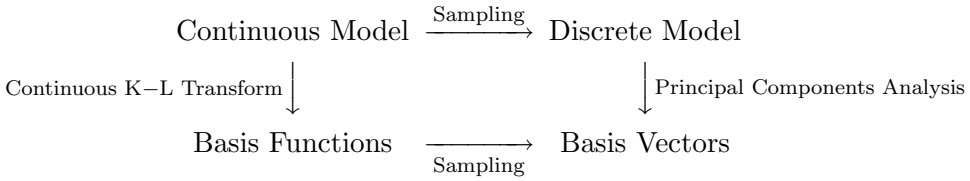

Figure 6.1: Interrelationships between 'discrete' and 'continuous' views of dimension reduction via the Karhunen-Loève transform.

Using once more the notation from Subsection A.5.3, it follows from Mercer's Theorem (Theorem A.4 in Subsection A.5.1) that the Karhunen-Loève approximation obtained by using the first $m \leq \text{rank}(\underline{X})$ empirical basis functions maximises the proportion of variability explained, among all sets of orthogonal functions with m elements. The maximum value of this proportion then equals ζ_m, defined at (A.5.6).

We now turn to a preliminary discussion of the database used in our experiments. This database, on which more elaboration will be given in Subsection 6.3.1, consisted of over 3000 images of low-resolution navigation radar data on various target objects. These images will subsequently also be referred to as radar *signatures*. The axes had the physical interpretation of azimuth and range (i.e., the distance of the object from the location of the radar antenna, which was stationary). It was a technological feature, and of importance in later considerations, that the information content in each of the sampling bins (resolution pixels) was appreciably larger in range than in azimuth direction. The strength of the return within each azimuth/range bin was measured by an eight-bit integer ranging from 0 to 255. In a preamble to each datafile, physical information such as target range, speed, bearing and course was usually also available. Together with this preamble, the image comprised what we shall call a *dataframe*.

For the classification experiment in Subsection 6.3.4, the range and speed information were used directly. Bearing and course could be combined, using a simple formula, to obtain the angle of view relative to the observing platform. The angle of view is highly relevant in radar image acquisition (see Acker, 1988, Robinson, 1996, and Inggs and Robinson, 1999). However, it does not seem easy to incorporate this information into the classifier. This may be seen as a drawback in comparison to purely Bayesian or neural network methods. For a discussion of various other approaches to classification of data of the type considered here, see Gibbins, Gray and Dempsey (1999). It should be noted that Bayesian elements are relatively easy to incorporate into the classifier to be developed here. This could be done, for example, by using non-uniform prior probabilities in the weighting of the odds-ratio at (1.3.5) that is used here, just as in Hall, Poskitt and Presnell (2001), to choose the most likely class. It is also conceivable to embed the

method presented here as a module into a different (e.g., tree-based) classifier.

The targets themselves were vessels of different kinds, which for classification purposes were divided into three 'superclasses,' referred to below as A1, A2 and A3. These superclasses were in turn subdivided into seven sub-categories, here labelled s1, s2,. . ., s7. The class frequencies are listed in Table 6.1.

Although the issues arising from the *spatial* nature of the data do not profoundly alter the methodology of the classifier itself, greater problems arise from the influence of 'noise' (the term is used in a somewhat loose sense as it does not necessarily refer to a strictly probabilistic notion), and sea clutter. The resolution of the individual scatterers on a vessel was limited by several factors. These included prominently the convolution with the approximately rectangular-shaped pulse (see the discussion in Subsection 6.3.1), and the *multipass* and *ghosting* effects that occurred as the radar bounced around the vessel and on the sea surface, especially at larger distances. The angle of view also had significant impact on the quality of the radar signature.

In the appearance of the noise as well as the signal, a crucial role is played by the technology of the radar. However, it seems a rather difficult task to exploit the dependencies in order to improve the signal-to-noise ratio (SNR). (Note that the SNR is defined here as the ratio of mean-square-root error of the signal and the noise respectively, in contrast to the definition at (4.2.5) in the context of edge estimation). In spite of their theoretical strengths, many statistical methods that could be considered for the purpose of this chapter have the disadvantage that the choice of their parameters is not easy to motivate from the physical characteristics of the imaging process. Automated parameter selection methods are often grounded on assumptions on large sample size and/or stochastic independence which are only approximately satisfied, at best, in the present case study. These considerations are of relevance both in improving the SNR and in the classification itself. Much of the subsequent development was motivated by the goal to use enhancement methods that are not only well-suited to the context, but which are also relatively insensitive to changes in ill-defined 'parameters,' notably weather and sea conditions. They should also work with only a small number of input parameters. The popular tool of wavelet thresholding is well suited to this task.

6.2 Methodological Issues

In order to obtain the input quantities for the classifier, the radar signature (that is, the matrix of scatterer strength for each range/azimuth bin within the observation win-

dow) is subjected to several pre-processing steps which are the topic of the subsequent sections. This serves to achieve a degree of uniformity in sampling, especially as the observation window, which was chosen automatically, contained a highly variable amount of noisy (sea) background surrounding the tracked object. (This amount tended to be especially large for weak signals at large ranges.) Moreover, after discarding the background part (reusing it, however, to set noise attenuation parameters; see Subsection 6.2.2), the segmented, or rather cropped, image is the one from which relevant measurements, for example pixel extension in range, are extracted to assist classification.

The purpose of this section is to discuss in some detail two salient problems that arise in pre-processing of the image data, namely segmentation from the noisy sea background (Subsection 6.2.1) and noise suppression/reduction via wavelet thresholding (Subsection 6.2.2). Being a third important topic, *registration* appears to be the most elusive in relation to this dataset. Subsection 6.2.3 outlines the problems involved here, alongside with general comments and references on registration of functional data.

6.2.1 Preliminary Segmentation

In order to segment the images from the ocean background noise, we use the morphological operations of dilation and erosion, $\oplus$ and $\ominus$, defined in Section 1.1. With sets approximated on a pixel grid, the set B from that definition, in this context sometimes called the *structuring element*, typically comprises only a few pixels, and its shape is crucial in determining the preferential features to be eroded (i.e., thinned out). Operators $\oplus, \ominus$ and various others are built-in functions in the Image Processing Toolbox of MATLAB, from Version 6.

It should be noted that the goal in the present section is to remove only those parts of the background which are readily identified as noise. Thus, we accept at this early processing stage that the complementary part may still overestimate the bounding box, and hence the actual pixel dimensions $\dim_{\mathrm{az}}^{\mathrm{true}}$ and $\dim_{\mathrm{rg}}^{\mathrm{true}}$ of the target.

Because there is latitude in the choice of parameters for the morphological operations on the images, we shall limit ourselves to a cursory exposition. Before everything else, the Canny edge detector was applied. For the scale parameter we chose the default value, $\sigma = 1.0$. The output of the edge detector is a binary image, as required by the subsequent MATLAB routines for the $\oplus$ and $\ominus$ operations.

In our implementation, we used twice the value of the optimiser for the low and high threshold in the Canny detector as returned by the MATLAB implementation of the

Canny edge detector. This stabilised the performance of edge detection at the loci of the bounding box to be determined.

The upper row of panels in Figure 6.2 illustrates the previously made points. The leftmost panel shows the unprocessed signature of a ship from class A2, observed at a range of 10.5km. In this and later plots, ranges and azimuths correspond to the horizontal and vertical axis, respectively. The two panels to the right show, respectively, the result of applying the Canny edge detector with $\sigma = 1.0$ with the low and high threshold values calculated by MATLAB, and the same with those thresholds doubled, showing a much better rendering of the edges. With regard to the preceding discussion, it should be remarked that no problems with edge streaking occurred. Presumably, this was due to the coarse pixelisation of the images. Depending on the size of the ship and its range, the values of $\mathrm{dim}_{\mathrm{az}}$ and $\mathrm{dim}_{\mathrm{rg}}$ rarely exceeded a value of about 80.

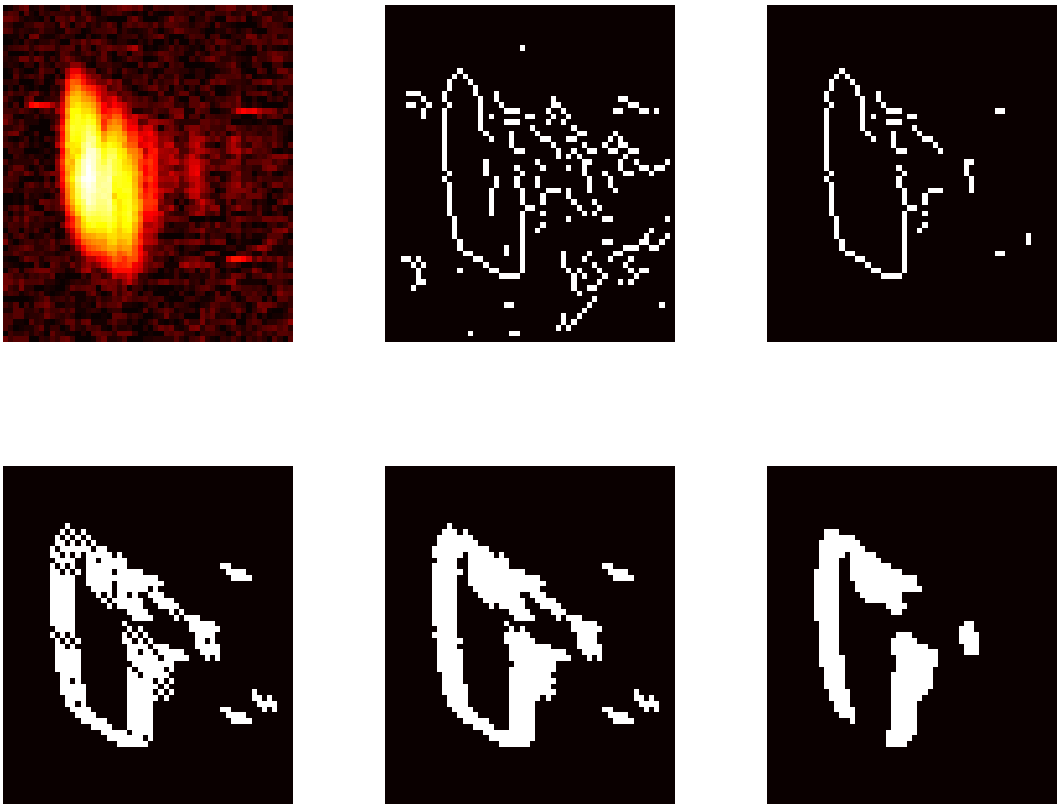

Figure 6.2: Panels show, from top to bottom and left to right in a row-wise fashion: The unprocessed signature of a ship from class A2 at range 10.5km, and the morphological operations on the binary version of that image, as described in Subsection 6.2.1.

In the next step, we used erosion operators in the horizontal (range) direction, and subsequently, a dilation by 3-pixel diagonal lines oriented at angles of 45° and 135°. The output

of this procedure can be seen in the fourth panel (bottom left) of Figure 6.2. Having thus obtained more 'solid' edges, the MATLAB filling operator `BWdfill` (with the 'holes' specification) was applied. This step is illustrated in the fifth panel. Another erosion operation, by a horizontal three-pixel line element, removed further clutter. Finally, the bounding box was determined as the minimum rectangle which enclosed the ship's outline, calculated in the illustrating example from the image in the last panel of Figure 6.2. The boundary regions were extracted and retained for later use; see Subsection 6.2.2.

It should be remarked that the morphological operations could likely be simplified, while maintaining or even improving overall performance, by using *linear* processing as a prelude. This could include wavelet methods in the spirit of Subsection 6.2.2, or (to circumvent the difficulties that led to the present order of pre-processing steps) the use of median filtering. Both of these alternatives appear to be interesting avenues for future research.

Finally, we should remark that we have also experimented with other non-linear approaches for segmentation, especially random-walk methods as described in Smolka and Wojciechowski (2001). While these may have some advantages in that they use the full data rather than a binary caricature of it, and require fewer parameters, their considerably higher computational expense recommended against using them in this application, where the need for online calculation is acute.

6.2.2 Wavelet Shrinkage

From the sizeable literature, as well as preliminary experiments for example with Wiener filtering, it appeared that wavelet methods were the most attractive for improving the SNR. To address similar tasks in one, and also to a fair extent in two, dimensions, Donoho, Johnstone and Silverman introduced and developed in a series of papers (e.g. Donoho and Johnstone, 1994, and Johnstone and Silverman, 1997) the now well-established concept of *wavelet shrinkage*, which operates by:

- Taking the Discrete Wavelet Transform (DWT) of the input image,
- Applying a thresholding rule to the resulting wavelet coefficients, and
- Taking the inverse DWT to render the reconstructed or *denoised* output image.

The composition of these three operations yields a denoising operator, which will be denoted $\mathcal{D}$, on the space of signals or images. In the rest of this subsection we discuss how to specify the parameters of $\mathcal{D}$.

With regard to the first two of the above points, there is considerable freedom left to the data analyst. The most important issue, however, arises with the choice of threshold. This should actually be a *vector*-valued quantity, so as to account for two features, the first of which is known to be prevalent in most problems in image and signal processing:

- As the resolution level of the DWT increases (corresponding to increasingly fine details), the SNR of the coefficients of that level decays.
- There is considerably more information contained in the refinement resolution channels for range than for the azimuth dimension, a feature that was clear in advance (and which can also be gleaned from Figure 6.3).

Before discussing our adaptation of the theory to the ship data, in the next section we give a brief review of the pertinent methodology of orthonormal wavelets. A strong reason for the use of orthonormal wavelets was the availability of a mature and well-tested software package to perform the computations.

Orthogonal Wavelets for Images

We begin with notation, closely aligned to that used in the papers cited earlier in this subsection. An implementation of the methods was provided by the authors cited there, together with collaborators, with the `WaveLab802` package for MATLAB 5.x and above, available from the URL

$$\text{http://www-stat.stanford.edu/}\tilde{}\text{wavelab/}\,,$$

referred to below as 'WaveLab.' The present spatial nature of the data required some alterations to that code, which were generally straightforward. We proceed with a brief account of methodological background, and use the notation (x, y) rather than $(x^{(1)}, x^{(2)})$ to facilitate notation. Suppose that a single radar signature is represented as a realisation of the following statistical model:

$$Z_{ij} = f(X_i, Y_j) + \epsilon_{ij}\,, \quad i, j = 1, \ldots n\,, \tag{6.2.1}$$

where the error variables ϵ_{ij} are drawn from some noise process (note that a nontrivial correlation structure of this process is thus permitted), and $n = 2^J$ for some $J \in \mathbb{N}$, implying that J represents the number of levels in the orthonormal wavelet decomposition. The case where the input image (i.e., the output of the procedures described in

Subsection 6.2.1) is of arbitrary rectangular or non-dyadic dimensionality may be treated by zero-padding.

Let $W = (w_{ij}) = \mathcal{W}(Z)$ denote the orthogonal wavelet transform of the input signal (image) $Z = (Z_{ij})$. For example, this could be the periodised wavelet transform implemented as the function `FWT2_PO.m` in WaveLab, using a Daubechies filter of order 4 as the quadratic mirror filter. A small filter order seems adequate in view of the coarse pixelisation and often low SNR in the images that are to be enhanced. We will also need the notation $W(A, B) = \{w_{ij} : i \in A, j \in B\}$ for $A, B \subseteq \{1, \ldots, n\}$. Arrangement of coefficients w_{ij} follows standard conventions, with the low-pass residual towards the top left, and the refinement coefficients shown as 'wedges' towards the lower-right corner. Figure 6.3 shows the output of the decomposition of the output image of Figure 6.2. For an illustration of the method applied to 'folklore' images in the literature, see Mallat (1999).

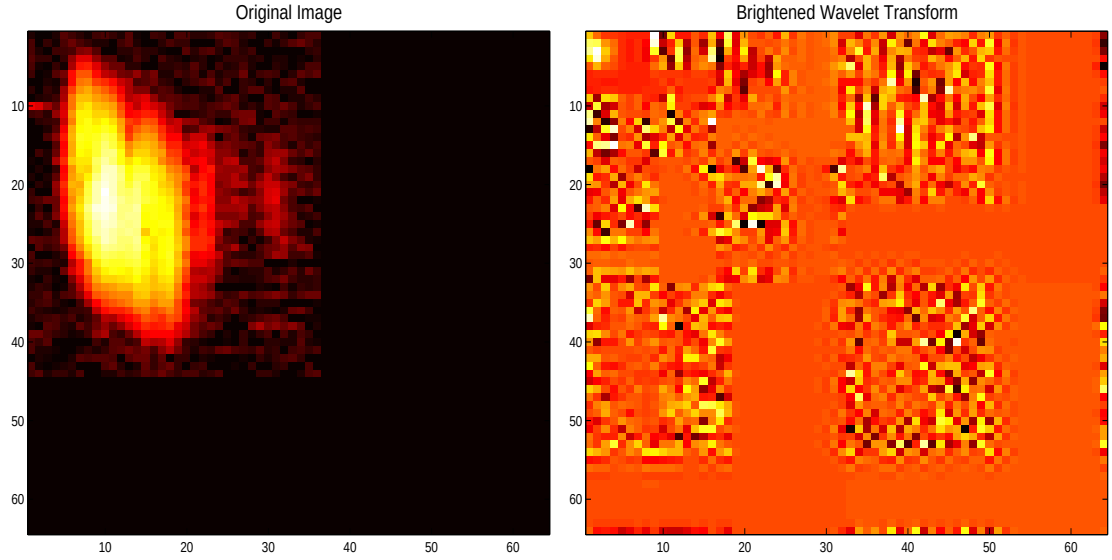

Figure 6.3: The left panel shows the cropped image from Figure 6.2 after zero-padding to the next-highest dyadic dimension (64×64). The right panel shows the result of applying a Daubechies wavelet transform, using three resolution scales. Coefficients were brightened for better visibility.

Since we are dealing with *spatial* rather than linear data, the multiresolution cascade comprises three channels instead of only one. Specifically, if we let $\mathrm{dyad}(j) \equiv \{2^j + 1, \ldots, 2^{j+1}\}$ for $j \in \mathbb{N}_0$, then the decomposition at each level j, with $L \leq j < J$, involves

the three sets of coefficients $\{w_j^H, w_j^V, w_J^D\}$ with

$$\begin{aligned} \{w_j^H\} &= W(\mathrm{dyad}(j-1), \mathrm{dyad}(j))\,, \\ \{w_j^V\} &= W(\mathrm{dyad}(j), \mathrm{dyad}(j-1))\,, \\ \{w_j^D\} &= W(\mathrm{dyad}(j), \mathrm{dyad}(j))\,, \end{aligned} \tag{6.2.2}$$

and the remaining elements for the case $L = 0$ are called the *low-pass residual* of Z. In equation (6.2.2) and later, superscripts H, V, D are used to indicate the relation of these sets of coefficients to what could be called, more informally, the 'horizontal,' 'vertical,' and 'diagonal' refinement details of the image. The coefficient sets $w_j^{\{V,H,D\}}$ with $j < L$ are usually thought of as being predominantly exhibiting signal rather than noise characteristics, hence they are not thresholded. At the outset of the analysis, the so-called *depth* of the decomposition was chosen as a maximum of five, namely $L = \max(J-5, 1)$. However, not all of the considered channels were thresholded later on.

To choose a threshold vector, we apply methods employed first in the context of correlated noise by Johnstone and Silverman (1997); see also Coifman and Donoho (1995). Before we proceed, we briefly comment on our choice of threshold function; see Section 1.1 for the definition of thresholding rules. We used the rule η_{Hard}; the *soft* thresholding rule η_{Soft} gave less satisfying results. This can be made plausible by observing that the predominant interest here is not in visual appearance, but in L^2 (near-)optimality (in the Bayesian risk setting as described in the aforementioned papers), which is known to recommend η_{Hard} over η_{Soft}. From our experiments, similar remarks apply to some continuous compromise between η_{Hard} and η_{Soft}. This includes raised cosine thresholding, which has gained some popularity in image denoising in other settings; see for example Kingsbury (1999).

Choice of Threshold

For the problem of setting the threshold, Johnstone and Silverman (1997) discussed the relative merits of two approaches. Both of these require preliminary estimation of noise standard deviation $\sigma_j^{\{V,H,D\}}$ in each decomposition channel. In order to do so, and following the suggestion of Johnstone and Silverman (1997), we used a robust estimator based on the median absolute deviation (MAD) from zero, defined earlier at (2.2.3):

$$\hat{\sigma}_j^{\{H,V,D\}} = \mathrm{MAD}\big(\{w_j^{\{H,V,D\}}\}\big) \,/\, 0.6745\,.$$

Note that robustness is essential because signal will generally still be present, if only sparsely, in the channels considered. Due to the decorrelation properties of the wavelet transform, the latter assumption is much more tenable in the wavelet domain, even if the noise was moderately correlated in the spatial domain (see Johnstone and Silverman, 1997). As mentioned earlier, the routines that are available in the WaveLab package were modified so as to take the two-dimensional nature of the index set into account. In calculating the MAD, this is simply meant to string out the data as a vector. Less trivial changes applied to the routines used below, but these were evident enough.

The threshold methods that we shall consider shortly all suffer from the problem that they are quite far from being conservative. Hence, we used the estimate of $\hat{\sigma}_j^{\{H,V,D\}}$ that was the maximum of the four estimates that were calculated from the (up to) four boundary regions that were identified as pure noise in the preceding segmentation step (cf. Section 6.2.1). Jansen and Bultheel (1999) proposed a generalised cross-validation criterion to obtain the threshold value that was found to perform well enough, but was still inferior in comparison to the approach discussed here.

The assumption of Normally distributed wavelet coefficients of the noise also underlies the first threshold definition, through a consideration based on extreme-value theory:

$$\lambda_{j,\mathrm{Univ}}^{\{H,V,D\}} = \sigma_j^{\{H,V,D\}} \cdot \sqrt{2 \log n_j^{\{H,V,D\}}} \,,$$

where $n_j^{\{H,V,D\}} = \mathrm{card}\big(\{w_j^{\{H,V,D\}}\}\big)$, and 'Univ' is the customary label 'universal' for this threshold. In other contexts this threshold is known to be rather conservative, in that it tends to yield larger estimates than the second rule to be described. However, this proved to be an advantage in the context of denoising the ship data.

The second rule draws strongly on the formulation of the thresholding problem in terms of L^2 risk. We content ourselves with the definition; the detailed derivation of this result may be found in the aforementioned papers, and the references cited therein. Suppressing the superscripts V, H, D here and below for ease of notation, the unbiased risk criterion for data $w_j = \{w_{jk}\}$ is

$$\hat{U}_j(t) = \hat{\sigma}^2 \, n_j + \sum_k \big\{ (w_{jk}^2 \wedge t^2) - 2 \, I \, (|w_{jk}| \leq t) \big\} \,,$$

where $a \wedge b = \min\{a, b\}$. Define

$$\hat{t}(w_j) = \mathrm{argmin}_{0 \leq t \leq \sqrt{2 \log n_j}} \, \hat{U}_j(t) \,.$$

The SURE (an acronym for Stein's Unbiased Risk Estimate) threshold is then obtained as follows:

$$\lambda_{j,\mathrm{SURE}} = \hat{\sigma}_j \cdot \hat{t}\big(w_j/\hat{\sigma}_j\big)\,, \qquad L \le j < J\,.$$

Before going further, in this paragraph we explain a further expedient that is applied to the data. This consists in calculating the threshold not from the original image itself, but from an average over a total of K^2 circulant shifts in azimuth and range directions. Specifically, with

$$\big(\mathcal{S}_{k_1,k_2}\, Z\big)_{ij} = z_{i_*,j_*}\,, \quad 1 \le k_1, k_2 \le K\,,$$

where $i_* = (i+k_1) \bmod n$, $j_* = (j+k_2) \bmod n$, we consider the *cycle-spinned* image, with 'TI' standing for (the approximately fulfilled property of being) 'translation invariant':

$$Z_{\mathrm{TI}} = K^{-2} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \mathcal{S}_{-k_1,-k_2} \circ \mathcal{D} \circ \mathcal{S}_{k_1,k_2} Z\,.$$

We adopted the value $K = 8$ that was also used in the WaveLab implementation. Complete translation invariance would be achieved by letting $K = \max\big\{\dim_{\mathrm{az}}, \dim_{\mathrm{rg}}\big\}$, but the less costly approximation was found preferable. By the cycle-spinning procedure, the tendency of the ordinary wavelet transform to produce spurious (Gibbs) phenomena in regions of abrupt changes is ameliorated. In addition to the above references, these issues are addressed by Coifman and Donoho (1995).

As remarked in Section 6.1, the information content in range direction is higher than in azimuth. That feature warrants the use of different depth in thresholding the channels w_j^H and w_j^V. (Recall that depth is defined via the minimum resolution level considered, denoted L in Subsection 6.2.2.) This motivated our choice of two and three for the respective values of depth in azimuth and range. In addition, the channel w_j^D is treated in the same fashion as w_j^H (depth three). In each channel, bearing in mind the properties of $\lambda_{j,\mathrm{Univ}}$ and $\lambda_{j,\mathrm{SURE}}$ mentioned earlier, we applied the threshold $\lambda_{j,\mathrm{Final}} = \max\big\{\lambda_{j,\mathrm{Univ}}, \lambda_{j,\mathrm{SURE}}\big\}$. It seems possible to justify these choices by non-purely data analytic means, as details that are physically distant less than 5m in range can not be unambiguously detected by the radar (the azimuth bin width depends on range). We shall not dwell on such issues here. Suffice it to say that the study of the connections between wavelet decomposition and physical parameters of the radar is an interesting avenue for further work.

The wavelet denoising carried out in this manner was expected to sufficiently suppress noise so as to complete the segmentation (cf. the corresponding remark in Sub-

section 6.2.1). To this end, a rather simple procedure was used, operating as follows. For each azimuth bin $\left\{ Z_{i,j_0}\,, 1 \leq i \leq \dim_{\mathrm{az}} \right\}$ represented by the column index j_0 with $1 \leq j_0 \leq \dim_{\mathrm{rg}}$, the intensities were arranged in ascending order: $Z_{(1),j_0} \leq Z_{(2),j_0} \leq \ldots \leq Z_{(\dim_{\mathrm{az}}),j_0}$, so that the peaks of the scatterers which corresponded to the object were in the bottom row of the matrix $(Z_{(i),j})_{ij}$ of order statistics. The index j_0 was flagged as belonging to a signal scan if it satisfied the following conditions:

$$\begin{aligned} &Z_{(\dim_{\mathrm{az}}-1),j_0} > 0.9\, Z_{(\dim_{\mathrm{az}}),j_0}\,, \qquad Z_{(\dim_{\mathrm{az}}-2),j_0} > 0.9\, Z_{(\dim_{\mathrm{az}}-1),j_0}\,, \\ &Z_{(\dim_{\mathrm{az}}-3),j_0} > 0.85\, Z_{(\dim_{\mathrm{az}}-2),j_0} \quad \text{and} \quad \sum_i Z_{i,j_0}^2 > 0.2 \max_j \Big\{ \sum_i Z_{ij}^2 \Big\}\,. \end{aligned} \tag{6.2.3}$$

The endpoints of the support of the ensuing zero-one vector of length $\dim_{\mathrm{rg}}$ was then taken as the final estimate of the range extension of the pixels. Similar computations were also applied to determine the azimuth extension, however in a much simplified version, as the only condition retained from the above was the fourth one regarding the gain/maximum gain ratio. Because the azimuth dimension of the bounding box represents only little information further to the strength of the peak scatterer, this measurement was not used for the analysis in Section 6.3.1.

The nine panels in Figure 6.4, to be read from the top left in a row-wise fashion, illustrate the pre-processing scheme from Subsections 6.2.1–6.2.2 in its application to the image of a ship from class A1. The second and third panels show the noisy border regions that were identified in the step described in Subection 6.2.1. Panel four shows the cropped image, zero-padded to the next binary dimension which equals in this case $n = 32$ for each axis. The next two panels show a spikes plot of the amplitudes of the wavelet data. A similar style of display is given by Donoho and Johnstone (1994), and is implemented for one dimension in WaveLab as `IMJPlotWaveCoeff.m`. The version developed for this report uses the maximum (black spikes) and minimum values (blue spikes) of the wavelet coefficients, taken over the appropriate dimension, in order to condense the two-dimensional set of coefficients into a linear collection. The seventh panel shows the denoised version of the third, the eighth is the result after dropping the zero padding, and the last panel is obtained after applying the steps described around equation (6.2.3).

6.2.3 Registration

The Karhunen-Loève analysis in Subsection 6.3.4 requires a unified format for each datum, which necessitates interpolation to a norm pixel size in both azimuth and range.

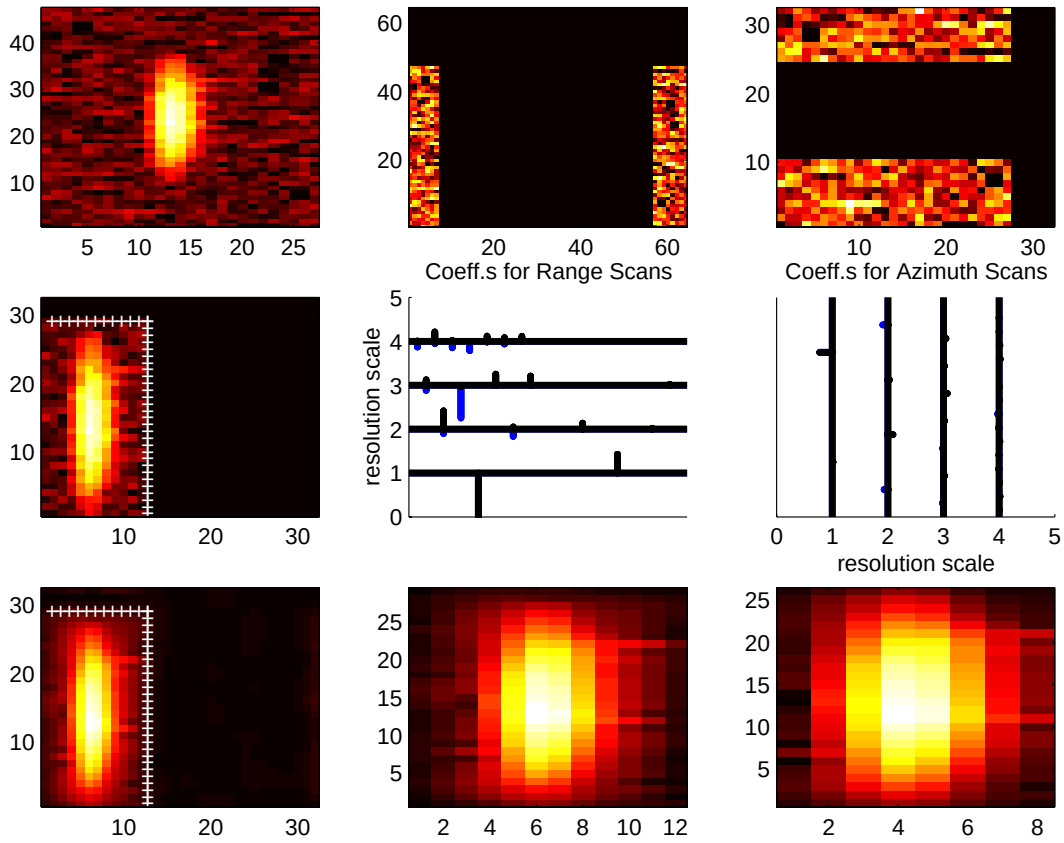

Figure 6.4: Illustration to the image pre-processing steps, as described in Section 6.2.2.

For those parameters we chose $\dim_{\text{az}} = 35$ and $\dim_{\text{rg}} = 40$, so that after stringing out the data as row vectors, the result was an $n \times 1400$ dimensional matrix corresponding to the radar signatures. Here and in what follows, n denotes the total sample size, to be given in Section 6.3.4, and which should not be confused with the use of n in (6.2.1).

An important issue that should be explored at this stage is the possibility of aligning a given image to a reference image. Even without the 'submersion' phenomenon that affects the strength, and in fact the sheer presence, of scatterer peaks as range increases and curvature of the Earth is no longer negligible, changes in the angle of view could lead to erroneous conclusions about the number of such peaks. Hence it is hard to find so-called *landmarks* that would otherwise be very useful for registration; see for example Ramsay and Silverman (2002, Chapter 4, pp. 57ff). From our experiments it appeared that if the data are split, according to their angle of view, into two extremes of 0° and 180°, thereby creating two sub-classes, classification does not significantly improve and in fact often deteriorates. This might be due to the amount of training data becoming too small. If the database was expanded, one could naturally consider using more than

two aspect angles. It is pertinent to note that the change in aspect angle leads to a *sinusoidal* scaling of the image, where the best recognition is possible for the 0° and 180° angles.

In view of these difficulties, the data from the first paragraph were augmented by the estimated pixel extension along the range axis, the estimated range of the vessel, and the maximum speed that was observed within the training set. It is the last of these measurements that is of crucial importance in distinguishing the superclasses A1 and A3, notwithstanding the fact that difficulties with distinguishing these two groups remain; see Figure 6.7.

Problems concerning curve registration were discussed by Ramsay and Silverman (1997, Chapter 5, pp. 67ff) at some length. We refer to that monograph, and the references cited therein, for a comprehensive account of issues not discussed here.

6.3 Case Study

In this section we combine the processing techniques from the previous section with tools from functional data analysis to tackle the navigation radar dataset. Some comments on the raw dataset are given in Subsection 6.3.1. The question of how to obtain basis functions that subsume both the signal, and the measurements of discrete parameters such as range and speed from the preamble to each datafile, is addressed in Subsection 6.3.2. In order to select the basis functions, in Subsection 6.3.3 we employ a criterion that, although unlikely to be really new, seems to have been used rarely under comparable circumstances. Classification results are presented in Subsection 6.3.4. Concluding remarks and future problems relating to the specific data of the case study, but probably applying in a considerably wider context, are given in Subsection 6.3.5. A flowchart of the processing and classification scheme as a whole is given in Figure 6.5.

6.3.1 The Raw Dataset

The dataset on which the subsequent analysis was based is shown in Table 6.1, and is of the same low-resolution type as analysed in Robinson (1996), Inggs and Robinson (1999), and Gibbins, Gray and Dempsey (1999). All data were sampled at a pulse recurrence frequency of 2kHz and at a pulse length of 80ns, and the '$\geq$' signs were due to the combination of various vessels for which no own names were available in the preamble of the dataframe. The shape of the pulse was approximately (but not exactly) rectangular.

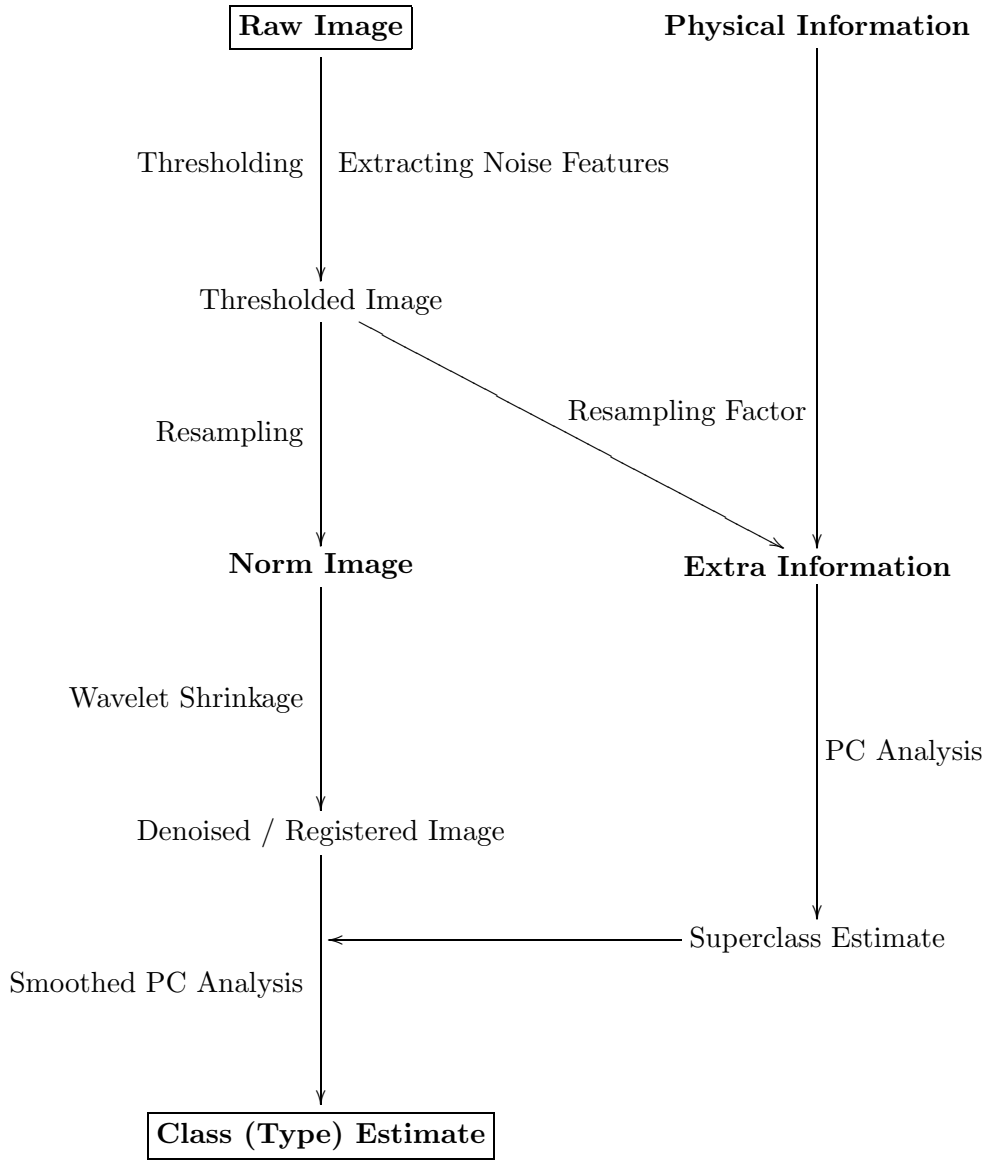

Figure 6.5: The processing and classification scheme. ('PC' stands for 'Principal components.')

Data on vessels observed at ranges below 1500m were excluded from the analysis, due to the distortion effects of *sensitivity time control* (cf. Acker, 1988, pp. 8–9), which is an inherent part of the radar technology to improve clutter performance. The class `s4'`, which mainly comprised unidentified vessels, was likewise excluded.

Superclass / Vessel Type	# Vessels	# Dataframes
A1 = s1	9	920
A2		
s2	9	507
s3	5	183
s4	6	380
s4'	($\geq$) 5	288
A3		
s5	($\geq$) 5	154
s6	($\geq$) 5	244
s7	($\geq$) 6	679
Total	($\geq$) 50	3355

Table 6.1: The raw data.

6.3.2 Determining the Basis Functions: Regularisation

It was found efficacious to perform further smoothing through imposing a roughness penalty on the basis functions, as described in Ramsay and Silverman (1997, Chapter 7, pp. 111ff). Here it may suffice to give a very condensed account. The structure of this penalty is that of the differential operator D^2 (i.e., the second derivative of a functional interpolant of the data), conveniently computed in a Fourier representation. The matrix manipulations can then take advantage of the fact that the Fourier functions are the eigenfunctions of D^2. Control over the size of the penalty is exercised through a coefficient (multiplier) $\lambda \geq 0$, with the case $\lambda = 0$ corresponding to an ordinary principal components analysis. The value of λ was chosen differently for each superclass, but typically of the order of 10^{-4}, or even smaller. It is also possible to choose λ by cross-validation, but this requires very substantial computational effort. The sensitivity of classification results with respect to the choices of λ for A2 and A3 did not seem to be high. The implicit assumption of stationarity could be seen as a theoretical disadvantage of using the Fourier transform. However, for the size of our dataset this drawback was more than outweighed by the computational cost that is associated with alternative representations, as these would involve densely-populated matrices of high order. We refer to Ramsay and Silverman, 1997, pp. 117ff, with the following corrigendum by the authors: In equation (7.10) and the next two displayed equations, $\mathbf{V}$ should be replaced by $\mathbf{JVJ}$; and in stage 2 of the algorithm, c_i should be replaced by $\mathbf{J}c_i$ on both occasions.

Of further help in classifying the ship types were the individual recorded speeds. Appending the speed data, the ith datum was represented by a vector $(x_i, v_i) \in \mathbb{R}^{1401}$ where the

last component v_i could be called, following Ramsay and Silverman (1997), the *numerical part* as opposed to the *functional* part x_i, i.e. the vectorisation of the image. In order to allow the meaningful analysis of the joint variability of these two parts (recall that the Karhunen-Loève expansion takes recourse to the singular value decomposition of the grand covariance matrix, which is not invariant to scale of its coordinates), one needs to first calibrate the respective scales. To this end, we multiplied v_i, for all $i = 1, \ldots, n$, by the three respective scaling constants $C_{\{\mathsf{A1},\mathsf{A2},\mathsf{A3}\}}$. Ramsay and Silverman (1997, p. 130) suggested the choice ($\bar{x} = \mathrm{card}(\mathsf{A1})^{-1} \sum_{k\in\mathsf{A1}} x_i \,, \bar{v} = \mathrm{card}(\mathsf{A1})^{-1} \sum_{k\in\mathsf{A1}} v_i$)

$$C^2_{\mathsf{A1}} = \frac{\sum_{k\in\mathsf{A1}} \|x_k - \bar{x}\|^2}{\sum_{k\in\mathsf{A1}} \|v_k - \bar{v}\|^2} \,, \tag{6.3.1}$$

and analogous definitions for C_{A2} and C_{A3}. There is, however, the problem that while $\|\cdot\|$ in the denominator denotes the ordinary Euclidean norm, in the numerator the norm is to be taken in the functional (L^2) sense. Since the length of the definition interval of the images x_i is ill-defined in the present case, and because the x_i have furthermore been up- or down-sampled by varying rates, questions are raised in justifying the use of (6.3.1) on its own. Also, as in the related discussion in Ramsay and Silverman (1997, *ibid.*), there are questions as to which normalisation should be used in (6.3.1). Experiments showed that by using an additional multiplier of $(\dim_{\mathrm{az}} \cdot \dim_{\mathrm{rg}})^{-1}$ in (6.3.1), and using half of the ensuing value of $C_{\{\mathsf{A1},\mathsf{A2},\mathsf{A3}\}}$, we obtained good approximations to the 'optimal' choices of these values, in the sense that they yielded the smallest error (i.e., misclassification) rates.

6.3.3 Determining the Basis Functions: Subset Selection

The techniques discussed previously yield a realisation of the random vector of scores $\{\Xi_{ir}\}$ for the ith observation, where r is the order of the coefficient in the expansion (A.5.2), and we also substitute Ξ_{ir} for ξ_r there. Two major issues arise in connection with the dimension reduction technique that is based on the Karhunen-Loève expansion:

- How should the integer M be chosen such that the vectors $\{\Xi_{ij},\, 1 \le j \le M\}$ are (close enough to being) optimal for classification purposes? If M is chosen too large then the phenomenon of *overfitting* will in fact lead to a deterioration of classification rates.

- As a point subordinate to the previous one, it is *a priori* reasonable to allow for the possibility that noise in (realisations of) the scores Ξ_{ij} does *not* consistently

become dominant over signal as r increases. Hence, we are interested in a method that is capable of selecting a subspace $\text{span}\{\psi_{j_1}, \ldots, \psi_{j_{M'}}\}$, with $j_1 < j_2 < \ldots j_{M'}$ and $j_\nu > j_{\nu-1} + 1$ for some ν.

In order to address the first of these issues, Hall, Poskitt and Presnell (2001) proposed cross-validatory estimation of M. In the context of the present data, however, this approach was found not to be satisfactory when used by itself. From preliminary trials, it was observed that the number of basis functions to consider is generally small, and indeed we found it sufficient to retain no more than ten basis functions. The eventual choice of the basis was therefore left with the second of the above problems, which was tackled by a method that was simple yet effective. Specifically, assume that an upper limit on the value of M, for example $M = 10$, is given. In the sequel, a generic subset of $\{1, \ldots, M\}$ will be denoted by calligraphic letters. For two class assignments (functions of the test data) $\mathcal{A}$, $\mathcal{B}$, define

$$c_{kl}(\mathcal{A}, \mathcal{B}) = (\text{number of observations that are in class } k \text{ under assignment } \mathcal{A} \text{ which fall in class } l \text{ under assignment } \mathcal{B}).$$

If $\mathcal{A}$ represented the true class assignments, this would yield the standard confusion matrix.

For $\mathcal{S}' \subseteq \mathcal{S}$, let $\mathcal{A}_{\mathcal{S}'}$ denote the (self-)classification on the training set when using the subset $\{\psi_\nu,\, \nu \in \mathcal{S}'\}$ of the Karhunen-Loève basis. For brevity, write $c_{kl} = c_{kl}(\mathcal{S}, \mathcal{S}')$, and let $0/0 \equiv 0$. Let

$$\texttt{FLOW}(\mathcal{S}, \mathcal{S}') = \sum_{\binom{k}{l}} \mathcal{I}_{kl}(\mathcal{S}, \mathcal{S}')\, I\left\{c_{kl} + c_{lk} \geq n_0\right\}\,, \tag{6.3.2}$$

where $\binom{k}{l}$ stands for summation over all combinations of size l drawn from k elements, $n_0 \in \mathbb{N}_0$ is a threshold (a void condition for $n_0 = 0$), and

$$\mathcal{I}_{kl} = (c_{kl} + c_{lk})^{-1} \left(\frac{c_{kl}}{c_{kl} + c_{lk}}\right)^{c_{kl}} \left(\frac{c_{lk}}{c_{kl} + c_{lk}}\right)^{c_{lk}}\,.$$

Note that the last expression is somewhat similar to the definition of the (log-)likelihood at (3.3.1). For the two superclasses A2, A3 which split up into several types, we applied the FLOW criterion in order to reject basis functions in the range $m = 4, \ldots, M$ with $M = 10$. With somewhat more simplicity than suggested by (6.3.2), we used $\mathcal{S} = \{1, \ldots, M\}$, $\mathcal{S}' = \mathcal{S} \setminus \{m\}$, $n_0 = 2$, and rejected the basis function of index m if $\texttt{FLOW}(\mathcal{S}, \mathcal{S}') < 2.5 \cdot 10^{-3}$.

6.3.4 Classification

In order to divide the dataset described in Section 6.3.1 into the shiptypes `s1`–`s7`, we used the sequential pre-processing techniques described in Section 6.2. Apart from the output images, normed by resampling to dimension $\dim_{\text{az}} \times \dim_{\text{rg}} = 35 \times 40 = 1400$, the only features used were related to the physical measurements. Namely, in distinguishing the three superclasses, we utilised:

- the estimated ship's pixel size in range,
- the ship's range, and
- the maximal speed observed in the training set.

Figure 6.6 shows pairwise scatterplots of these measurements for the original data. These are to be read in pairs, the upper two panels corresponding to the training data, which in this case comprised $n_{\text{train}} = 2230$ ships, and the test data with the corresponding number $n_{\text{test}} = 815$. The speed is given in knots, with one knot equalling 1.852 kilometres per hour. Note that for illustrative purposes, and because of the significantly larger amount of data at small distances, the ranges are displayed on a logarithmic scale. A listing of the training and the test datasets for each type is shown in Table 6.2.

Using the speed maxima instead of individually measured speeds in the classifier is beneficial for classification, notwithstanding the fact that the Gaussian assumption that underpins the QDA classifier (see the definition below) is clearly less tenable. Certainly, speed should not be regarded as a very reliable input, especially if the training sample used only a relatively short time window.

	Training Data		Test Data	
	#Vessels	#Data	#Vessels	#Data
`s1`	7	684	3	228
`s2`	5	290	4	223
`s3`	5	149	1	34
`s4`	9	303	2	77
`s5`	3	100	2	54
`s6`	3	121	3	117
`s7`	5	583	2	82

Table 6.2: Sizes of training and test datasets for the example in Subsection 6.3.4.

In order to distinguish the three superclasses, we used a standard (non-smoothed) principal components analysis applied to the above three measurements. The maximum

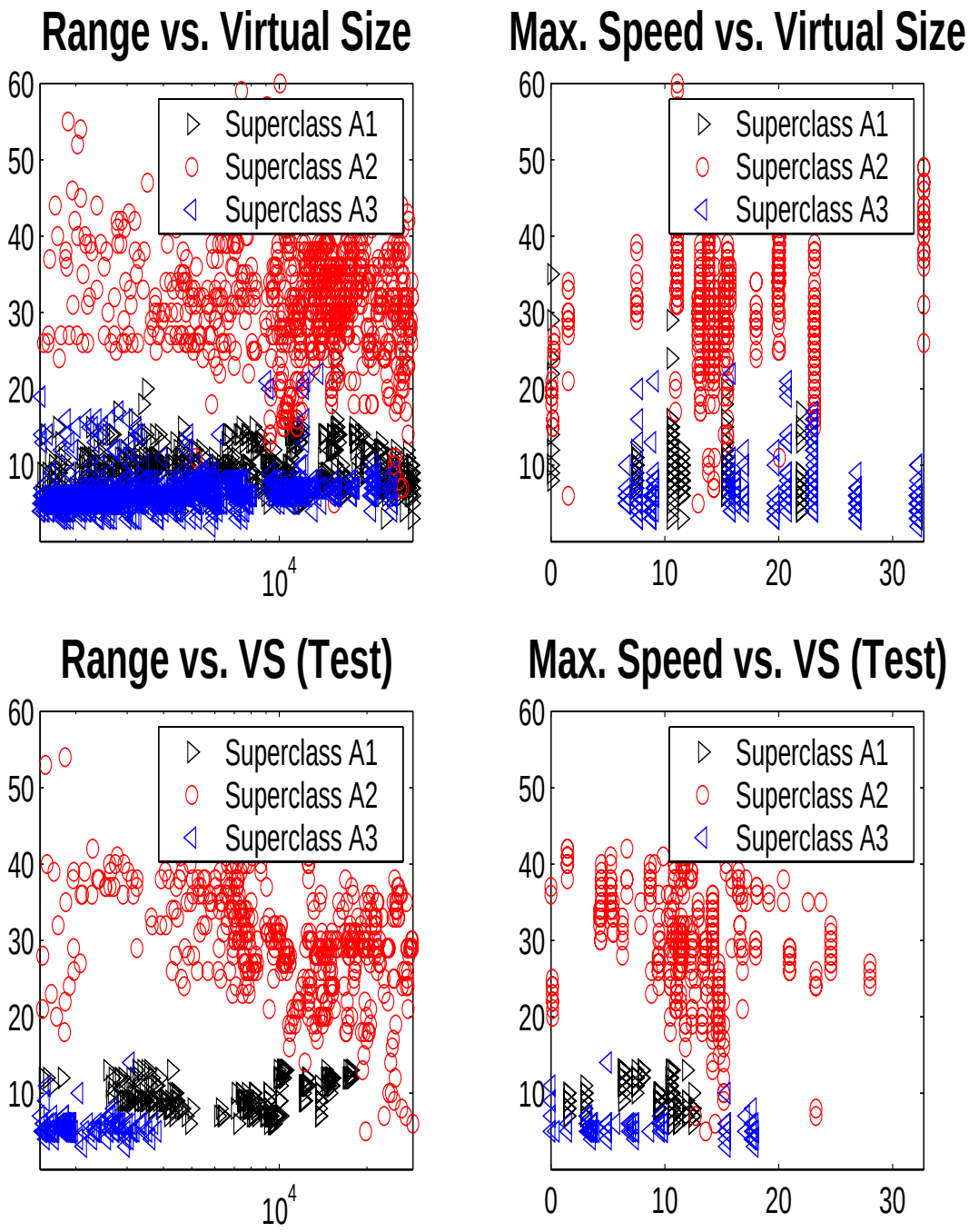

Figure 6.6: Scatterplot of physical measurements for the three superclasses.

number of basis functions that are calculable (namely, three), was used. We obtained good separation of the superclass A2 from the other two, and varying but generally satisfactory performance in distinguishing A1 from A3. An example output is shown in Figure 6.7; note that the main diagonal of the confusion 'matrix' in Figures 6.7–6.8 runs from the bottom-left to the upper-right corner.

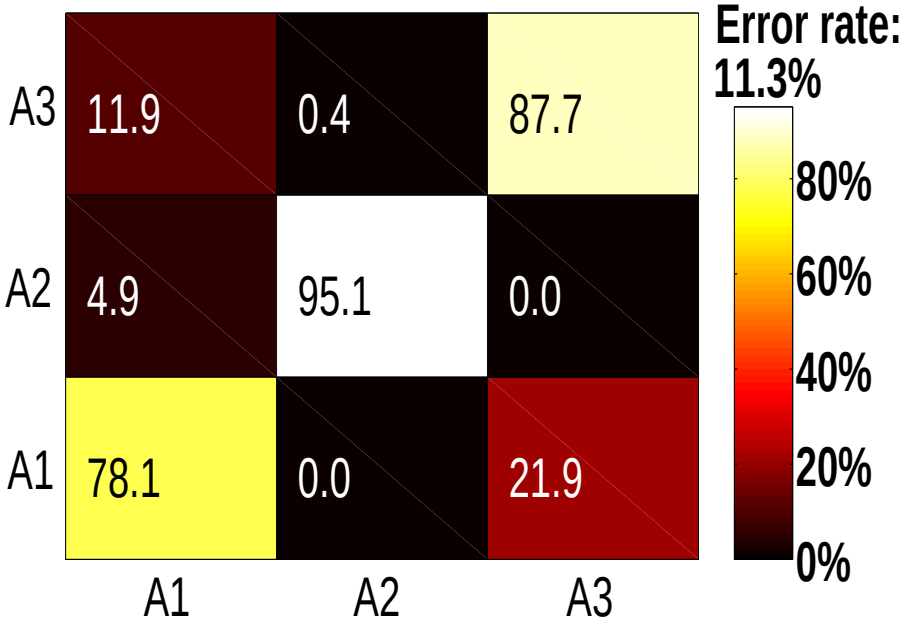

Figure 6.7: Confusion matrix for classification into the three superclasses.

It should be noted that improved error rates are likely achievable by using the original images, as well as the processed ones, in the analysis. This would mainly serve to ameliorate the occasional failures of the pre-processing techniques when determining the bounding box of very noisy data at distant ranges.

The training and test datasets were chosen according to the rules given below. It should be noted that the sensitivity of the results to the design of the train/test combination is considerable, but one would expect that this variability decreases as the database is enlarged.

- If there was more than one run on a vessel available, the number of test runs was chosen to be the largest integer less than a third of the number of runs.
- In the opposite case, the single run was assigned to the training or the test pool with probabilities 0.7 and 0.3 respectively. This implies that in particular, no data were discarded after selection for the training/test pools.

- In either of the above situations, upper bounds of 250 and 50 were imposed on the sizes of the training and test datasets, respectively.
- As there exists large fluctuation of the number of frames within the runs, the size of the test dataset could possibly exceed the size of the training data for a given ship type. This was remedied by interchanging these sets where necessary.

Figure 6.8 illustrates classification results obtained in this setup, obtained by quadratic discriminant analysis in the same way as in Hall, Poskitt and Presnell (2001). Recall that quadratic discriminant analysis estimates the class density for the lth type using a multivariate normal density. The mean and covariance matrix for the lth type are estimated by their sample analogues,

$$\bar{X}_{(l)}^{(m)} = n_l^{-1} \sum_{i=1}^{n_l} X_{(l),i}^{(m)} \qquad \text{and} \qquad \hat{\Sigma}_{(l)} = n_l^{-1} \sum_{i=1}^{n_l} (X_{(l),i}^{(m)} - \bar{X}_{(l)}^{(m)})(X_{(l),i}^{(m)} - \bar{X}_{(l)}^{(m)})',$$

respectively. The density estimate is then

$$\hat{f}_{(l),m}(x) \propto \det(\hat{\Sigma}_{(l)})^{-1/2} \exp\Big\{ -\frac{1}{2} \Big(x^{(m)} - \bar{X}_{(l)}^{(m)} \Big)' \hat{\Sigma}_{(l)}^{-1} \Big(x^{(m)} - \bar{X}_{(l)}^{(m)} \Big) \Big\}.$$

As noted earlier, the shipclass `s4'` from Table 6.1 was excluded from the analysis because it appeared to be too vaguely defined to permit being grouped separately; acronyms are as in Table 6.1. The respective sets of basis functions were chosen according to the rule described in Subsection 6.3.3, based on the `FLOW` criterion of (6.3.2). While only the first three basis functions were retained for the superclass A3, the ninth was in addition retained for the superclass A2. The inclusion of this basis function may show the presence of superstructure, even if only at a weak level.

The result seems to be representative of the error rates that were obtained in multiple trials, at least if very disproportionate training/test configurations obtained under the above conventions were disregarded. Note that due to varying weather conditions for separate runs, and in view of the fluctuations in the number of frames contained in a run, disproportionality is not straightforward to describe. These circumstances were often detected only after the experiment. The existence of otherwise distorted data further complicates a quantification of the frequency of occurrence of such unfavourable configurations. Sometimes the error rates could be as low as 30%, but on other occasions also in excess of 40%. For vessels for which multiple runs were available, the transfer of a training set to a test set worsened the results quite dramatically. This seems to support the conjecture that expansion of the database is crucial in order to obtain better, or at

least in practice more reliable, results with the Karhunen-Loève (as probably with any other) classifier.

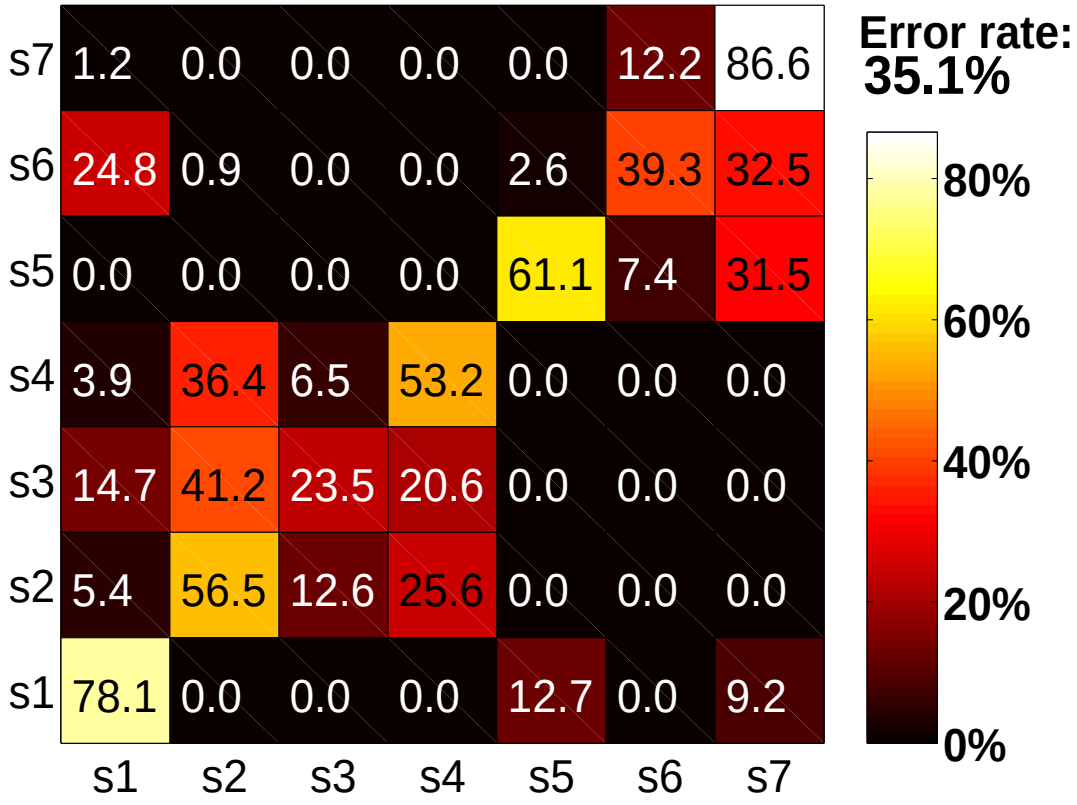

Figure 6.8: Estimated (mis)classification rates (row percentages) using quadratic discriminant analysis.

6.3.5 Conclusion and Further Research

The research in this chapter indicates that by dimension reduction via the Karhunen-Loève approach, it is possible to detect shape differences which characterise ship types. These differences are not obvious to the eye, even with knowledge of the true types which the letters `s1`–`s7` represent. Problems which arise from varying angles of view, as well as other sources, are felt quite strongly. Nevertheless, a relatively unsophisticated approach that performs images enhancement on each datum in turn, without taking sequential features into account, seems to be capable of reducing such distortions while retaining salient features for classification. At least, mingling a moderate proportion of poor-quality data into the training set does not seem to severely confound the basis functions or, by implication, the classifier.

Expanding the database seems to be an essential step in the implementation of more sophisticated methods which utilise the sequential aspect of the data. This applies to both the image processing, notably the registration, and the classification levels. But even with the present database, it seems worthwhile to attempt a more judicious use of the viewing angle information.

Appendix:
Miscellaneous Background Topics

A.1 Overview

In this appendix, we collect some background material from elementary differential geometry, probability theory and statistics that are instrumental for the results presented in Chapters 2–4 of this thesis. The style of the survey will be aimed not at completion, but rather at indicating how our results relate to existing work, especially regarding the links to the now well-developed area of empirical process theory. Because clarity is given preference over completeness, frequent references to the literature will be given.

The study of edge curves necessitates the use of concepts from elementary differential geometry, which are recapitulated in Section A.2. We turn to the problem of framing the sequence of edge estimators in rigorous probabilistic terms in Section A.3. Due to their appearance as limits of summation processes such as our likelihoods, in quite complex contexts such as here, Section A.4 is devoted to Gaussian processes. Computational treatment of Gaussian, and in fact more general L^2 processes, is greatly facilitated by the so-called *Karhunen-Loève expansion*, discussed in Section A.5. In Section A.6 we utilise theory of martingales indexed by classes of planar sets in order to address the convergence of stochastic summation processes, which are especially important in Chapter 4. Being a special case of M-estimation, the likelihood theory from Chapter 4 can be related to results from the wider class of M-estimators that derive from a stochastic process such as considered here, which are the subject of Section A.7. Finally, in Section A.8 we discuss some aspects of minimax theory, which has become the by far most important tool to assess and compare performance of edge estimators statistically.

A.2 Descriptors for Planar Curves and Surfaces

As pointed out in Section 1.2, we regard the edge detection problem as primarily relating to curves, rather than point sets. This enables the description of the original edge, as well as the estimators to be presented in Chapters 2 and especially 3 and 4 of this thesis, in terms of differential-geometric rather than set-theoretic notions. In describing our estimators, an important role is played by the intrinsic differential-geometric coordinate system, commonly (as in this section) referred to as the *gauge coordinate system*, which is determined by the local analytic properties of the curve as described in the following. The estimators proposed in Chapters 3 and 4 make essential use of the notion of a *ridge* in a planar surface. There are several plausible definitions for a ridge (see e.g. Hall, Qian and Titterington, 1992). In order to give a concise description of our estimator, and for

occasional use elsewhere, it is convenient to collect in the present section some notions from elementary differential geometry of planar curves and surfaces. More details on the subsequent topics may be found in do Carmo (1976), Hall *et al.* (1992), and Haralick (1983).

We start with the basic descriptors of a planar curve, for which we will use the letter $\mathcal{C}$. The curve $\mathcal{C}$ can be represented by a parametric equation,

$$u \mapsto x(u) = \big(x^{(1)}(u), x^{(2)}(u)\big)\,, \quad u \in [0,T]\,, \quad T < \infty\,. \tag{A.2.1}$$

The points $x(0)$ and $x(T)$ may be called the 'beginning' and 'end' of $\mathcal{C}$. Recall that $\mathcal{C}$ is a *Jordan curve* if $x(0) = x(T)$ and $x(\cdot)$ is injective on $[0,T)$, in which case any point on $\mathcal{C}$ may function as the beginning of $\mathcal{C}$ as well as its end. If $\mathcal{C}$ is a continuously-differentiable curve, its (*arc*) *length*,

$$\text{length}(\mathcal{C}) = \int_0^T \|\dot{x}(u)\|\,du = \int_0^T \big\{(\dot{x}^{(1)}(u))^2 + (\dot{x}^{(2)}(u))^2\big\}^{1/2}\,du\,, \tag{A.2.2}$$

is well-defined. (Here and in the sequel, the usual dot notation is used for derivatives.) Considering the upper integration limit in equation (A.2.2) as variable, denoted by w say, yields a strictly increasing function $\Gamma(w)$ on $[0,T]$. A special role is played by the *natural* parametrisation, which arises when $\Gamma(s) \equiv s$. In this case, the parameter s is related to an arbitrary parametrisation as at (A.2.1) via the equation

$$s = s(u) = \int_0^u \|\dot{x}(v)\|\,dv\,.$$

For any parametrisation, the tangent vector at parameter value u is given by $\theta = \theta(u) = \big(\dot{x}^{(1)}(u), \dot{x}^{(2)}(u)\big)/\|\dot{x}(u)\|$, and the corresponding normal vector $\theta^{\perp}(u)$ is obtained by rotating $\theta(u)$ counter-clockwise through $\pi/2$ radians, *viz.*, $\theta^{\perp}(u) = \big(-\dot{x}^{(2)}(u), \dot{x}^{(1)}(u)\big)/\|\dot{x}(u)\|$. The *curvature* of $\mathcal{C}$ is given by the formula

$$\kappa(u) = \frac{\begin{vmatrix} \dot{x}^{(1)}(u) & \ddot{x}^{(1)}(u) \\ \dot{x}^{(2)}(u) & \ddot{x}^{(2)}(u) \end{vmatrix}}{\|\dot{x}(u)\|^3}\,, \tag{A.2.3}$$

where $|\cdot|$ denotes the determinant. Formula (A.2.3) simplifies for the natural parametrisation since in that case $\|\dot{x}\| \equiv 1$. The sign of κ is determined by the (simplified) *Serret-Frenet equation*, $\theta(s) = \kappa(s)\,\theta^{\perp}(s)$. Usually we will only be interested in the magnitude of curvature, $|\kappa|$. Henceforth the qualifier 'absolute' is tacitly understood when referring

to curvature. The vectors θ and $\theta^{\perp}$ define the *gauge coordinates* of $\mathcal{C}$. The advantage of the gauge coordinate representation is that the description of points is invariant under rotation of the curve.

In cases where a representation of the curve in Cartesian coordinates, say of the form

$$\mathcal{C} = \big(t, f(t)\big)\,, \qquad f : [0,T] \longrightarrow \mathbb{R}\,, \tag{A.2.4}$$

or in polar coordinates:

$$\mathcal{C} = \big(r(\phi)\,\cos\phi, r(\phi)\,\sin\phi\big)\,, \qquad r : J \subseteq [0,2\pi] \longrightarrow \mathbb{R}_+$$

exists, the following respective versions of (A.2.3) may be used:

$$\kappa(t) = \big\{1 + f'(t)^2\big\}^{-3/2} f''(t)\,, \tag{A.2.5}$$

$$\kappa(\phi) = \big\{r(\phi)^2 + r'(\phi)^2\big\}^{-3/2} \big\{r(\phi)^2 + 2r'(\phi)^2 - r(\phi)r''(\phi)\big\}\,. \tag{A.2.6}$$

Because the curve estimators to be developed in Chapters 4 and 5 derive from surfaces in the form of ridge curves, the curvature of $\mathcal{C}$ may, under some extra conditions, be estimated through the partial derivatives of that surface. Using such two-dimensional operators in this context is appealing for three reasons, mentioned in Koenderink and Richards (1988) and which we rephrase as follows:

1. Estimation of curvature of a fault line in a smoothed surface is not equivalent to the two sequential steps of extraction of that line, and application of a smoothing method for curves. The former process is likely to better preserve the information on curvature.

2. The robustness to noise can be expected to be superior for an area-based operator.

3. From a psychophysical point of view, an area-based operator is more plausible as well: human vision appears to discern ridge curves in this manner.

With regard to item 1 in the foregoing list, it should be appreciated that the two mentioned procedures are not even equivalent if the smoothing parameters for the surface and the curve are chosen to be optimal. To illustrate this fact, recall that the optimal estimator of the mode of a density is *not* recovered by using the optimal bandwidth for the density estimate. The verdict from item 3 seems not to have been superseded by research since the time of writing of Koenderink and Richards (1988).

In this paragraph we rederive, under the mentioned extra conditions, the area-based curvature formulae. For simplicity, assume that the fault line has a representation as at (A.2.4), and for notational convenience we use xy notation for points in $\mathbb{R}^2$ until the paragraph containing (A.2.8). We now stipulate that the regression surface has a constant limit on one, say the '$-$' side, of $\mathcal{C}$,

$$g_* = \lim_{\substack{(x,y)\to(x_0,y_0)\\(x,y)\in\mathcal{R}_-}} g_-(x,y)\,, \qquad \text{for all } \big(x_0,y_0\big)\in\mathcal{C}\,. \tag{A.2.7}$$

Assumption (A.2.7), although restrictive, is often satisfied in practical applications where the image is obtained through thresholding at some level of the response variable. Our second assumption is that $g_y \equiv (\partial/\partial y)g$ has a continuous extension from $\mathcal{R}_-$ to $\bar{\mathcal{R}}_- = \mathcal{R}_- \cup \mathcal{C}$ and does not vanish on $\mathcal{C}$. Under these conditions, the image set of $\mathcal{C}$ can be expressed, in a sufficiently small neighbourhood of $\bar{\mathcal{R}}_-$, as the *level curve* with equation $g\big(x,y(x)\big) = g_*$. By implicit differentiation, it follows that

$$\begin{aligned} y' &= -\frac{g_x}{g_y}\,,\\ y'' &= \big(-g_{xx}g_y^2 + 2\,g_{xy}g_xg_y - g_{yy}g_x^2\big)/g_y^3\,, \end{aligned} \tag{A.2.8}$$

which in combination with (A.2.5) gives the announced alternative formula for the curvature of $\mathcal{C}$.

We now review briefly the basic differential geometric properties of surfaces, building on the previously introduced concepts. In an abstract context, such a surface is often also referred to as a (two-dimensional) *manifold*, customarily denoted by the letter $\mathcal{L}$, and defined on some compact domain $\mathcal{D}$. In the present instance, and for each $x \in \mathcal{D}$, the (first-order) gauge coordinates $u = u(x)$ and $v = v(x)$, often also called the first and second *principal directions*, are given by the normalised gradient vector, $m_0 = m_0(x)$ say, and its perpendicular $m_0^\perp$ parallel to the x plane, choosing any of the two possible orientations. Moving along $m_0^\perp$ on the surface, in an infinitesimal context, amounts to tracing a contour of the surface defined by g. The vectors $u(x)$ and $v(x)$ span the eigenspace of the Hessian of $\mathcal{L}(x)$. Moreover, the eigenvalues of $\mathcal{L}(x)$ represent the respective (local) curvatures of the planar curves that are obtained by restricting $\mathcal{L}$ to the projection of the one-dimensional affine subspace generated by either $u(x)$ or $v(x)$, on $\mathcal{D}$. Along these directions, curvature on the surface is respectively minimal and maximal.

We now give the definition of a ridge that suits our purposes.

Definition A.1. *Assume the same notation as above. A point* $(x, \mathcal{L}(x))$ *is a* ridge point

if $\mathcal{L}_u(x) = 0$ *and* $\mathcal{L}_{uu}(x) < 0$. *A* ridge line *is the projection of a ridge onto* $\mathcal{D}$. *An* antiridge (line) *is a ridge* (*line*) *in the surface defined by* $(-\mathcal{L})$.

Where a contour line meets a ridge line, it does so necessarily at a right angle. At the intersection point, the value of $\|m_0\|$ is minimal; this property relates to the fact that successive contour lines, as one moves away from a ridge in the perpendicular directions, have local maxima along these directions. A consequence of the previously-mentioned property is that by following a contour line, and (in our applications, numerically) computing the value of $\|m_0\|$ at each step, ridge points can be found. This is important in the estimation of fault lines as ridges, as described in Section 3.2 as well as Chapter 4.

It should be recognised that although the tasks of blurring, which is instrumental in the scale-space representation of an image, and smoothing are similar in many ways, they have different aims and should be distinguished. We also note that the motives for *wavelet decomposition* are related to an appreciable extent to scale-space theory. The use of wavelets in connection with edge estimation will not be addressed in this thesis.

For the purposes of image analysis and estimation of curves, a very important concept is that of the so-called *scale space*. Scale space theory is instrumental in both the Canny and curvature scale space algorithms, which is used in Section 5.4. The rationale for considering an image at various scales simultaneously is, in the words of Lindeberg (1998, p. 122),

> ... the basic fact that image structures, like objects in the world, exist as meaningful entities over certain ranges of scale, and that one, in general, cannot expect to know in advance what scales are appropriate for describing those.

Below we shall give only the basic definitions of scale spaces that are relevant to us; for more details see Lindeberg (1998) and the references cited therein. We resort to the notation $I = I(x)$ for the original image, so to be consistent with the standard use of the letter L to denote the scale-space mapping of I. Formally, this is a map $L : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$, or more specifically $L : \Pi \times \mathbb{R}_+ \to \mathbb{R}$ if Π denotes the observation window, usually a square. The image $I_\sigma \equiv L(\cdot, \sigma)$ is thought of as the representation of I at scale or blurring level defined by the *scale parameter* σ. (Here and below, the parameter σ is not to be confused with a statistical standard deviation, although these notions will be seen to be connected.) It is well known that under fairly general assumptions, including linearity, spatial shift invariance, and the avoidance of creating, in some sense, new detail

when σ is increased, the Gaussian kernel function

$$\varphi(t,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{t^2}{2\sigma^2}}\,, \quad t \in \mathbb{R}\,, \tag{A.2.9}$$

is singled out as a canonical choice for constructing L via convolution. (That φ satisfies the last of the properties listed above has been proved by Silverman (1981); see also the related discussion in Hirst (2002, p. 5)). In the case of two-dimensional imagery, the convolution is usually done in a separable fashion, tolerating a cost in performance along directions off the main axes. Accordingly, at scale σ, the image is defined as $L(\cdot,\sigma) = (I * (\varphi \otimes \varphi))(\cdot,\sigma)$. Partial first and second-order derivatives can be computed using the formula

$$\frac{\partial^{\alpha+\beta}}{\partial^\alpha x^{(1)} \partial^\beta x^{(2)}}\, L(x,\sigma) = \left(I * \frac{\partial^{\alpha+\beta}}{\partial^\alpha x^{(1)} \partial^\beta x^{(2)}}\, (\varphi \times \varphi)(x,\sigma) \right)\,, \quad \alpha,\beta \in \mathbb{N}_0\,. \tag{A.2.10}$$

We always assume that these derivatives exist up to order two in the parts of Π where the regression surface is continuous.

Using a similar idea, Mokhtarian and Mackworth (1992) introduced the concept of the *curvature scale space*. Let $\{(X^{(1)}(s), X^{(2)}(s) : 0 \le s \le T\}$ denote the natural parametrisation of $\mathcal{C}$, and let $\sigma > 0$ denote a scale parameter. An *evolved* version of $\mathcal{C}$, denoted $\mathcal{C}_\sigma$, is defined by $\{(X_\sigma^{(1)}(s), X_\sigma^{(2)}(s)) : 0 \le s \le T\}$, where

$$X_\sigma^{(i)}(s) = X^{(i)}(s) * \varphi(s,\sigma)\,, \qquad i = 1,2\,, \tag{A.2.11}$$

and φ is as defined at (A.2.9). An illustration of the effect of applying the evolution operator to Jordan curves, such as one that approximates the circumference of the African shoreline, is given in Mokhtarian and Mackworth (1992). In Section 5.4 we present an example of the usage of scale-space representations of curves.

We finally note that in Section 4.2 we also use the notation $X_h = \{X_h(s) : 0 \le s \le T\}$, for $h > 0$ a bandwidth parameter. Notwithstanding the similarities of their definitions and properties (the second operator does also have a tendency to smooth out peaks and thereby reduce total arc length), these two curve operators should be distinguished.

A.3 Convergence and Measurability

This section is devoted to topological and measure-theoretic properties of the state spaces on which the summation processes in which we are interested are defined. Our exposition

is mainly based on Van der Vaart and Wellner (1996), especially Chapters 1.5 and 1.6. First, we give a brief explanation how measure-theoretic issues arise with the summation processes considered in this thesis.

It was first noted in the mid-sixties (see e.g. Billingsley (1968) and the references cited therein) that when the space $D[0,1]$ (see the definition on p. 25) is endowed with the supremum metric, then the empirical d.f. of an i.i.d. Uniformly(0,1) distributed sequence fails to be measurable with respect to the Borel σ-field. To remedy this problem, two different main approaches were subsequently developed. The first of these uses the Skorohod metric (or a suitable modification thereof) on $D[0,1]$. This permits the use of the classical theory on weak convergence on metric spaces in the lines of Billingsley (1968). A second corrective, originally proposed by Dudley (1966), was to abandon the Borel σ-field on $D[0,1]$ in favour of the σ-field generated by the open balls with respect to the supremum metric. Due to the non-separability of $D[0,1]$, it may be suspected that the latter σ-algebra is strictly coarser. This indeed turns out to be true, and in particular, the empirical d.f. is rendered by a measurable map. Pollard (1984) gave a fine description of this method, which is easier to work with than the one mentioned earlier. However, empirical and partial-sum processes in more general settings, such as those used in this thesis, cannot be handled by Dudley's approach.

The main invention in Hoffman-Jørgensen's concept (1991) of weak convergence was to drop the requirement of measurability of the sequence, and to maintain it only for the limiting element. This assumption is not as artificial as it might seem at first. Indeed, in the example of the empirical d.f. from the previous paragraph, it is a pivotal result (although not trivial to prove) that the limiting process may be chosen to have sample paths in the space of *continuous* functions on the unit interval. Thus, the limit may be taken to be the standard Brownian motion on $[0,1]$. It is appropriate to remark already here that in a similar manner, Gaussian processes whose covariance functions are 'well behaved' admit continuous modifications. Hoffman-Jørgensen's definition of weak convergence has widely been accepted as the appropriate notion to handle the general empirical process, as well as most aspects of large sample theory in statistics. It will be referred to below as *modern weak convergence* in the sequel whenever explicit mention is made.

Modern weak convergence is framed in terms of so-called *outer integrals* (or *outer expectations*), which refer to a well-established concept in measure theory. If $X : \Omega \to [-\infty, \infty]$ is an arbitrary extended real-valued map defined on a probability space $(\Omega, \mathcal{F}, P)$, then

the outer integral of X with respect to P is defined as

$$E^*X = \inf\{EU : U \geq X,\, U : \Omega \to [-\infty, \infty] \text{ measurable and } EU \text{ exists}\}\,.$$

As usual, existence of EU means that $\min\{E[UI(U > 0)], -E[UI(U < 0)]\} < \infty$. Modern weak convergence of a sequence $\{X_n\}$, which takes its values in a metric space with metric d, to a limiting element X is then defined by requiring that X be measurable, and that

$$E^*f(X_n) \to Ef(X)\,, \qquad \text{for every bounded and continuous } f\,.$$

In this case we use the notation $X_n \Longrightarrow X$, which is consistent with the classical use of the arrow symbol. We shall also need the concept of *outer probability* of an arbitrary subset B of Ω:

$$P^*(B) = \inf\{P(A) : A \supset B\,, A \in \mathcal{F}\}\,,$$

on which the definition of convergence in outer probability builds. Specifically, with the same notation as above, X_n converges in outer probability to X if for every $\epsilon > 0$, it holds that $P^*\big(d(X_n, X) \geq \epsilon\big) \to 0$. The standard apparatus of weak convergence theory (most importantly, Portmanteau's theorem and the characterisation of infinite-dimensional convergence by convergence of marginals plus a tightness/equicontinuity condition, or Prohorov's theorem), largely carries over to the new definitions. In this way, a convergence theory may be established which is sufficiently rich to accommodate most of the present-day statistical applications. Due to the sacrifice of measurability it is necessary to rephrase the definition even of almost sure convergence, but this programme may be carried out as well. In general, we will be slightly sloppy by dropping the asterisk in probabilities and expectations.

The lack of total order in the indexing sets of general empirical processes, and those considered here, make it useful to formulate the theory with the use of nets rather than sequences. Thus, the index sets only need to be directed rather than totally ordered. For our purposes, the only relevant example is a collection of subsets $A \subseteq B$ of a given planar set, which is a $\preceq$-directed class via

$$A \preceq B :\Longleftrightarrow B \subseteq A\,,$$

because every pair (A_1, A_2) with $A_i \subseteq B$ has a $\preceq$-successor:

$$A_3 \equiv A_1 \cap A_2 \succeq A_i\,, \quad i = 1, 2\,.$$

We do not expand further on these topics and instead refer to any standard textbook on topology, for example Munkres (2000).

We now recall the definition of the (vector) space of bounded functions on an arbitrary set T, denoted $\ell^\infty(T)$, which consists of all functions $z : T \to \mathbb{R}$ such that

$$\|z\|_\infty = \|z\|_{\infty,T} = \sup_{t\in T} |z(t)| < \infty\,.$$

As in the classical theory, weak convergence in the space $\ell^\infty(T)$ may be characterised by a Prokhorov-type theorem (Van der Vaart and Wellner, 1996, Theorem 1.5.4, p. 35). It is also possible to give a relatively simple characterisation of tightness of a stochastic process on this space (see Van der Vaart and Wellner, Theorem 1.5.7, p. 37).

The space $\ell^\infty_{\text{loc}}(T) \supseteq \ell^\infty(T)$ of locally bounded functions, which is the one of greatest relevance to us, is defined as follows. Assume that T is σ-compact, that is, there exist $T_1 \subseteq T_2 \subseteq \ldots \subseteq T$ such that each T_i is compact and $T = \cup_{i=1}^\infty T_i$. The space $\ell^\infty_{\text{loc}}(T)$ is defined as the set of all functions $z : T \to \mathbb{R}$ that are uniformly bounded on every T_i (but not necessarily on T). If each T_i is compact then the norm $\|\cdot\|_\infty$, defined by

$$\|z\|_\infty = \sum_{k=1}^\infty 2^{-k} \min(1, \|z\|_{\infty,T_k})\,,$$

makes $\ell^\infty_{\text{loc}}(T)$ a Banach space where the metric is that of compact convergence. Weak convergence in the space $\ell^\infty_{\text{loc}}(T)$ can be characterised by convergence in the spaces $\ell^\infty(T_i)$, for any sequence $\{T_i\}$ as above, provided that the limit in each T_i is a tight r.v. (Van der Vaart and Wellner, 1996, Theorem 1.6.1, p. 43).

We conclude this section with a definition and a theorem which pertain to switching the underlying probability space, while preserving the properties of the sample paths up to sets of arbitrarily small probability. This will serve to simplify the argumentation at the close of the proof of Theorem 4.2 in Subsection 4.3.2. As noted in Van der Vaart and Wellner (1996, p. 17), the separability condition in Theorem A.1 is slightly weaker than tightness, and the latter will suffice for our purposes, where also D is a subset of $\mathbb{R}^2$.

Definition A.2. (Van der Vaart and Wellner, 1996, p. 52) *Let $X_n, X : \Omega \to (D, d)$ be arbitrary maps, taking values in a metric space. The sequence X_n converges* almost uniformly *to X, written $X_n \to_{au} X$, if for every $\epsilon > 0$, there exists a measurable set A with $P(A) \geq 1 - \epsilon$ and $d(X_n, X) \to 0$ uniformly in A.*

Theorem A.1. (Van der Vaart and Wellner, 1996, p. 59) *Let $X_n : \Omega \to (D, d)$ be arbitrary maps into a metric space and X be Borel measurable and separable (i.e., there*

exists a measurable separable set that has probability 1 under the law of X). If $X_n \Longrightarrow X$, then there exists a probability space $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$ and maps $\tilde{X}_n : \tilde{\Omega} \to D$ with (i) $\tilde{X}_n \to_{au} \tilde{X}_\infty$ *and* (ii) $E^* f(\tilde{X}_n) = E^* f(X_n)$*, for every bounded $f : D \to \mathbb{R}$ and every n.*

A.4 Gaussian Processes

In this section we define Gaussian processes, and discuss some of their properties which will be of relevance later on. Gaussian processes typically arise as weak limits of empirical and partial-sum processes, and it is especially the connection to the latter that is of importance to us. As is well-known, empirical processes generally have limits that are 'tied-down' Gaussian processes, the prototype being the Brownian bridge on the unit interval, i.e. when $T = [0, 1]$. We are especially interested, however, in Gaussian processes on general fixed compact subsets of $\mathbb{R}^2$. While there has been an ongoing vivid research interest in the case $T = [0, 1]$ or more generally, $T = [0, 1]^d$ for $d \geq 1$, due to the aforementioned intimate connection with empirical process theory (see e.g. Alexander and Pyke, 1986, Shorack and Wellner, 1986, and Van der Vaart and Wellner, 1996), there is a scarcity of available results for the kind of Gaussian processes that are of interest to us. The development in this subsection mainly follows Adler (1990).

A.4.1 Definition and Fundamental Properties

We start with the basic definition.

Definition A.3. *Let T denote a semimetric space. A stochastic process $\{X_t, t \in T\}$ is called* Gaussian *if for every finite subset $T_0 \subseteq T$ and collection $\{\alpha_t, t \in T_0\}$ of real numbers, the r.v. $Z_{T_0} = \sum_{t \in T_0} \alpha_t X_t$ is Normally distributed. If $E(Z_{T_0}) = 0$ for all choices of T_0, the process $\{X_t\}$ is called* centrally Gaussian.

Unless stated otherwise, centrality will be assumed throughout this subsection. Gaussian processes with (quadratic) drift will be considered in Subsection A.4.2. Notice that the assumption on T is fairly weak, which allows us to consider the following class of processes under the Gaussian framework.

Definition A.4. (Adler, 1990, p. 6) *Let $(E, \mathcal{E}, \nu)$ be a σ-finite measure space, and let $\mathcal{E}_0 \subseteq \mathcal{E}$ denote the subring of those sets with finite ν-measure. A* Gaussian white noise based on ν *is a random set function $W = W_\omega : \Omega \to [-\infty, \infty]^{\mathcal{E}_0}$, where $(\Omega, \mathcal{F}, P)$ is some probability space, such that for all $A, B \in \mathcal{E}_0$,*

1. *$W(A)$ is Normally distributed with mean zero and variance $\nu(A)$.*
2. *If $A \cap B = \emptyset$ then $W(A \cup B) = W(A) + W(B)$ with probability 1.*
3. *If $A \cap B = \emptyset$ then $W(A)$ and $W(B)$ are independent.*

The measure ν is also referred to as the variance measure *of W.*

From the above definition, it is tempting to think of W in terms of the representation

$$W(A) = \int_A dW(t)\,, \quad A \in \mathcal{E}_0\,, \tag{A.4.1}$$

which indicates how one would integrate simple $\mathcal{E}$-measurable functions. However, it is generally impossible to interpret (A.4.1) in a pathwise sense. To appreciate why, it suffices to recall textbook facts on trajectories of standard Brownian motion on $T = \mathbb{R}_+$. The process W is also known under the names *isonormal Gaussian process* (Dudley, 1973, Shorack and Wellner, 1986) or just *Brownian motion* (Ivanoff and Merzbach, 2000, p. 67). The first of these synonyms is explained by extending W through the natural embedding $\mathcal{E}_0 \hookrightarrow L^2(E, \mathcal{E}, \nu)$, $A \mapsto I_A$. By the usual method of extending to linear combinations and taking limits, it is then readily verified that $EW(f)W(g) = \int fg\, d\nu$ for $f, g \in L^2(E, \mathcal{E}, \nu)$. Thus, the process W defines an isometry between the spaces $L^2(\Omega, \mathcal{F}, P)$ and $L^2(E, \mathcal{E}, \nu)$, and a standard construction can be carried out to rigorously define a stochastic integral with respect to W, and so to give meaning to (A.4.1).

The covariance function of the process W may easily be calculated from Definition A.4, which gives

$$\begin{aligned}
\operatorname{cov}\{W(A), W(B)\} &= \operatorname{cov}\{W\left((A \cap B) \cup (A \setminus B)\right), W\left((A \cap B) \cup (B \setminus A)\right)\} \\
&= \operatorname{var}\{W(A \cap B)\} \\
&= \nu(A \cap B)\,.
\end{aligned}$$

The nomenclature of Definition A.4 is justified by the following argument. To construct W, define the function $R = R_\nu : \mathcal{E}_0 \times \mathcal{E}_0 \to \mathbb{R}_+$ via $R(A, B) = E\{W(A)W(B)\} = \nu(A \cap B)$. The function R is easily be shown to be positive semi-definite; i.e., for all $k \geq 1$, all real $\alpha_1, \ldots, \alpha_k$, and all $A_1, \ldots, A_k \in \mathcal{E}_0$, it holds that $\sum_{i,j} \alpha_i \alpha_j R(A_i, A_j) \geq 0$. Hence, by well-known arguments based on Kolmogorov's extension theorem, there exists a probability space $(\Omega, \mathcal{F}, P)$ which is the domain of a Gaussian white noise based on ν. By definition, R is the covariance function of that process. As is true for any second-order stochastic process, pathwise mean-square continuity of a Gaussian process is equivalent

to continuity of its covariance function. The values of W over general rectangles, say $R = [a_1^1, a_2^2] \times \ldots \times [a_1^k, a_2^k] \subseteq \mathbb{R}_+^k$, may be reconstructed by means of iterated application of the difference operator $\Delta_{\cdot,\cdot}$ on each of the k respective coordinates, *viz.*,

$$\begin{aligned} W(R) &= \Delta_{a_1^1,a_1^2} \cdot \ldots \cdot \Delta_{a_k^1,a_k^2} W(t) \\ &= W(a_1^2, \ldots, a_k^2) - W(a_1^2, \ldots, a_k^1) + - \ldots + (-1)^k W(a_1^1, \ldots, a_k^1)\,. \end{aligned}$$

As noted in the paragraph containing (A.4.1), the reconstruction of W on sets other than rectangles requires use of the stochastic integral. For our purposes it suffices to note that such a construction is possible.

The case which has been most extensively studied in the literature (see Shorack and Wellner, 1986; Van der Vaart and Wellner, 1996; and the references cited therein) is where E is a subset of $\mathbb{R}^k$, like the positive orthant $\mathbb{R}_+^k$ or the unit cube $I_k \equiv [0,1]^k$. In either of these cases, one usually considers the stochastic process $\{W(t) \equiv W([0,t))\}$, which defines the *Brownian sheet* on either of the above parameter sets. (Here we use the customary notation, $[0,t) = [0,t_1) \times \ldots \times [0,t_k)$.) Of special importance is the case $T = I_k$, due to the frequent appearance of 'tied-down' Gaussian processes, i.e. processes which, in analogy to the Brownian bridge on $[0,1]$, vanish on one or several boundaries of I_k. Examples of such processes are given in Van der Vaart and Wellner (1996).

In all practically relevant applications the space E from Definition A.4 has some topological structure. By forming equivalence classes, E can then be made a *metric* space. Even if this is not done, the semimetric d which is defined by

$$d(s,t) = \{E[(X_s - X_t)^2]\}^{1/2}\,,$$

is often (as here) called the *canonical metric* for W_T (or T). The canonical metric is easier to work with than the original one (if different), but still insufficient to describe the 'size' of the parameter set T. Entropy conditions are needed to characterise most of the important properties of general Gaussian processes, notably continuity. We shall not further dwell on the topic, on which more information can be found in Adler (1990); see also item 3 in Example 7.1.3 of Ivanoff and Merzbach (2000, pp. 135–136). Lemmas A.1 and A.2 below, of which especially the second will be of importance later on, give a flavour of the nature of conditions required, and the subtleties that arise even in the case where the measure ν equals Lebesgue measure.

For the Gaussian white noise W based on ν, the canonical semimetric is given by

$$\begin{aligned} d^2(A,B) &= E\{W(A) - W(B)\}^2 \\ &= \nu(A) + \nu(B) - 2\nu(A \cap B) \\ &= \nu(A \cup B) - \nu(A \cap B) \\ &= \nu(A \triangle B)\,. \end{aligned}$$

In Section A.8 we introduce the semimetric $d_1 = d^2$, which arises in the important case where ν equals Lebesgue measure.

The last two lemmas that we reproduce here relate to continuity properties of the Gaussian process, with the second lemma showing that the restriction to polygonal sets in Lemma A.1 cannot be dispensed with. Call the set $A \subseteq \mathbb{R}^d$ a *lower set* if ($s \leq t$ and $t \in A \Rightarrow s \in A$).

Lemma A.1. (Dudley, 1973, Theorem 4.1) *Let W denote a Gaussian white noise based on d-dimensional Lebesgue measure on a set $E \subset \mathbb{R}^d$. For $m > d$ and $x_1, \ldots, x_m \in E$, let $g(x_1, \ldots, x_m; \omega) = W_\omega(\text{conv}\{x_1, \ldots, x_m\})$ where* conv *denotes the convex hull. Then the following assertions hold:*
(a) *The sample functions $g(\cdot; \omega)$ are continuous on E^m.*
(b) *If $x_1, \ldots, x_m \in K \subseteq E$, where K is compact, then g has the sample modulus of continuity equal to $h(u) = (u|\log u|)^{1/2}$.*

Lemma A.2. (Adler, 1990) (a) *The Brownian sheet on lower sets in $[0,1]^2$ is a.s. discontinuous and unbounded.* (b) *If considered on the convex subsets of $[0,1]^2$, however, it* is *a.s. continuous.*

It will prove important, especially for numerical calculations, that Gaussian processes admit an alternative construction as a *random biorthogonal series*. This expansion, which is well known from the area of signal processing, also applies to general L^2 processes, and hence will be utilised in Chapter 6. The theoretical background of this expansion is the topic of Section A.5. First we collect some results about the location of the maxima of Gaussian processes in the next subsection.

A.4.2 Maximisers of Gaussian Processes

In the sequel, let $V \in \mathbb{R}^{d \times d}$ ($d \in \mathbb{N}$) be a fixed positive-definite matrix and $\{W(t),\, t \in T \equiv \mathbb{R}^d\}$ be a centred Gaussian process whose covariances have the following rescaling

property:

$$\operatorname{cov}\{W(\lambda t_1), W(\lambda t_2)\} = \lambda \operatorname{cov}\{W(t_1), W(t_2)\}\,, \tag{A.4.2}$$

for all $\lambda > 0$ and $t_1, t_2 \in T$. The following result, due to Kim and Pollard (1990), establishes that the maximiser of a Gaussian process minus a drift parameter is a Borel measurable map. This fact will be essential in the considerations of weak convergence, in the sense of Section A.3, of a sequence of M-estimators derived from a partial-sum process; see Section A.7. As noted by Ferger (1999), unimodality may also be deduced from an older result (Tsirel'son, 1975, cited by Lifshits, 1982). Using Tsirel'son's result, it is possible to replace the non-degeneracy condition of the subsequent theorem by the weaker condition

$$E\{[Z(s) - Z(t)]^2\} \neq 0 \quad \text{for all pairs } s \neq t \in T\,. \tag{A.4.3}$$

Theorem A.2. (Kim and Pollard, 1990) *Assume the previously stated conditions with the exception of* (A.4.3)*, and define the Gaussian process Z by $Z(t) = W(t) - \frac{1}{2}t'Vt$. If Z has non-degenerate increments, i.e.,*

$$\operatorname{var}\{Z(s) - Z(t)\} \neq 0 \quad \textit{for all pairs } s \neq t \in T\,,$$

then $Z \in \ell^\infty(T)$ a.s. Moreover, Z is a.s. unimodal, i.e. no sample path of Z can achieve its supremum at two distinct points of T.

Section A.7 gives a brief account on the approach that is generally taken in deriving the limit distribution of modes as appearing in Theorem A.2. That limit distribution is very often of the quadratic-drift type on which our previous discussion focused (see Kim and Pollard, 1990, and Van der Vaart and Wellner, 1996, Theorem 3.2.10, p. 293). The arguments used to prove Theorem 4.2 of this thesis, however, will establish the limit form by direct Taylor-series type expansion with the aid of Markov-inequality bounds.

The most commonly encountered case is when in the setup of Theorem A.2, one has $d = 1$ (i.e., the scalar case) and $V = 2$. This distribution has made a frequent appearance since the 1960s, commencing with the work on the mode of a probability density by Chernoff (1964), Venter (1967), Prakasa Rao (1969), and Wegman (1971). As a result of a deep analysis, Groeneboom (1985, 1989) and independently Daniels and Skyrme (1985) obtained the Fourier transform of the distribution of this r.v. The latter authors were the first to give tabulated values. Groeneboom and Wellner (2001), who named the distribution in honour of Chernoff, presented an effective method for its simulation, after previous attempts by other authors cited in that paper.

Below we state Groeneboom's main result, which determines more generally the distribution of the r.v. $T_c^{*,0}$ (say) which is obtained for $V = 2c$. The term "Chernoff's distribution" will be reserved for the case $c = 1$.

Theorem A.3. (a) (Groeneboom (1989), Corollary 3.3) *The density of* $T_c^{*,0}$ *is given by*

$$f_{T_c^{*,0}}(u) = \tfrac{1}{2}\,\varphi_c(u)\varphi_c(-u)\,,$$

where the function $\varphi_c(u)$ *has the Fourier transform*

$$\bar{\varphi}_c(\lambda) = \int_{-\infty}^{\infty} e^{i\lambda s}\varphi_c(s)\,ds = \frac{2^{1/3}c^{-1/3}}{\mathrm{Ai}(i(2c^2)^{-1/3}\lambda)}\,,$$

and Ai *denotes the Airy function as defined, for example, in Abramowitz and Stegun* (1970).

(b) (Groeneboom (1989), Corollary 3.4 (iii)) *The density* $f_{T_c^{*,0}}(u)$ *has the following asymptotic behaviour:*

$$f_{T_c^{*,0}}(u) \sim \tfrac{1}{2}(4c)^{4/3}|u| \exp\left\{-\tfrac{2}{3}c^2|u|^3 + (2c^2)^{1/3}a_1|u|\right\} \mathrm{Ai}'(a_1)^{-1} \qquad \text{as } |u| \to \infty\,, \quad \text{(A.4.4)}$$

where $a_1 \approx -2.3381$ *is the largest zero of the Airy function* Ai *and where* $\mathrm{Ai}'(a_1) \approx 0.7022$ (*see Abramowitz and Stegun,* 1970).

As Groeneboom notes, it follows from part (b) that the distribution of $T_c^{*,0}$ has very thin tails. In the present context of edge estimation, it will become evident later on that this means that a quite narrow confidence region could be constructed. We shall discuss the consequences of this fact simultaneously with our main result (Theorem 4.2), which involves a close relative of $T_c^{*,0}$ with $c = 1/2$.

A.5 The Karhunen-Loève Expansion

In Subsection A.4.1 (p. 148) we have noted that a Gaussian process $\{W(t),\, t \in T\}$ with a covariance function that is continuous on $T \times T$ may be treated as a random element in the space $L^2(T)$. In order to compute a functional of W, as shall be the case in this thesis (Theorem 4.2 in Section 4.2), one needs to find an approximation which is capable of reproducing the behaviour of the process W. The L^2 condition makes it possible to use tools from functional analysis. A central role in this approach is played by Mercer's theorem (Theorem A.4). From a computational viewpoint, this approach has often been

argued to be more appealing than direct random sampling in some underlying probability space (see e.g. Ghanem and Spanos, 1991).

Importantly, Mercer's theorem and its corollaries apply to general L^2 processes. In the modelling of physical phenomena such as radar image acquisition, the 'continuous' viewpoint is usually the most natural to employ, especially if we do not want to be relatively specific about the model that generates the data, such as is implicit in the Fourier and (albeit to a lesser extent extent) wavelet transforms.

A.5.1 Generalities

Throughout what follows, T denotes a compact subset of the Euclidean space $\mathbb{R}^k$ for some $k \in \mathbb{N}$, and $\{X(t),\, t \in T\}$ is an L^2-integrable stochastic process with

$$E\{X(t)\} = 0\,, \quad t \in T\,, \tag{A.5.1}$$

and covariance function $\Gamma_X(t,t') = E\big[X(t)X(t')\big]$ for $t, t' \in T$. Heuristically, we are looking for an expansion

$$X(t) = \sum_{r=1}^{\infty} \xi_r\, \psi_r(t) \tag{A.5.2}$$

with an orthonormal system of (deterministic) functions $\{\psi_r\}$, and uncorrelated r.v.s $\{\xi_r\}$, which makes the right-hand side of (A.5.2) a *random biorthogonal series*. For now, the convergence is assumed only to be in L^2. Under the foregoing assumptions, we have the representation

$$\xi_r = \int X(s)\psi_r(s)\, ds \tag{A.5.3}$$

(which exists as a mean-square integral), and hence by uncorrelatedness,

$$\iint \psi_r(t)\psi_u(s)\, \Gamma_X(t,s)\, dt\, ds = E\big(\xi_r^2\big)\, \delta_{ru}\,,$$

δ_{ru} denoting the Kronecker delta. Conversely, for a given orthonormal system $\{\psi_r\} \subset L^2$, one can project the data (*viz.*, the realisations of X as at (A.5.3)) such that (A.5.2) holds. Among the infinity of possible choices for the sequence $\{\psi_r\}$ in (A.5.2), there is one which is distinguished by the requirement that for each $K \in \mathbb{N}$, the Kth *residual*, obtained by summing in (A.5.2) over indices $r > K$ only, has minimal L^2 norm. This corresponds to the special case of the *Karhunen-Loève expansion* of X. (By what we have stipulated, the enumeration of the terms in the sum is completely specified, which is of relevance later on; see the remarks at the end of Subsection A.5.3.) In this special case, the previous

discussion is made valid through the following central result, which is of purely functional-analytic nature and has been known for quite a while; for a proof, see e.g. Zaanen (1956). We formulate it for Gaussian processes on unit cubes. By means of a rescaling, this does not imply a serious restriction. We also simplify notation as $K \equiv \Gamma_X$.

Theorem A.4. (Mercer's Theorem). (Adler, 1990, p. 75). *Let X be a centred Gaussian process on $I_k = [0,1]^k$ with continuous covariance function $K(s,t)$, and let $\lambda_0 \geq \lambda_1 \geq \ldots$, and $\psi_0, \psi_1, \ldots$, be, respectively, the eigenvalues and normalised eigenfunctions of the integral operator $\mathcal{K} : L^2(I_k) \to L^2(I_k)$ defined by*

$$\mathcal{K}\psi(t) \equiv \int K(s,t)\psi(s)\, ds\,. \tag{A.5.4}$$

Then it holds that

$$K(s,t) = \sum_{n=0}^{\infty} \lambda_n\, \psi_n(s)\psi_n(t)\,,$$

where the sequence converges absolutely and uniformly on $I_k \times I_k$.

In the situation of Theorem A.4, and with the notation used earlier, the random coefficients $\{\xi_r\}$ in (A.5.2) may be normalised as

$$\xi_r = \lambda_r^{1/2}\, \eta_r\,, \quad \|\eta_r\| = 1\,, \tag{A.5.5}$$

and the quantity

$$\zeta = \zeta_m = \left(\sum_{i=1}^{m} \lambda_r\right) \Big/ \left(\sum_{i=1}^{\infty} \lambda_r\right) \tag{A.5.6}$$

may be interpreted as the proportion of the variance explained by the first m terms in the expansion of X. A plot of ζ_m (or $1 - \zeta_m$) as a function of m is often, especially in applied contexts, referred to as a 'scree plot' (Sharma, 1996).

By Theorem 3.8 in Adler (1990, p. 71), the convergence in (A.5.2) is automatically even uniform a.s., provided that the sample paths of X are a.s. continuous.

As evinced by (A.5.4), the eigenfunctions are the solutions of a Fredholm integral equation of the second kind. The number of known examples where the solutions are obtainable analytically is rather small. Nevertheless, a discretised version of the expansion may be expected to perform well, and to yield stable behaviour. This is important in the usual situation where the process X is observed under the presence of additive random noise, as we assume in Chapter 6. The underlying model that will be of relevance to us is where X is a *finite* mixture of $L \in \mathbb{N}$ independent processes with mutually different probability

laws, defined on a space of square-integrable functions over a common probability space:

$$P_X = \sum_{l=1}^{L} \omega_l P_{\widetilde{X}_l}\,, \tag{A.5.7}$$

where P_Y denotes the probability law induced by the stochastic process $Y = \big\{Y(t),\, t \in T\big\}$, $\omega_l > 0$ for $l = 1, \ldots, L$ with $\sum_{l=1}^{L} \omega_l = 1$ are the mixing weights, and for each l, in the customary differential notation and using the notions from Section A.4,

$$d\widetilde{X}_l(t) = X_l(t)\, dt + \sigma_l(t)\, dW(t)\,, \quad t \in T\,, \tag{A.5.8}$$

where the $\sigma_l(t)$ are deterministic variance processes. The process W may be assumed to be white noise, but this is taken rather as a calibration model for a wider class of random error processes later on. In the situation that will be of interest, a sample of size $n_l \geq 1$ is available for each of the l 'types' for $l = 1, \ldots, L$. A datum from the sample consists of a realisation of the relevant process X_l on a discrete set of points, usually a regular grid.

For a given set of observations, dimension reduction is achieved by projecting the observed data onto a basis $\{\psi_r\}$ as in (A.5.2). Realisations of the sequence $\{\lambda_r\}$ from (A.5.2) are then extracted from realisations of the sequence of r.v.s, *viz.*, $\lambda_r = \mathrm{var}(\xi_r)$ for $r \geq 1$. In this context, interest is in the sequence $\{\lambda_r\}$ from (A.5.2), and especially the sequence of associated *scores*, that is, the realisations $\{\xi_r = \xi_r(\widetilde{\omega})\}$ from (A.5.2) and (A.5.5). The latter are taken as surrogates in estimation problems pertaining to the population process.

The question of how many basis functions are to be retained in the expansion at (A.5.2) is both important and usually, such as in the case study of this thesis, highly non-trivial. If the goal is classification, and even if differences between various methods for this are deemed of no importance, the sequence $\{\lambda_r\}$ or the proportion of variance calculated therefrom (see the paragraph containing (A.5.6)), is not always reliable, as even basis functions of small index may contain a large amount of noise.

There may also be a need to compute the actual approximant to the stochastic process X, that is the right-hand side of (A.5.2) minus the residual. This case will be considered in the next subsection.

A.5.2 Karhunen-Loève Expansion for Gaussian Processes

There is a discrete form of the K-L expansion which has already been mentioned in the introduction. This (linear) operator is called the *K-L Transform*, and will be considered in

more detail in Subsection A.5.3. Unlike the situation for the Fourier transform, there is no fast algorithm such as the Fast Fourier Transform (FFT) available for computation of the K-L Transform (Ahmed and Rao, 1975, p. 191). This seems to have had a disencouraging effect on its use. More recently however, Courmontagne (1999) has described a method which utilises the previously described 'continuous' origins of the K-L Transform in order to devise an algorithm that speeds up the computations significantly, with additional benefits in terms of numerical precision.

We limit ourselves to the exposition of the main points in Courmontagne (1999), referring to the original paper for a complete account. In addition to the general assumptions of Section A.5, we specialise to the case $T = [-R, R]$ for finite R. Courmontagne (1999) also assumes stationarity of the process X, so that Γ_X becomes a circulant matrix. The key idea is to symmetrise the underlying stochastic process X, now assumed to be defined on the cube $T = [-R, R]$ and hence its kernel:

$$\begin{aligned} X_{\text{symm}}(t) &= X(R - |t|)\,, \\ \Gamma_{X_{\text{symm}}}(t, t') &= \Gamma_X(R - |t|, R - |t|)\,. \end{aligned}$$

By inspection, if the eigenvalues are ordered according to magnitude and signs are disregarded, the solutions of the following *modified* Fredholm integral equation:

$$\frac{1}{2} \int_{-2R}^{2R} \Gamma_{X_{\text{symm}}}(t, s)\, \psi_{\text{symm}}(s)\, ds = \lambda\, \psi_{\text{symm}}(t)\,,$$

are connected to the solutions of (A.5.4) via the relation

$$\psi_{\text{symm}}(t) = \psi(T - |t|)\,, \quad |t| \leq T\,. \tag{A.5.9}$$

The Fourier expansions of $\Gamma_{X_{\text{symm}}}$ and each element in the solution set $\left\{\psi^{(n)}_{\text{symm}},\, n \in \mathbb{N}\right\}$ retain only the even (i.e., cosine) terms, whence the Gibbs phenomenon is not present, and the convergence

$$\begin{aligned} \Gamma_{X_{\text{symm}}}(t, t') &= \frac{K_0(t)}{2} + \sum_{l=1}^{\infty} K_l(t) \cos\left(\frac{\pi m t'}{2R}\right)\,, \quad \text{where} \\ K_l(t) &= \frac{1}{2R} \int_{-2R}^{2R} \Gamma_{X_{\text{symm}}}(t, t') \cos\left(\frac{\pi l s}{2R}\right) dt', \end{aligned}$$

is even uniform. Thus, for any truncation index $N \in \mathbb{N}$, and defining the matrix $\Omega =$

$\big(\Omega_{l,m}, 0 \le l, m \le N\big)$ with entries

$$\Omega_{0,m} = \tfrac{1}{2}\,\Omega'_{0,m}\,, \qquad \Omega_{l,m} = \Omega'_{l,m} \quad \text{for } l \ge 1\,, \tag{A.5.10}$$
$$\Omega'_{l,m} = \frac{1}{2R}\int_{-2R}^{2R}\int_{-2R}^{2R} \Gamma_{X_{\mathrm{symm}}}(t,s)\,\cos\left(\frac{\pi m t}{2R}\right)\cos\left(\frac{\pi l s}{2R}\right)\,dt\,ds\,,$$

the eigenvectors $\big\{\alpha_l = \alpha_l^{(n)},\, l \in \mathbb{N}\big\}$ are such that

$$\sum_{l=0}^{N} \alpha_l^{(n)}\, K_l(t) \propto \psi_{\mathrm{symm}}^{(n)}\,, \quad n \in \mathbb{N}\,. \tag{A.5.11}$$

The ordering in n in (A.5.11) is again according to the sequence of associated eigenvalues, ordered with respect to their moduli, which coincide for the two cases; our interest, however, is limited to the sequence $\big\{\alpha_l\big\}$. From (A.5.9) it follows that the approximations of the solutions of (A.5.4) evaluate to

$$\psi_{\mathrm{approx}}^{(n)}(t) = \sum_{l=0}^{N} \alpha_l^n\,\cos\left(\frac{\pi l(t-T)}{2R}\right)\,, \quad |t| \le R\,. \tag{A.5.12}$$

With alterations of a mainly notational kind, the previous discussion also applies to the case of random fields. Similarly as before, it may be assumed that $T = [-R, R]^2 \subset \mathbb{R}^2$ is a rectangle. The added complexity shows up mainly in the analogue of (A.5.10), which now involves a four-dimensional Fourier or cosine transformation. In order to keep the computational expense feasible, we use a *sparse grid* approach that was described by Hallatschek (1992). The four-dimensional Fourier Transform of the covariance kernel $\Gamma_{X_{\mathrm{symm}}}$ is then replaced by two consecutive two-dimensional Fourier Transforms operating on a different pair of dimensions. Thus the saving in terms of sparsity, and hence of order of complexity, is considerably less than for a four-dimensional FFT.

A.5.3 More General L^2 Signals

For reasons mentioned earlier, in the area of signal processing it is particularly appealing to treat observations as discrete samples from a stochastic process that is intrinsically continuous in time or, as in the case study of Chapter 6 of this thesis, in a two-dimensional space of images. Estimators of discretised versions of the series of functions $\{\psi_r(t)\}$ and of the non-negative sequence $\{\lambda_r\}$, defined in Subsection A.5.1, are then obtainable through the time-honoured approach of singular value decomposition (SVD). We confine ourselves to a rather sketchy review of the theory, because our main interest will be

in those aspects which distinguish the case of functional data from classical (discrete) multivariate analysis.

We envisage the situation described in the paragraphs containing (A.5.7) and (A.5.8), and hence let the data consist of a sample from a population with L distinct sub-populations or *types* (determined by laws of stochastic processes), such that a sample of size n_l is available on type l, for $l = 1, \ldots, L$. In each instance the discretisation is assumed to be on the same finite grid, $\mathcal{T} = \{t_1, \ldots, t_p\}$ with $p < \infty$. For the purposes of the present section only, and in order to avoid ambiguities in notation, we will use underlined letters to denote matrices. The SVD is computed for the matrix $\underline{X} \in \mathbb{R}^{n \times p}$ which contains, in each row, an observation of the process $\{X(t)\}$. In the case where X is $\mathbb{R}^2$-valued, as is the case in the analysis of Chapter 6, the data will need to be vectorised as described in Section 1.2 (see p. 9). The total number of observations available is $n = \sum_{l=1}^{L} n_l$. As is standard practice in the literature (e.g. Ramsay and Silverman, 1997), and in parallel to the analogous stipulation at (A.5.1), a centring at the overall mean is assumed to have been performed so that the column sums of $\underline{X}$ are all equal to zero. In customary notation, write the SVD of $\underline{X}$ as follows:

$$\underline{X} = \underline{U}\,\underline{L}\,\underline{V}',$$

with row-orthonormal matrices $\underline{U} \in \mathbb{R}^{n \times q}, \underline{V} \in \mathbb{R}^{p \times q}$ and the diagonal matrix $\underline{L} = \mathrm{diag}\{d_1, d_2, \ldots, d_q\} \in \mathbb{R}^{q \times q}, d_1 \geq d_2 \ldots \geq d_q$, where $q = \mathrm{rank}(\underline{X})$. The sample covariance matrix equals $\underline{R} = \underline{V}\,\underline{L}^2\underline{V}'/n$, and 'natural' estimators of the quantities in (A.5.5) are given by

$$\{\widehat{\eta}_r(t),\, t \in \mathcal{T}\} = e_r\,\underline{U}\,, \quad \widehat{\lambda}_r = \underline{L}^2_{rr}/n\,, \qquad 1 \leq r \leq q\,, \tag{A.5.13}$$

where e_r denotes the rth unit (row) vector of the appropriate dimension. Reasons why these estimators can be indeed considered 'natural' are given in Hall, Poskitt and Presnell (2001), where a more comprehensive exposition of the theory, alongside with references, may be found. Additionally, we point to the observation of Gerbrands (1981, p. 379) that it is because of the choice of the covariance estimator $\underline{R}$ that the notions of the singular value decomposition on the one hand, and the (random) techniques of Karhunen-Loève expansion and principal components analysis on the other hand, essentially coincide.

In a statistical context with the presence of noise, the quantity ζ defined at (A.5.6) is suited to choosing the most appropriate number q of basis functions: at the point where it levels, it is likely that the variability is contributed mostly by unwanted noise (see also the related discussion in Li and Shedden, 2002). Including such basis functions could

actually increase classification error, due to overfitting. In our case study, we resort to a cross-validatory approach (see the paragraph containing (6.3.2)) to select the basis functions.

A.6 Central Limit Theorem for Set-Indexed Martingales

In this section we state a central limit theorem for set-indexed processes, such as the one that we are concerned with in Chapter 4. The key feature to exploit is the martingale property. Subsection A.6.1 gives a brief exposition of the theory of set-indexed martingales. The martingale of specific interest in Chapter 4 is introduced in Subsection A.6.2, and the central limit theorem itself is stated in Subsection A.6.3.

A.6.1 Framework and Notation

The exposition in this subsection gives some background material on the central limit theorem (CLT) which will be instrumental in the proof of Theorem 4.2 in Section 4.3. The CLT is given in Ivanoff and Merzbach (2000), referred to below as IM, and we refer to their monograph for a detailed account.

First we recall some definitions pertaining to the notion of a set-indexed martingale. Using this framework in the present context is particularly appealing because both cases of gridded or Poisson-distributed design turn out to be tractable by this approach.

As we have seen in Subsection A.4.1 (see Lemma A.2), care must be taken with regard to the size of the class of sets over which a Brownian motion is considered. The notion of an *indexing set*, defined in (IM, p. 10), is tailored to this need. We reproduce it below for later reference, suppressing the definition of the somewhat lengthy and technical condition which these authors call "separability from above." As is usually the case in practically relevant examples, this condition is later shown in our application to be satisfied.

Definition A.5. (Ivanoff and Merzbach, 2000, p. 10) *Let $T \subseteq \mathbb{R}^2$ denote a closed (but not necessarily bounded) connected set. A nonempty class $\mathcal{A}$ of subsets of T is called an* indexing collection *if it satisfies the following, where $\mathcal{A}(u)$ denotes the class of finite unions of sets in $\mathcal{A}$:*

1. *$\emptyset \in \mathcal{A}$, and $A^\circ \neq \emptyset$ if $A \neq \emptyset$ or T. In addition, there is an increasing sequence (B_n) of sets in $\mathcal{A}(u)$ such that $T = \bigcup_{n=1}^{\infty} B_n$.*

2. *$\mathcal{A}$ is closed under arbitrary intersections and if $A, B \in \mathcal{A}$ are nonempty, then $A \cap B$ is nonempty. If (A_i) is an increasing sequence in $\mathcal{A}$ and there exists n such that $A_i \subseteq B_n$ for every i, with (B_k) the same as mentioned in point* 1, then $\overline{\bigcup_i A_i} \in \mathcal{A}$.

3. The σ-algebra generated by $\mathcal{A}$ coincides with the class of Borel sets of T.

4. $\mathcal{A}$ is separable from above.

The classical example, for which all the assumptions of Definition A.5 may be readily verified, is $T = \mathbb{R}^d$ and $\mathcal{A}$ equal to the class of lower sets in $\mathbb{R}_+$. In the next subsection we introduce a somewhat more complex example which will play a major role later on (see Section 4.2).

Among the various alternative definitions of a martingale indexed by a subset of $\mathbb{R}^d$ with $d > 1$, we shall need the one given presently. Following IM (p. 25), an $\mathcal{A}$-indexed process X is said to be *additive* if $X_\emptyset = 0$ and if $X(C_1) + X(C_2) = X(C)$ a.s. for any $C, C_1, C_2 \in \mathcal{V}$ with $C = C_1 \cup C_2$ and $C_1 \cap C_2 = \emptyset$. Here the class $\mathcal{V}$ is defined as $\mathcal{V} \equiv \{A \backslash B,\, A \in \mathcal{A},\, B \in \mathcal{A}(u)\}$. As usual, we write $\bigvee_i \mathcal{E}_i$ for the smallest σ-algebra which contains $\cup_i \mathcal{E}_i$. In the sequel we occasionally use the alternative notation $X_A = X(A)$.

Definition A.6. (Ivanoff and Merzbach, 2000, p. 54) *Let $(\Omega, \mathcal{F}, P)$ be a complete probability space with a filtration $\{\mathcal{F}_A,\, A \in \mathcal{A}\}$ which satisfies $\mathcal{F}_T = \mathcal{F}$ and the conditions:*

- *For all $A \in \mathcal{A}$, we have $\mathcal{F}_A \subseteq \mathcal{F}$ and $\mathcal{F}_A$ contains the P-null sets.*
- *If $A, B \in \mathcal{A}$ satisfy $A \subseteq B$, then $\mathcal{F}_A \subseteq \mathcal{F}_B$.*
- *Monotone outer-continuity: $\mathcal{F}_{\cap A_i} = \bigcap \mathcal{F}_{A_i}$ for any decreasing sequence (A_i) in $\mathcal{A}$.*

Let $X = \{X_A,\, A \in \mathcal{A}\}$ denote an additive integrable process such that X_A is $\mathcal{F}_A$-measurable, for every $A \in \mathcal{A}$. Let further $\mathcal{G}^*_C = \bigvee_{B \in \mathcal{A}(u), B \cap C = \emptyset} \mathcal{F}_B$. The process X is called a strong martingale *if $E(X_C \mid \mathcal{G}^*_C) = 0$ for all $C \in \mathcal{C}$.*

An important example of a strong martingale is provided by a mean-zero process with independent increments, where the filtration is taken to be the minimal one (see IM, p. 66). Important examples are provided by the Poisson process (see the definition below Theorem 1.1) and, especially for our purposes, Gaussian white noise based on some variance measure ν (see Definition A.4 in Subsection A.4.1). For strong martingales in L^2, it can be shown that there exists a unique '*-predictable' quadratic variation process (IM, p. 74). We mention this fact merely in reference to the statement of Theorem A.5 in Subsection A.6.3. Other properties well-known from martingale theory on the real line, derived in IM and the references cited therein, shall not be of concern here.

A.6.2 A Polygonal Gaussian Process

In this subsection we construct a Gaussian process defined on $\mathbb{R}^2$ that plays a key role in Section 4.2. As noted in the paragraph following Definition A.6, this process is a martingale, and hence we also specify a suitable indexing collection of this martingale in the sense of Definition A.5.

Let $\mathcal{S} = \{x : |x^{(1)}| \leq M\} \subset \mathbb{R}^2$ denote the semi-infinite strip bordered by the straight lines $x^{(1)} = \pm M$, for $M > 0$. Let

$$\begin{aligned}
\mathcal{A} &= \{A = A_1 \cup A_2 : A_1, A_2 \in \{\emptyset\} \cup \mathcal{F}_0\,, A_1 \cap A_2 \text{ is empty or a singleton}\}\,, \\
\mathcal{F}_0 &= \{F_0 \subset \mathcal{S} : F_0 \subset \mathbb{R} \times \mathbb{R}_+ \text{ or } (-F_0) \subset \mathbb{R} \times \mathbb{R}_+\,, \\
&\qquad F_0 \text{ is compact, convex and satisfies } \big((x^{(1)}, x^{(2)}) \in F_0 \Rightarrow (x^{(1)}, 0) \in F_0\big)\}.
\end{aligned} \tag{A.6.1}$$

Lemma A.3. *The system $\mathcal{A}$ is an indexing collection.*

Proof. With each set $A \in \mathcal{F}_0$ we associate its boundary function h_A on the interval $[l_A, r_A]$, where $l_A = \min\{x^{(1)} : (x^{(1)}, 0) \in A\}$, $r_A = \max\{x^{(1)} : (x^{(1)}, 0) \in A\}$ and

$$h_A(x^{(1)}) = \begin{cases} \sup\{x^{(2)} : (x^{(1)}, x^{(2)}) \in A\}\,, & A \cap \big(\{x^{(1)}\} \times \mathbb{R}_+^*\big) \neq \emptyset\,, \\ \inf\{x^{(2)} : (x^{(1)}, x^{(2)}) \in A\}\,, & (-A) \cap \big(\{x^{(1)}\} \times \mathbb{R}_+^*\big) \neq \emptyset\,, \\ 0\,, & \text{otherwise}\,, \end{cases}$$

this function having the property of being either concave and non-negative, or convex and non-positive. Hence the assertion may be proved by employing arguments that are very similar to those in Example 1.2.6 in IM (pp. 17–18). To verify condition 2 in Definition A.5, note that the intersection of an arbitrary number of convex sets is convex. Finally, let

$$\begin{aligned}
\mathcal{A}_n &= \{A = A_1 \cup A_2 : A_1, A_2 \in \{\emptyset\} \cup \mathcal{F}_0^n\}\,, \\
\mathcal{F}_0^n &= \{F_0 \in \mathcal{F}_0 : h_{F_0} \text{ is either non-positive and convex, or non-negative and concave} \\
&\qquad \text{on } [l_{F_0}, r_{F_0}], \text{ and linear on } [k\,2^{-n}\,M, (k+1)\,2^{-n}\,M], \text{ for } k \in \{-2^n, \ldots, 2^n - 1\}\}\ .
\end{aligned} \tag{A.6.2}$$

The sequence of sets $\{\mathcal{A}_n,\, n \in \mathbb{N}\}$ defined at (A.6.2) may be shown to meet the requirements of the 'separability from above' condition, stated as item 4 in (IM, p. 10). The required corresponding functions $\{g_n, n \in \mathbb{N}\}$ may be chosen as the dyadic-valued majorants of the function h appearing at (A.6.2), similar to (IM, p. 19) but with using a piecewise-linear instead of a piecewise-constant approximation. It follows that, for

example, $A = \cap_n g_n(A)$ because of the fact that any convex set is the intersection of enveloping tangents placed at a countable dense subset of boundary points. □

Consider now the Gaussian process W which is defined as follows. For $s, t \in \mathbb{R}$, let

$$\mathcal{A}(s,t) = \left\{x = (x^{(1)}, x^{(2)}) \in \mathbb{R}^2 : |x^{(1)}| \leq 1\,, x^{(2)} \leq t + sx^{(1)}\right\}, \tag{A.6.3}$$

and put $W(s,t) = W(\mathcal{A}(s,t) \triangle \mathcal{A}(0,0))$.

Lemma A.4. *The process W is continuous, with modulus of continuity equal to $h(u) = \big(u \log|u|\big)^{1/2}$ on bounded subsets of $\mathbb{R}^2$.*

Proof. The assertions follow directly from parts (a) and (b), respectively, of Lemma A.1. □

We conclude this subsection by giving the explicit form of the covariance function

$$\varphi = \varphi(s_1, t_1, s_2, t_2) = \mathrm{cov}\big\{W(s_1,t_1), W(s_2,t_2)\big\}\,. \tag{A.6.4}$$

The intimate connection to the indexing collection defined earlier, which is apparent from the definition of φ, motivates us to refer to the process W with this covariance function as the *polygonal Gaussian process*. In calculating φ, and without loss of generality, it may be assumed that $t_1 - s_1 \geq 0$ (otherwise take the negative of all arguments at (A.6.4)) and that if $t_1 - s_1 = t_2 - s_2$, then $t_1 + s_1 \geq t_2 + s_2$. Note too that by well-known properties of the Boolean ring of sets,

$$\begin{aligned}\varphi(s_1,t_1,s_2,t_2) = \nu\big(\{\mathcal{A}(s_1,t_1) \cap \mathcal{A}(0,0)\} \triangle \{\mathcal{A}(s_1,t_1) \cap \mathcal{A}(s_2,t_2)\} \\ \triangle \{\mathcal{A}(s_2,t_2) \cap \mathcal{A}(0,0)\} \triangle \mathcal{A}(0,0)\big)\end{aligned}$$

where ν in this case is Lebesgue measure. The exact description of φ requires the consideration of a number of nested cases, and it will be most convenient to display them in the form of computer pseudo-code. Let

$$\rho = \frac{t_1 - t_2}{s_2 - s_1}\,, \qquad s_1 \neq s_2\,,$$

denote the first coordinate of the intersection point in $\mathbb{R}$ of the two lines which define the (finite) boundaries of $\mathcal{A}(s_1,t_1)$ and $\mathcal{A}(s_2,t_2)$ in the $x^{(2)}$ direction, either upwards or downwards.

if $t_1 + s_1 > 0$ **then**
{Boundary completely above the axis $x^{(2)} = 0$}
if $(s_1 = s_2)$ or $(|\rho| > 1)$ **then**
if $t_2 + |s_2| > 0$ **then**
$\varphi = 0$
else
if $t_2 - s_2 > 0$ **then**
if $t_2 + s_2 > 0$ **then**
$\varphi = 2t_2$
else
$\varphi = \frac{1}{2s_2}(t_2 - s_2)^2$
end if
else
$\varphi = \frac{1}{2s_2}(t_2 + s_2)^2$
end if
end if
else
{Boundaries cross}
if $s_2 \neq 0$ **then**
if $t_2 - s_2 < 0$ **then**
$\varphi = t_2\Big(\rho + \frac{t_2}{s_2}\Big) + \frac{s_2}{2}\Big(\rho^2 - \frac{t_2^2}{s_2^2}\Big) + t_1(1-\rho) + \frac{s_1}{2}(1-\rho^2)$
else
$\varphi = t_2(\rho+1) + \frac{s_2}{2}(\rho^2 - 1) + t_1(1-\rho) + \frac{s_1}{2}(1-\rho^2)$
end if
else
$\varphi = t_2(\rho+1) + t_1(1-\rho) + \frac{s_1}{2}(1-\rho^2)$
end if
end if
else
{The case $t_1 + s_1 < 0$}
if $(s_1 = s_2)$ or $(|\rho| > 1)$ **then**
{Boundaries of the polygons do not cross}
if $t_2 - s_2 > 0$ **then**
$\varphi = t_2\Big(1 - \frac{t_2}{s_2}\Big) + \frac{s_2}{2}\Big(\frac{t_2^2}{s_2^2} - 1\Big) + \Big|t_1\Big(1 + \frac{t_1}{s_1}\Big) + \frac{s_1}{2}\Big(1 - \frac{t_1^2}{s_1^2}\Big)\Big|$
else
$\varphi = \Big|t_1\Big(1 + \frac{t_1}{s_1}\Big) + \frac{s_1}{2}\Big(1 - \frac{t_1^2}{s_1^2}\Big)\Big|$

end if
else
{Boundaries cross}
if $t_2 - s_2 > 0$ **then**
if $\rho < -\frac{t_1}{s_1}$ **then**
if $(s_2 \neq 0)$ and $\Big(-\frac{t_2}{s_2} < 1\Big)$ and $\Big(-\frac{t_2}{s_2} > -\frac{t_1}{s_1}\Big)$ **then**
$\varphi = t_2(\rho+1) + \frac{s_2}{2}(\rho^2-1) - t_1\Big(\frac{t_1}{s_1}+\rho\Big) + \frac{s_1}{2}\Big(\frac{t_1^2}{s_1^2}-\rho^2\Big) + \Big|t_2\Big(1+\frac{t_2}{s_2}\Big) + \frac{s_2}{2}\Big(1-\frac{t_2^2}{s_2^2}\Big)\Big|$
else
$\varphi = t_2(\rho+1) + \frac{s_2}{2}(\rho^2-1) - t_1\Big(\frac{t_1}{s_1}+\rho\Big) + \frac{s_1}{2}\Big(\frac{t_1^2}{s_1^2}-\rho^2\Big)$
end if
else
$\varphi = t_2\Big(1-\frac{t_2}{s_2}\Big)+\frac{s_2}{2}\Big(\frac{t_2^2}{s_2^2}-1\Big)+\Big|t_1\Big(\rho+\frac{t_1}{s_1}\Big)+\frac{s_1}{2}\Big(\rho^2-\frac{t_1^2}{s_1^2}\Big)\Big|+\Big|t_2(1-\rho)+\frac{s_2}{2}(1-\rho^2)\Big|$
end if
else
if $\rho < -\frac{t_1}{s_1}$ **then**
$\varphi = t_2\Big(\rho+\frac{t_2}{s_2}\Big) + \frac{s_2}{2}\Big(\rho^2-\frac{t_2^2}{s_2^2}\Big) - t_1\Big(\frac{t_1}{s_1}+\rho\Big) + \frac{s_1}{2}\Big(\frac{t_1^2}{s_1^2}-\rho^2\Big)$
else
if $(s_2 = 0)$ or $\Big(\Big|\frac{t_2}{s_2}\Big| > 1\Big)$ **then**
$\varphi = \Big|t_1\Big(1+\frac{t_1}{s_1}\Big) + \frac{s_1}{2}\Big(1-\frac{t_1^2}{s_1^2}\Big)\Big| + \Big|t_2(1-\rho)+\frac{s_2}{2}(1-\rho^2)\Big|$
else
$\varphi = \Big|t_1\Big(1+\frac{t_1}{s_1}\Big) + \frac{s_1}{2}\Big(1-\frac{t_1^2}{s_1^2}\Big)\Big| + \Big|t_2\Big(\frac{t_2}{s_2}+\rho\Big) - \frac{s_2}{2}\Big(\frac{t_2^2}{s_2^2}-\rho^2\Big)\Big|$
end if
end if
end if
end if
end if

A.6.3 Central Limit Theorem

In this subsection we cite a version of a central limit theorem for set-indexed martingales. For this purpose, the martingales are considered as elements of a function space that is a set-indexed analogue of the Skorokhod space $D(\mathbb{R}_+)$. A precise definition of this space can be found in IM (p. 134). The study of set-indexed processes is greatly facilitated

by the concept of a *flow*, which provides a link to related theory for stochastic processes on $\mathbb{R}$ or $\mathbb{R}_+$, and in fact underpins the definition of convergence for the set-indexed martingales (see Definition A.8) that would usually, such as here, be established first. As noted before, functional convergence may not always be the suitable concept due to the discontinuity of Gaussian processes over classes of sets that are too large.

Definition A.7. *Let $\tilde{\mathcal{A}}(u)$ denote the class of countable intersections of sets in $\mathcal{A}(u)$, and let $a, b \in \mathbb{R}$ with $a < b$. A* flow *(on $[a,b]$) is an increasing function $f : [a,b] \to \tilde{\mathcal{A}}(u)$.*

Write $S(\mathcal{A})$ for the class of all *simple* flows (see IM, p. 148), and denote by $D[S(\mathcal{A})]$ the space of all additive $\mathcal{A}$-indexed processes X such that for every $f \in S(\mathcal{A})$, there exists a modification of $X \circ f$ which is in $D[0,1]$ and unique up to indistinguishability. (Two stochastic processes on $[0,1]$ are called indistinguishable if their sample paths are identical a.s.) In what follows, this version of $X \circ f$ will be denoted $M_f(X)$.

Definition A.8. *Let $X, X_1, X_2, \ldots \in \mathcal{D}[S(\mathcal{A})]$. The sequence (X_n) converges semi-functionally to X, denoted $X_n \to_{sf} X$, if $M_f(X_n) \Longrightarrow M_f(X)$ in $D[0,1]$, for every $f \in S(\mathcal{A})$.*

Remark A.1. *Connections between semi-functional and weak convergence.* If the semi-functional convergence to X is such that $M_f(X)$ is even *continuous* for every simple flow f, then the finite dimensional distributions of X_n converge (see IM, Proposition 7.3.7, p. 149). Convergence in distribution of X itself, with respect to the Skorohod-type metric defined in IM (p. 134), requires an additional condition on the approximability of X_n by a sum of purely atomic and continuous processes, given as condition 5 of IM (Theorem 9.1.4, pp. 180–181). We shall only need the following weaker result, given as Theorem 7.2.7 in IM (pp. 146–147), and which involves less technical conditions. The *Hausdorff distance* d_H is defined below at (A.8.2), or see IM (p. 19). Assume that

1. the finite dimensional distributions of X_n converge to those of X,
2. $\forall\, \eta > 0\ \exists\, a > 0$ such that $P\big(\sup\{X_n(A) : A \in \mathcal{A}\} > a\big) < \eta\ \forall\, n$, and
3. $\forall\, \eta > 0$ and $\epsilon > 0\ \exists\, \delta > 0$ such that $P\big(w(X_n, \delta) \geq \epsilon\big) < \eta\ \forall\, n$,

where $w(X_n, \delta) = \sup\big\{|X_n(A) - X_n(B)| : d_H(A,B) < \delta,\ A, B \in \mathcal{A}\big\}$. Then $X_n \Longrightarrow X$.

With the preparations from the previous subsection we are now able to formulate the CLT for set-indexed martingales as given in IM. In the formulation of Theorem A.5 we use the so-called *jump functional* of a process $y \in D[0,1]$, defined by $J(y) = \sup\big\{|y(s) - y(s-)| : 0 \leq s \leq 1\big\}$.

Theorem A.5. (Ivanoff and Merzbach, 2000, p. 176) *Let $(X_n, \mathcal{F}_n, P_n)$ be a sequence of strong martingales defined on the indexing collection $\mathcal{A}$, and let (Q_n) be any sequence of corresponding *-quadratic variation processes. Assume that*

1. $\sup_n E\left[|X_n(T)|^{2+\delta}\right] < \infty$ *for some* $\delta > 0$,
2. $J(M_f(X_n)) \to_P 0$ *as* $n \to \infty$ *for every simple flow* f,
3. $\{Q_n(T)\}$ *is uniformly integrable,*
4. *for every* $A \in \mathcal{A}$, $Q_n(A) \to_P \Lambda(A)$, *where* Λ *is a deterministic, increasing, monotone inner- and outer-continuous function* (IM, p. 27) *on* $\mathcal{A}$.

Then there exists a Brownian motion X defined on $\mathcal{A}$, based on the measure Λ, such that

$$X_n \to_{sf} X\,.$$

The combination of Theorem A.5 and Remark A.1, together with minor adjustments, yields the central limit theorem which will be instrumental in the proof of Theorem 4.2.

A.7 M-Estimators

Likelihood estimators are special instances of M-estimators, which are reviewed in the present section. The theory for the latter allows us to embed the results of Chapter 4, especially Theorem 4.2, in this context through identification of the limit as a functional of a multi-indexed Gaussian process with quadratic drift. We shall need an argmax-continuous mapping theorem which, after some general remarks in Subsection A.7.1, is stated in Subsection A.7.2.

A.7.1 Introduction

The boundary estimators which will be introduced in Chapters 3–5 are based on maximisation of a suitable criterion function, evaluated at a local part of the given data. Such estimators are called *M-estimators*, and their paramount importance among statistical estimators has been the main rationale for their extensive study. The prototype case is the ordinary (parametric) maximum-likelihood estimation where the criterion function is the logarithm of the product density of the sample. Specifically, if $\{X_1, \ldots, X_n\}$ are

i.i.d. with density p_θ, where θ ranges over a subset of $\mathbb{R}^k$ for some $k \in \mathbb{N}$, one seeks to maximise the function

$$\theta \mapsto F_n(\log p_\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \big(p_\theta(X_i)\big)\,, \tag{A.7.1}$$

where F_n is the empirical d.f. pertaining to $\{X_1, \ldots, X_n\}$. It is a folk theorem (see e.g. Schervish, 1995, pp. 418–424) that if the map at (A.7.1) satisfies regularity conditions, then the maximum likelihood estimator is asymptotically $\mathrm{N}(0, I_\theta^{-1})$ distributed, where I_θ is the Fisher information matrix. In an example in Van der Vaart and Wellner (1996, pp. 288–289) it is shown how to derive this result from quite far-reaching results on asymptotic behaviour of M-estimators of processes that satisfy an equicontinuity condition in expectation (in the sense of Section A.3); see equation (A.7.2). Note, however, that the main result for which we shall use M-estimation theory (Theorem 4.2 of Chapter 4) involves a decreasing sequence of bandwidths, and thus the theory on estimators of the form as at (A.7.1) cannot be applied.

We are concerned with the asymptotic behaviour of maximisers of functionals of Gaussian processes of the general type defined in Section A.4. However, we shall need only the case where the parameter set equals a subset of $\mathbb{R}^2$. In view of the central limit theory recapitulated in Section A.6, it is not surprising that the estimators in which we are interested are asymptotically defined via the location of the maximum of a Gaussian process, which exhibits strong (quadratic) negative drift away from the origin.

A.7.2 The Argmax-Continuous Mapping Theorem

In Subsection A.4.2 we have noted that under weak conditions, the maximiser of a Gaussian process gives rise to a well-defined and, in fact Borel-measurable, map from $\ell^\infty(T)$ into $\mathbb{R}^d$. In order to establish (modern) weak convergence of maximisers belonging to a sequence of stochastic processes with a Gaussian limit, one seeks to use an appropriate version of the continuous mapping theorem. Thus, it is necessary to impose a continuity condition on the (nonlinear) argmax functional,

$$\hat{t} = \mathrm{argmax}_{t \in T}\, f(t)\,, \qquad f \in \ell^\infty_{\mathrm{loc}}(T)\,,$$

with respect to the uniform metric on $\ell^\infty_{\mathrm{loc}}(T)$. However, it turns out that it is enough to impose the condition of *upper semicontinuity* on the stochastic processes in order to prevent erratic behaviour of the argmax functional which is defined through it. We quote

the following theorem from Van der Vaart and Wellner (1996), which is a refinement of a result given by Kim and Pollard (1990).

Theorem A.6. (Argmax-Continuous Mapping Theorem) (Kim and Pollard, 1990; Van der Vaart and Wellner, 1996, p. 286) *Let Γ_n, Γ be stochastic processes indexed by a metric space T such that $\Gamma_n \Longrightarrow \Gamma$ in $\ell^\infty(K)$ for every compact $K \subseteq T$. Suppose that almost all sample paths $t \mapsto \Gamma(t)$ are upper semicontinuous and possess a unique maximum at a (random) point $\hat{t}$, which as a random map in T is tight. If the sequence $\{\hat{t}_n\}$ is uniformly tight and satisfies $\Gamma_n(\hat{t}_n) \geq \sup_t \Gamma_n(t) - o_P(1)$, then $\hat{t}_n \Longrightarrow \hat{t}$ in T.*

In the study of the limit behaviour of maximisers, the next question of interest is the rate of convergence. To this end, it is customary to consider a rescaling of the original process, Γ_n say, as

$$\widehat{\Gamma}_n(t) \equiv \Gamma_n(\hat{t} + r_n^{-1}t) - \Gamma_n(\hat{t})\,. \tag{2.4.5}$$

Then, if $\hat{t}_n$ is assumed to exactly maximise Γ_n, it is immediate that

$$r_n(\hat{t}_n - \hat{t}) = \operatorname{argmax}_t \widehat{\Gamma}_n(t)\,, \quad \text{for all } n.$$

Hence, $r_n^{-1} \searrow 0$ may be seen as (an upper bound to) the rate of convergence of the sequence $\{\hat{t}_n\}$. This setup is envisaged in the application of the next theorem we cite, formulated with somewhat more restrictive conditions and adapting notation to our own. The critical condition to check there is the bound on the expected continuity modulus of $\Gamma_n - \Gamma$. Note that Theorem A.7 makes only minimal assumptions on the 'distance' d. Moreover, its last statement makes it possible to use it in parallel, rather than in sequence, to Theorem A.6. This will be the situation in which we shall use the two theorems in Section 4.3.2.

Theorem A.7. (Van der Vaart and Wellner, 1996, pp. 289–290) *Let Γ_n be stochastic processes indexed by a metric space T, and $\Gamma : T \to \mathbb{R}$ be a deterministic function with unique maximiser $\hat{t}$. Suppose that there exists $\delta_0 \in (0, \infty]$, a universal constant $C_1 > 0$, and a function $d : T \to \mathbb{R}_+$ such that for all δ in a neighbourhood of δ_0,*

$$\Gamma(t) - \Gamma(\hat{t}) \leq -C_1\, d^2(t, \hat{t})\,.$$

Furthermore, assume that there exists a constant $C_2 > 0$ such that for every n and all sufficiently small $\delta > 0$, the process $\Gamma_n - \Gamma$ satisfies

$$E^* \sup_{t : d(t,\hat{t}) < \delta} \left|(\Gamma_n - \Gamma)(t) - (\Gamma_n - \Gamma)(\hat{t})\right| \leq C_2\, \frac{\phi_n(\delta)}{n}\,, \tag{A.7.2}$$

where the functions $\phi_n : (0, \delta_0) \to (0, \infty)$ *are such that* $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ *is decreasing for some* $\alpha < 2$ *(not depending on* n*). Let* $\{r_n\}$ *be a sequence of positive constants with the properties* $r_n^{-1} \in (0, \delta_0)$ *and*

$$r_n^2 \, \phi_n\left(\frac{1}{r_n}\right) \leq \sqrt{n}\,, \qquad \textit{for every } n\,.$$

If the sequence $\hat{t}_n$ *satisfies* $\Gamma_n(\hat{t}_n) \geq \Gamma_n(\hat{t}) - O_P(r_n^{-2})$ *and converges in outer probability to* $\hat{t}$*, then* $r_n \, d(\hat{t}_n, \hat{t}) = O_P^*(1)$*. If the displayed conditions are valid for every* t *and* δ*, then the condition that* $\hat{t}_n$ *is consistent is unnecessary.*

The third step in the approach that Van der Vaart and Wellner (1996) suggest (p. 288) consists in establishing the limit distribution of the relevant M-estimator, normalised by the appropriate rate as may be obtained by Theorem A.7. Results in that direction from their monograph pertain to empirical processes where decreasing bandwidths play no role. Hence, in Section 4.3.2 the limiting form of the estimator as the functional of a Gaussian process with drift will be established by direct arguments.

A.8 Minimax Theory and Optimal Convergence Rates

The main goal of the present section is to give a definition of the mean-square and several other relevant distance measures in Subsection A.8.1, and subsequently introduce the risk function used in minimax theory recapitulated in Subsection A.8.2. As evinced by the formula of the key result (Theorem A.8), the convergence rate approaches the parametric rate n^{-1}, known from change-point problems on the line, as the regression surface and the edge simultaneously belong to a higher smoothness class. For our own estimators, and with our smoothness assumptions, this optimal benchmark rate becomes $O(n^{-2/3})$.

A.8.1 Distance Measures for Boundary Curves

In the formulation of the edge estimation problem as one pertaining to curves, as is done in this thesis, an important problem is the specification of the underlying distance or *similarity measure* (s.m.) between curves which define boundaries. We often avoid the term 'metric' such as here, as it turns out that many practically relevant s.m.s fail to satisfy the positive definiteness condition, or even the symmetry in its arguments. It is obvious that closeness and consistency of boundary curves crucially depend on the s.m. that is being used, and to the extent to which it satisfies natural postulates. A

plausible set of such postulates will be given shortly. Unlike in conventional statistical problems of point estimation, where Euclidean distance usually suffices, there is a range of competing s.m. s which fulfil the postulates to varying degrees. In this subsection we briefly review some of the most common s.m.s. We also introduce the s.m. that is implied by our method of confidence band estimation in Chapter 4.

As we noted in passing in Chapter 1, edge detection may frequently be seen as the first step in reconstructing a more complex image. For example, the edge may arise as a binary image, with the edge comprising the contour line in a preliminary thresholding procedure; the contour line then of course completely characterises the planar set it encloses. Thus, each s.m. on the set of images (possibly with regularity constraints) gives rise to an s.m. on the corresponding set of contours or edges. Although such s.m.s apply directly to closed curves only, they may readily be adapted to situations where only sections of curves are to be estimated, although with the absence of natural coordinate axes, there will usually be no canonical way of doing this. For defining s.m.s on images, L^p-type distances, especially the cases $p = 1$ and $p = 2$, are of special importance.

Hagedoorn and Veltkamp (1999) list desirable properties of s.m.s pertaining to patterns (which are, in the present context, either images or the boundaries defining them). We redisplay these criteria for s.m.s, with some minor alterations.

Desirable properties of similarity measures:

- Fulfilment of the properties of a metric.
- Linearity in the reciprocal of scale, and invariance with respect to translation and rotation.
- Robustness for perturbation.
- Robustness for partial occlusion.

Marron and Tsybakov (1995) discuss several "visual error criteria" which show that the visual concept of distance is often inappropriately reflected by the metric being used, and that different concepts (possibly not requiring some of the properties of a metric) may yield more satisfying results. Apart from its independent theoretical appeal, the second point in the above list may be seen as responsible for the potential improvement provided by those alternatives.

Next we list and briefly discuss three basic s.m.s which will be important in this thesis. These apply mainly to parametrised curves, finite point sets, and planar regions respec-

tively. Our references are Marron and Tsybakov (1995) and Hagedoorn and Veltkamp (1999).

The Fréchet distance. This notion of distance generalises the idea that underlies the definition of the Skorohod metric on $D[0,1]$. The definition given here considers Jordan curves only, but this entails no real loss of generality. Recall that Jordan curves are the homeomorphic images of the unit sphere S^1. Now, for any two Jordan curves $\mathcal{C}_1, \mathcal{C}_2$, the Fréchet distance is given by

$$d_{\mathcal{F}}(\mathcal{C}_1, \mathcal{C}_2) = \inf \left\{ \|\phi_1 - \phi_2\|_\infty : \phi_1 \in \mathrm{Hom}(S^1, \mathcal{C}_1),\, \phi_2 \in \mathrm{Hom}(S^1, \mathcal{C}_2) \right\} .$$

This defines a semimetric which is invariant under isometries. It is also robust for small deformations such as humps. Such deformations would have a potentially much more serious effect on distance if instead of the infimum in the above definition, one took specific functions ϕ_i $(i = 1, 2)$ such as, for example, rescaled versions of the natural parametrisations of the $\mathcal{C}_i$, calculated from intersection points of these curves with a given line. The Fréchet distance is fairly expensive to compute: for two polygonal curves with m and n defining points, it requires $O(mn \log mn)$ arithmetic operations. The logarithmic factor disappears when it is only to be determined whether the two curves are less than $\epsilon > 0$ apart with respect to $d_{\mathcal{F}}$.

The Hausdorff distance. This is a distance measure (and in fact a metric) which has been used much more widely in the literature than the Fréchet distance. Let $\mathcal{K}$ denote the class of nonempty compact subsets of the plane. The *directed Hausdorff distance* of the nonempty sets $A, B \in \mathcal{K}$ is defined as

$$\vec{d}_H(A, B) = \max \{ d(a, B) : a \in A \} , \tag{A.8.1}$$

where $d(x, M) = \min\{\|x - m\| : m \in M\}$ denotes the distance between x and the closed (nonempty) set M. Marron and Tsybakov (1995) presented this as their "visual error criterion" VE_∞; they considered A and B as the sets below graphs of functions, but do point out the extension to the general case. The *Hausdorff metric* on $\mathcal{K}$ is defined as

$$d_H(A, B) = \max \left\{ \vec{d}_H(A, B),\, \vec{d}_H(B, A) \right\} , \tag{A.8.2}$$

which corresponds to the SE_∞ (where 'S' stands for 'symmetric') error criterion of Marron and Tsybakov. Computation of $d_H(A, B)$ for finite sets A, B is known to be executable in $O((m + n) \log(m + n))$ time, where $m = \mathrm{card}(A)$, $n = \mathrm{card}(B)$. We also point to the modification of the Hausdorff distance that is presented in Qiu (2002b), who argues that

the modification is better suited to image processing applications where outlier pixels may render d_H awkward to use.

The distance of symmetric difference. This is defined as the area (two-dimensional Lebesgue measure) of the set-theoretic symmetric difference:

$$d_1(A, B) = \|A \triangle B\| . \tag{A.8.3}$$

The s.m. d_1 is indeed a distance if A and B are *solid sets*; a set C is called solid if $(C^\circ)^- = C$. In the notation d_1, the subscript 1 indicates the relationship to the functional L^1 norm of the associated boundary fragments (if existent) as defined in Subsection A.8.2, where the risk function will be the expected value of the L^2 distance. Note that it is problematic to define a metric d_2, in analogy to d_1 at (A.8.3), via the functional L^2 norm. Therefore, in order to formulate the results of Subsection A.8.2, a piecewise representation as a function in Cartesian coordinates is employed for the edge. In global assessment of edge estimator performance, the generally meandering nature of the fault lines considered in this thesis necessitates the use of the above set-theoretic, rather than the classical functional, version of the various metrics.

A.8.2 Decision Theory for Curve Estimators

In this subsection we develop, in a now conventional manner, the decision-theoretic underpinnings that are used to assess, and especially compare, the asymptotic performance of boundary estimators. The use of decision theory requires the specification of the class of images to which the true model is assumed to belong, and from which the estimator is chosen; and selection of a suitable risk function, which is most commonly (as here) taken to be the minimax risk. We confine ourselves to a brief outline of the methodology and refer to Korostelev (1991) and Korostelev and Tsybakov (1993, 1994) for more details.

Because of the assumption of compactness of the image, it entails no loss of generality to restrict attention to images which are subsets of the unit cube $I_2 = [0, 1]^2$. Following Korostelev and Tsybakov (1993, 1994), we define the class of *boundary fragments* as those images whose domains admit a representation

$$\Gamma = \left\{ x = (x^{(1)}, x^{(2)}) \in I_2 : \, 0 \le x^{(2)} \le g(x^{(1)}) \right\} ,$$

with $g \in \Sigma(\beta, L, h)$, where

$$\Sigma(\beta, L, h) = \Sigma(\beta, L) \cap \left\{ g : \, g(x^{(1)}) \in [h, 1-h], \, x^{(1)} \in [0, 1] \right\}$$

for $h \in (0, \frac{1}{2})$, and $\Sigma(\beta, L)$ is the class of univariate Hölder continuous functions of order $\beta \geq 1$ with Hölder constant $L \in \mathbb{R}_+^*$. The following regularity assumptions from Korostelev and Tsybakov (1993, 1994) define the class of domains which will be considered. They contain the class of boundary fragments, and it is a helpful guiding principle, from both the theoretical and practical viewpoints, to split a domain of the below defined class into boundary fragments.

Let $\mathcal{G}$ denote the class of domains defined by the following properties (see Korostelev and Tsybakov, 1993, p. 143):

1. All members of $\mathcal{G}$ are closed connected compact subsets of I_2. Euclidean distances between the edges and the boundary of I_2 are bounded away from zero uniformly over $\mathcal{G}$.

2. There exists a universal constant $C_1 \in \mathbb{R}_+^*$ such that $|\Gamma| = \text{length}(\Gamma) \leq C_1$ for any edge Γ in the class.

3. In the natural parametrisation $\Gamma = \left(x^{(1)}(s), x^{(2)}(s)\right)$, $0 \leq s \leq |\Gamma|$, the functions $x^{(1)}(s/|\Gamma|), x^{(2)}(s/|\Gamma|)$ belong to the class $\Sigma(\gamma, L, h)$ with some fixed γ, L, h.

4. The curve Γ has no singular points, i.e. there exists a constant $C_2 > 0$ such that $\|\dot{x}(s)\| \geq C_2$ for all $s \in [0, |\Gamma|]$.

5. $\gamma \geq 2$, so that curvature is uniformly bounded.

For any $\beta \in \mathbb{R}_+^*$ and γ as above, a class of *grey-scale images* $\Phi_{\beta,\gamma} \subseteq \{f : \text{ domain}(f) = I_2\}$ is defined by those functions with a representation that generalises equation (1.2.2):

$$f(x) = f_0(x) I\{x \notin G\} + f_1(x) I\{x \in G\}\,,$$

where G denotes an arbitrary element of $\mathcal{G}$,

$$\inf_{x \in G} f_1(x) \geq b\,, \quad \sup_{x \notin G} f_0(x) \leq a\,, \quad 0 < a < b < 1\,, \tag{A.8.4}$$

and f_i are bivariate Hölder continuous functions of order β. In (A.8.4), the inequality $b < 1$ is introduced merely for technical reasons, and may be achieved by rescaling the response variable. Grey-scale boundary fragments are defined analogously.

In line with the general assumptions of this thesis, the observations are assumed to be of the form (X_i, Y_i). As before, $\mathcal{X} = \mathcal{X}_n = \{X_i\}$ denotes the explanatory data, either located on a grid or uniformly distributed over the unit square I_2; and the responses Y_i

are assumed to follow the additive model given at (1.2.3), with the additional requirement that the errors ϵ_i be independent $\mathrm{N}(0, \sigma^2)$ distributed r.v.s. The best studied risk function is the *integrated* (*square*) *risk*, which is given by

$$R(f, T_n) = R_I(f, T_n) = E\left(\int \{f(x) - T_n(x)\}^2\right) dx\,,$$

where T_n is an image estimator (see Section 1.1), and the expectation is with respect to the joint d.f. of the pairs (X_i, Y_i). Assessment of a given edge estimator is now treated as an image estimation problem, as follows.

Definition A.9. (Korostelev and Tsybakov, 1994, p. 50) *A positive sequence ψ_n is called a* minimax rate of convergence *on the class of images Φ if for all sufficiently large n and some constants $C_0, C_1 \in \mathbb{R}_+^*$,*

$$C_0 \leq \inf_{T_n} \sup_{f \in \Phi} \psi_n^{-2} R(f, T_n) \leq C_1\,,$$

where Φ is a class of images such as, for example, the previously defined classes $\Phi_{\beta,\gamma}$ for some β, γ. An estimator f_n is called an optimal image estimator *if for all n,*

$$\sup_{f \in \Phi} \psi_n^{-2} R(f, f_n) \leq C_1\,,$$

where ψ_n is the minimax rate.

A.8.3 Minimax-Optimal Convergence Rates

The minimax-optimal convergence rates for image models in which we are interested have been obtained by Korostelev and Tsybakov (1993, 1994), whose results we display below; see also Mammen and Tsybakov (1995).

Theorem A.8. (Korostelev and Tsybakov, 1994, p. 65) *Assume that model* (1.2.3) *holds, and that the previously introduced notation is retained. Then, the minimax-optimal convergence rate for the model with images from the class $\Phi_{\gamma,\beta}$ is given by*

$$\psi_n = \max\left\{n^{-\gamma/(2(\gamma+1))}, n^{-\beta/(2(\beta+1))}\right\}\,, \gamma \geq 1\,,\ \beta \in \mathbb{R}_+^*$$

in the case of Poisson-distributed design, and by $n^{-1/2}$ in the case of gridded design, if the metric of symmetric difference is being used. For the Hausdorff metric, the (slower) rates are given by replacing n with $n/\log n$ in all previous statements.

Korostelev and Tsybakov (1993, 1994) also gave an explicit construction of an estimator that achieves these optimal rates. This construction, however, seems not easy to implement numerically. In the next three chapters we shall develop estimators which, while not nearly suffering from this drawback to the same extent, come within a logarithmic, or at least an arbitrarily small polynomial, factor of the minimax-optimal convergence rate.

References

Abramowitz, M. and Stegun, I.A. (1970). *Handbook of Mathematical Functions.* Dover, NY.

Acker, A.E. (1988). *How to Speak Radar: Basic Fundamentals and Applications of Radar.* Varian Associates, Palo Alto, CA.

Adler, R.J. (1990). *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes.* Institute of Mathematical Statistics Lecture Note Series **12**. Institute of Mathematical Statistics, Hayward, CA.

Ahmed, N. and Rao, K.R. (1975). *Orthogonal Transforms for Digital Signal Processing.* Springer, Berlin.

Alexander, K.S. and Pyke, R. (1986). A uniform central limit theorem for set-indexed partial-sum processes with finite variance. *Ann. Prob.* **14**, 582–597.

Alt, H. and Godau, M. (1995). Computing the Frechèt distance between two polygonal curves. *Int. J. of Computational Geometry and Applications* **5**, 75–91.

Altman, N.S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.* **85**, 749–759.

Altman, N.S. (1993). Estimating error correlation in nonparametric regression. *Statist. Prob. Lett.* **18**, 213–218.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, NY.

Burbeck, C.A. and Pizer, S.M. (1995). Object representation by cores: Identifying and representing primitive spatial regions. *Vision Research* **35**, 1917–1930.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Patt. Anal. Mach. Intell.* **8**, 679–698.

Carlstein, E., Müller, H.-G. and Siegmund, D. (1994). *Change-Point Problems.* Institute of Mathematical Statistics Lecture Note Series **23**. Institute of Mathematical Statistics, Hayward, CA.

Chernoff, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16**, 31–41.

Coifman, R.R. and Donoho, D.L. (1995). Translation-Invariant De-Noising. In: Antoniadis, A. and G. Oppenheim, Eds. (1995), Wavelets and statistics, Lecture Notes in Statistics **103**. Springer-Verlag, NY.

Copas, J.B. (1995). Local likelihood based on kernel censoring. *J. Roy. Statist. Soc. Ser. B* **57**, 221–235.

Courmontagne, Ph. (1999). A new formulation for the Karhunen-Loeve expansion. *Signal Processing* **79**, 235–249.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*, 2nd Edn. Wiley, NY.

Daniels, H.E. and Skyrme, T.H.R. (1985). The maximum of a random walk whose mean path has a maximum. *Adv. Appl. Prob.* **17**, 85–99.

Deprins, D., Simar, L. and Tulkens, H. (1984). Measuring labor inefficiency in post offices. In *The Performance of Public Enterprises: Concepts and measurements*, eds. M. Marchand, P. Pestieau and H. Tulkens, pp. 243–267. North-Holland, Amsterdam.

Do Carmo, M. P. (1976). *Differential Geometry of Curves and Surfaces.* Prentice-Hall Inc., Englewood Cliffs, NJ.

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Journ. Roy Statistical Society,* Series B, **81**, 425–455.

Dudley, R.M. (1966). Weak convergence of measures on nonseparable metric spaces and empirical measures on Euclidean spaces. *Illinois Journ. Math.* **10**, 109–126.

Dudley, R.M. (1973). Sample functions of the Gaussian process. *Ann. Probab.* **1**, 66–103.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance.* Springer, NY.

Etchison, T., Patula, S.G. and Brownie, C. (1994). Partial autocorrelation function for spatial processes. *Statist. Prob. Lett.* **21**, 9–19.

Eubank, R.L. and Speckman, P.L. (1994). Nonparametric estimation of functions with jump discontinuities. In: *Change-Point Problems*, Eds. E. Carlstein, H.-G. Müller and D. Siegmund, IMS Lecture Notes **23**, pp. 130–144. Institute of Mathematical Statistics, Hayward, CA.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

Ferger, D. (1999). On the uniqueness of maximizers of Markov-Gaussian processes. *Statist. Prob. Lett.* **45**, 71–77.

Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics.* Wiley, NY.

Gasser, T. and Kneip, A. (1995). Searching for structure in curve samples. *J. Amer. Statist. Assoc.* **90**, 1179–1188.

Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86**, 643–652.

Gayraud, G. and Tsybakov, A.B. (2002). Testing hypotheses about contours in images. *J. Nonparam. Statist.* **14**, 67–85.

Gerbrands, J.J. (1981). On the relationships between SVD, KLT and PCA. *Pattern Recognition* **14**, 375–381.

Ghanem, R.G. and Spanos, P.D. (1991). *Stochastic Finite Elements: A Spectral Approach.* Springer, NY.

Gibbins, D., Gray, D.A. and Dempsey, D. (1999). Classifying Ships Using Low Resolution Maritime Radar. Proceedings of the Fifth International Symposium on Signal Processing and Applications (ISSPA'99), Brisbane, Australia, August, 325–328.

Gijbels, I., Hall, P. and Kneip, A. (1999) On the estimation of jump points in smooth curves. *Ann. Inst. Statist. Math.* **51**, 231–251.

Gijbels, I., Mammen, E., Park, B.U. and Simar, L. (1999). On estimation of monotone and concave frontier functions. *J. Amer. Statist. Assoc.* **94**, 220–228.

Godtliebsen, F. and Sebastiani, G. (1994). Statistical methods for noisy images with discontinuities. *J. Appl. Statist.* **21**, 459–477.

Groeneboom, P. (1985). Estimating a monotone density. In: *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, Vol. II, Eds. E. Carlstein, H.-G. Müller and D. Siegmund, pp. 539–555.

Groeneboom, P. (1989). Brownian motion with a parabolic drift and Airy functions. *Probab. Th. Rel. Fields* **81**, 79–109.

Groeneboom, P. and Wellner, J.A. (2001). Computing Chernoff's distribution. *J. Comput. Graph. Statist.* **10**, 388–400.

Hagedoorn, M. and Veltkamp, R.C. (1999). Reliable and efficient pattern matching using an affine invariant metric. *Int. J. of Computer Vision* **31**, 203–225.

Hall, P. (1979). On the rate of convergence of normal extremes. *J. Appl. Probab.* **16**, 433–439.

Hall, P. (1982). On some simple estimates of an exponent of regular variation. *J. Roy. Statist. Soc. Ser. B* **44**, 37–42.

Hall, P. (1988). *Introduction to the Theory of Coverage Processes.* Wiley, NY.

Hall, P. and Owen, A.B. (1993). Empirical confidence bands in density estimation. *J. Comput. Graph. Statist.* **2**, 273–289.

Hall, P., Park, B.U. and Stern, S. E. (1998). On polynomial estimators of frontiers and boundaries. *J. Multivariate Anal.* **66**, 71–98.

Hall, P., Peng, L. and Rau, C. (2001). Local-likelihood tracking of fault lines and boundaries in spatial problems. *J. Roy. Statist. Soc. Ser. B* **63**, 569–582.

Hall, P., Poskitt, D. and Presnell, B. (2001). A functional data-analytical approach to signal discrimination. *Technometrics* **43**, 1–9.

Hall, P., Qian, W. and Titterington, D.M. (1992). Ridge finding from noisy data. *J. Comput. Graph. Statist.* **1**, 197–211.

Hall, P., Qiu, P. and Rau, C. (2002). Tracking edges, cornes and vertices in an image. Manuscript.

Hall, P. and Raimondo, M. (1997a). Approximating a line thrown at random onto a grid. *Ann. Appl. Probab.* **7**, 648–665.

Hall, P. and Raimondo, M. (1997b). Measuring the performance of boundary-estimation methods. In: *L_1-Statistical Procedures and Related Topics*, Proceedings of the Third International Conference on the L_1-Norm and Related Methods, Ed. Y. Dodge. IMS

Lecture Notes–Monograph Series, vol. 31, pp. 1–14. Institute of Mathematical Statistics, Hayward, CA.

Hall, P. and Raimondo, M. (1998). On global performance of approximations to smooth curves using gridded data. *Ann. Statist.* **26**, 2206–2217.

Hall, P., and Rau, C. (2000). Tracking a smooth fault line in a regression surface. *Ann. Stat.* **28**, 713–733.

Hall, P., and Rau, C. (2002). Likelihood-based confidence bands for fault lines in response surfaces. *Probab. Th. Rel. Fields* **124**, 26–49.

Haralick, R.M. (1983). Ridges and valleys on digital images. *Computer Graphics and Image Process.* **22**, 28–38.

Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82**, 267–270.

Hermann, E. (1997). Local bandwidth choice in kernel regression. *J. Comput. Graph. Statist.* **6**, 35–54.

Hirst, D.S. (2002). *Some Statistical Problems in Testing the Modality of Regression Curves.* PhD Thesis, The Australian National University.

Hjort, N.L. (1994). Minimum L2 and robust Kullback-Leibler estimation. In *Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Ed. P. Lachout and J.Á. Víšek, pp. 102–105. Prague: Academy of Sciences of the Czech Republic.

Hjort, N.L. and Jones, M.C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24**, 1619–1647.

Hoffmann-Jørgensen, J. (1991). Stochastic Processes on Polish Spaces. *Various Publication Series* **39**. Aarhus Universitet, Aarhus, Denmark.

Huber, P.J. (1981). *Robust Statistics.* Wiley, NY.

Huertas, A. and Medioni, G. (1986). Detection of intensity changes with subpixel accuracy using Laplacian–Gaussian masks. *IEEE Trans. Patt. Anal. Mach. Intell.* **8**, 651–664.

Huxley, M.N. (1996). *Area, Lattice Points and Exponential Sums.* London Mathematical Society Monographs **13**, Oxford University Press.

Inggs, M.R. and Robinson, A.D. (1999). Ship Target Recognition using low resolution radar and neural networks. *IEEE Transactions on Aerospace and Electronic Systems* **35**, 386–392.

Ingster, Y.I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. *Math. Methods of Statist.* **(2–4)** , 85–114, 171–189, 249–268.

Ivanoff, G. and Merzbach, E. (2000). *Set-Indexed Martingales.* Chapman and Hall/CRC, Boca Raton, FL.

Jain, A.K. (1989). *Fundamentals of Digital Image Processing.* Prentice Hall, NJ.

Jansen, M. and Bultheel, A. (1999). Multiple Wavelet Threshold Estimation by Generalized Cross Validation for Images with Correlated Noise. *IEEE Transactions on Image Processing* **8**, 947–953.

Johnstone, I.M. and Silverman, B.W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B* **59**, 319–351.

Jones, R.H. and Stewart, R.C. (1997). A method for determining significant structures in a cloud of earthquakes. *J. Geophysical Res.* **102**, 8245–8254.

Khmaladze, E., Mnatsakanov R. and Toronjadze, N. (2002). The change set problem and local covering numbers. Manuscript.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Stat.* **18**, 191–219.

Kingsbury, N. G. (1999). Image Processing with Complex Wavelets. *Philosophical Transactions of the Royal Society of London* A, on a Discussion Meeting on "Wavelets: the key to intermittent information?", London, 1999.

Kittler, J. and Young, P.C. (1973). A new approach to feature selection based on the Karhunen-Loeve expansion. *Pattern Recognition* **5**, 335–352.

Kneip, A., Park, B.U. and Simar, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory* **14**, 783–793.

Koenderink, J.J. and Richards, W. (1988). Two-dimensional curvature operators. *J. Opt. Soc. Am. Ser. A* **5**, 1136–1141.

Kohlmann, K. (1996). Corner detection in natural images based on the 2-D Hilbert transform. *Signal Processing* **48**, 225–234.

Korostelev, A.P. (1991). Minimax reconstruction of two-dimensional images. *Theory Probab. Appl.* **36**, 153–159.

Korostelev, A.P. and Tsybakov, A.B. (1993). *Minimax Theory of Image Reconstruction.* Springer Lecture Notes in Statistics **82**. Springer-Verlag, Berlin.

Korostelev, A.P. and Tsybakov, A.B. (1994). Asymptotically minimax image reconstruction problems. In: *Topics in Nonparametric Estimation*, Ed. R.Z. Khasminskii, pp. 45–86. AMS, Providence, RI.

Lachenbruch, P.A. (1975). *Discriminant analysis.* Hafner Press, London.

Leadbetter, M.R., Lindgren, G. and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, NY.

Li, K.-C. and Shedden, K. (2002). Identification of shared components in large ensembles of time series using dimension reduction. *J. Amer. Statist. Assoc.* **97**, 759–765.

Liu, X. and Ehrich, R. (1995) Subpixel edge location in binary images using dithering. *IEEE Trans. Patt. Anal. Mach. Intell.* **17**, 629–634.

Lifshits, M.A. (1982). On the absolute continuity of distributions of functionals of random processes. *Theory Probab. Appl.* **27**, 600–607.

Lindeberg, T. (1998). Edge detection and ridge detection with automatic scale selection. *Internat. J. Computer Vision* **30**, 117–154.

Loader, C.R. (1996). Local likelihood density estimation. *Ann. Statist.* **24**, 1602–1618.

Loève, M. (1963). *Probability Theory.* Van Nostrand, NY.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing.* 2nd Edition, Academic Press, San Diego, CA.

Mammen, E. and Tsybakov, A.B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23**, 502–524.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, London.

Marcel, B. and Cattoen, M. (1997) Edge and line detection in low level analysis. In *Third Workshop on Electronic Control and Measuring Systems*, pp. 89–97. Université Paul Sabatier, Toulouse.

Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proc. Roy. Soc. London, Ser. B* **207**, 187–217.

Marr, D. and Poggio, T. (1979). A computational theory of human stereo vision. *Proc. Roy. Soc. London, Ser. B*, **204**, 301–328.

Marr, D. (1982). *Vision.* W.H. Freeman, San Francisco, CA.

Marron, J.S. and Tsybakov, A.B. (1995). Visual error criteria for qualitative smoothing. *J. Amer. Statist. Assoc.* **90**, 499–507.

Mendel, J.M. and Fu, K.S. (1970). *Adaptive Learning and Pattern Recognition Systems.* Academic Press, NY.

Mokhtarian, F. and Mackworth, A. (1986). Scale-based description and recognition of planar curves and two-dimensional shapes. *IEEE Trans. Patt. Anal. Mach. Intell.* **8**, 34–43.

Mokhtarian, F. and Suomela, R. (1998). Robust image corner detection through curvature scale space. *IEEE Trans. Patt. Anal. Mach. Intell.* **20**, 1376–1381.

Müller, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.* **20**, 737–761.

Müller, H.-G. and Song, K.S. (1994). Maximin estimation of multidimensional boundaries. *J. Multivar. Anal.* **50**, 265–281.

Munkres, J.R. (2000). *Topology*, 2nd edition. Prentice Hall, NJ.

Nalwa, V.S. and Binford, T.O. (1986). On detecting edges. *IEEE Trans. Patt. Anal. Mach. Intell.* **8** 699–714.

Opsomer, J.-D. (1995). Estimating a function by local linear regression when the errors are correlated. Preprint 95-42, Department of Statistics, Iowa State University.

Opsomer, J.-D. (1997). Nonparametric regression in the presence of correlated errors. In: *Modelling Longitudinal and Spatially Correlated Data*. Springer Lecture Notes in Statistics **122**, 339–348. Springer-Verlag, Berlin.

Opsomer, J.-D., Wang, Y. and Yang, Y. (1999). Nonparametric regression with correlated errors. *Stat. Science* **16**, 134–153.

O'Sullivan, F. and Qian, M. (1994). A regularised contrast statistic for object boundary estimation – implementation and statistical evaluation. *IEEE Trans. Patt. Anal. Mach. Intell.* **16**, 561–570.

Park, B.U., Sickles, R.C. and Simar, L. (1998). Stochastic panel frontiers: a semiparametric approach. *J. Econometrics* **84**, 273–301.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, NY.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters — an excess mass approach. *Ann. Statist.* **23**, 855–881.

Prakasa Rao, B.L.S. (1969). Estimation of a unimodal density. *Sankhyā A* **31**, 23–36.

Preparata, F.P. and Shamos, M.I. (1975). *Computational Geometry: An Introduction*. Springer, NY.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edition. Cambridge University Press.

Qiu, P. and Yandell, B. (1997). Jump detection in regression surfaces. *J. Computat. Graph. Statist.* **6**, 332–354.

Qiu, P. (1997). Nonparametric estimation of jump surface. *Sankhyā A* **59**, 268–294.

Qiu, P. (1998). Discontinuous surfaces fitting. *Ann. Statist.* **26**, 2218–2245.

Qiu, P. (2002a). A nonparametric procedure to detect jumps in regression surfaces. *J. Computat. Graph. Statist.* **11**, 799–822.

Qiu, P. (2002b). Jump-preserving surface reconstruction from noisy data. Manuscript.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*. Springer, NY.

Ramsay, J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer, NY.

Rice, J.A. and Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53**, 233–243.

Robinson, A.D. (1996). *Ship Target Recognition*. M.Sc. Thesis, University of Cape Town.

Rousseuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, NY.

Rudemo, M. and Stryhn, H. (1994). Approximating the distribution of maximum likelihood contour estimators in two-region images. *Scand. J. Statist.* **21**, 41–55.

Schervish, M.J. (1995). *Theory of Statistics.* Springer, NY.

Seiford, L.M. (1996). Data envelopment analysis: the evolution of the state-of-the-art, 1978–1995. *J. Productivity Anal.* **7**, 99–138.

Sharma, S. (1996). *Applied Multivariate Techniques.* Wiley, NY.

Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, NY.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.

Simar, L. (1996). Aspects of statistical analysis in DEA–type frontier models. *J. Productivity Anal.* **7**, 177–185.

Simar, L. and Wilson, P. (1998). Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. *Management Science* **44**, 49–61.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics.* Springer, NY.

Sinha, S.S. and Schunck, B.G. (1992). A 2-stage algorithm for discontinuity-preserving surface reconstruction. *IEEE Trans. Patt. Anal. Mach. Intell.* **14**, 36–55.

Smolka, B. and Wojciechowski, K.W. (2001). Random walk approach to image enhancement. *Signal Processing* **81**, 465–482.

Stryhn, H. (1993). Spatial change point models applied to image segmentation. PhD dissertation, Department of Mathematics and Physics, Agricultural University, Copenhagen.

Titterington, D.M. (1985a). Common structure of smoothing techniques in statistics. *Internat. Statist. Rev.* **53**, 141–170.

Titterington, D.M. (1985b). General structure of regularization procedures in image reconstruction. *Astronom. Astrophys.* **144**, 381–387.

Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* Springer, NY.

Venter, J.H. (1967). On estimation of the mode. *Ann. Math. Statist.* **38**, 1446–1455.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

Wang, Y. (1998). Change curve estimation via wavelets. *J. Amer. Statist. Assoc.* **93**, 163-172.

Wegman, E.J. (1971). A note on the estimation of the mode. *Ann. Math. Statist.* **42**, 1909–1915.

Xu, G. and Zhang, Z. (1996). *Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach.* Kluwer, NY.

Yao, Y.-C. (1987). Approximating the distribution of the maximum likelihood estimate of the change-point in a sequance of independent random variables. *Ann. Math. Statist.* **15**, 1321–1328.

Zaanen, A.C. (1956). *Linear Analysis.* North Holland, Amsterdam.