DEVELOPING A SPLINE – SMOOTHED DENSITY

A thesis submitted in partial fulfilment of a Degree of Bachelor of Science with Honours

Department of Mathematics

Australian National University

Giles Hooker

Abstract

The problem of estimating a probability density function for a data set differs significantly from functional estimation in which a response variable is present and has for this reason been treated by substantially different methods. Here we seek to forge a connection between the two problems. We demonstrate that it is possible to apply spline-type functionals to the problem of density estimation for large data sets. The resulting estimators may be considered to be kernel estimators for which we have motivated a choice of kernel. In particular, they specify the degree of smoothness we expect of the underlying density function, and they preserve the empirical moments of the data up to that degree.

Importantly, the functional definition of these estimates, and of the related histosplines, allows the use of the finite element method as an approximation to the fundamental solution we seek. This extends to many of the usual finite element variations, including nonconforming methods. Such an approximation has consequences for the computational cost of the algorithms. In particular, we can demonstrate that the algorithm is scalable in the size of the data set used. It also allows us much faster routines for computing an automatic choice of parameters. We present some numerical demonstrations of the performance of these estimators on synthetic and real-world data and some preliminary applications to which they may be put.

Acknowledgements

I would like to express my deep gratitude to my supervisors, Markus Hegland and Stephen Roberts. Their generosity with their time, patience and ideas has been nothing short of magnificent. Nira Dyn must also be thanked for her encouragement of the project earlier in the year. I would like to thank the School of Mathematical Sciences, my lecturers and particularly the honours co-ordinator, John Hutchinson. They have always been helpful and were their programme less gruelling this work would be substantially diminished. I am indebted to the Research School of Information Sciences and Engineering and to CSIRO for the use of their facilities and data-sets and to the ACSYS CRC whose financial support allowed those facilities to exist.

I must thank my fellow honours students, Emma, Nick and Sharon as well as Dirk and Steffen for their camaraderie in lunacy and the honours students from 1998 whose mistakes I have managed to repeat despite their warnings. My house-mates, Mel and Andi, have fought a valiant year-long battle for my sanity and deserve a special thanks. Both Mel and my father spent many hours proofreading and I am well and truly in their debt.

I could not possibly have gotten so far today without the support of my parents. Their generosity allowed me to come here and their love and affection has sustained me and given me the confidence to reach so high. My thanks also go to my sister, Claire, for demonstrating before me that such feats are possible. Finally, I am indebted to the many others who have influenced me and who are far too numerous to name; without them, I could not be here.

Contents

Preface

Suppose we were to take a set of measurements from a population of objects - width and height of people in Canberra, for example. If we plot each object as a point in space determined by its measurements, then it is relatively easy to see that the resulting picture would have some sort of pattern in it. It may be that the points appear to be evenly spread across the space, or they may seem to be clumped around one or more regions. The reproduction and description of these features is the central problem of density estimation. In physical density the amount of mass in a given region is measured. For the statistical case we are interested in the proportion of points (thought of as the mass) in any given region. In fact, we are really interested in the probability of finding a point in a region, or the proportion of points which ought to fall into the region, rather than the number of points which happen to do so.

These ideas are not new. The census performed by Moses in 1490 B.C. groups the number of militarily fit males by tribe. The motivation and techniques involved are somewhat more primitive than the ideas we wish to put forth, but the central concept of seeing how much data lies in some place remains. A more sophisticated methodology appears in 1625 in the calculations of John Graunt, who used church records to produce a table of the probabilities of dying in a particular ten-year age bracket [TT90b]. The results of Graunt's study were not presented graphically, but they nevertheless represent one of the earliest genuine estimates of density; in this case, the density associated with the record of age at death in early seventeenth century London. Nevertheless, the real development of density estimation, originally as a tool for data visualisation, begins with the 1891 lectures of Karl Pearson at Gresham College in London [Sco92]. Pearson is credited with coining the term "histogram" and was mostly concerned with the representation of data in various bins. It was not until 1952 that Fix and Hodges introduced the kernel method, followed by Rosenblatt in 1956. Since then a plethora of possible methods have been developed, taking advantage of the recent increase in computer power.

Despite recent advances, density estimation remains an inherently difficult field. Essentially, this stems from the fact that it is difficult to gain any intuition about the value which a density function ought to take. A graphical plot of data points might give us some feeling for the general shape which a density should have (provided we do not use too many points) but not even this is available for higher dimensional data. This is substantially different from the case of interpolation problems in which we know the values, with some margin of error, that a function ought to take at the data points. In those cases, a good estimate requires a certain uniformity of data across a region; in density estimation we are trying to establish whether such uniformity exists. This difficulty has manifested itself in the theory surrounding density estimation. Many of the most powerful results place bounds on how good an estimate can get, whatever the estimator we use. Essentially, we can always find a distribution of data points which will appear to be similar to some density function but which will differ substantially close to one point in space. Unlike the case of fitting a function whose values at the data point are known, there is no way of avoiding such pathologies in this situation.

This lack of a reference for the accuracy of a given function has meant that despite their multitudinousness, the methods which have been developed for this problem do not fit together particularly coherently. In fact, it can be shown that they almost all represent some form of kernel estimator, but this neither helps to decide between them or motivate a global choice of estimate. The work in this thesis presents an attempt to overcome these difficulties. We present a new estimator, which has been labelled the density spline, that attempts to produce a functional motivation for the estimator. It also allows us access to the finite element method, which provides a substantial computational saving. The development of the density spline and the exploration of its behaviour should be regarded as the central focus of this work.

Outline of Structure

The thesis is split into seven chapters dealing with aspects of density estimation and density splines. Chapter 1 presents an overview of the subject of density estimation. We develop the idea of a density and examine some of the important applications to which density estimation may be put. We then go on to give a brief outline of some of the more influential and important methods in the field. In §1.3 we produce some global theorems which are useful in setting a context for the following chapters. This includes theorems concerning the generality of the kernel method and the rate at which estimates are able to converge. Whereas the preliminary sections of Chapter 1 are intended to be relatively clear, §1.3 contains some very dense and often obscure mathematics. As much explanation as possible has been incorporated into the proofs there, but the reader will still need to pay close attention if they are to follow the proofs fully. Some constructions and theorems have also been left to the appendices.

Chapter 2 introduces density splines as the central focus of the thesis. It begins with an overview of the usual problem of spline smoothing from which we wish to motivate the methods that follow. Much of the theory relies on the structure provided by reproducing kernel Hilbert spaces which have been further described in Appendix D. The mathematics in this case is technical, but will be used later in §2.3. The Chapter then goes on to motivate and define the density splines we are interested in working with. We produce a univariate example of an exact solution and discuss some of the consequences and properties of the estimator. We will then go on demonstrate the convergence of the method in a particular case. Section2.3 is concerned with the generalised histospline, which may be thought of as a related estimator to the density spline. The section demonstrates some of the basic theory of the histospline and the form of the exact solution. It then goes on to produce the method of approximation proposed in [DW82] and to demonstrate that convergence is achieved in some cases in one dimension. We then make some comparison of the two methods. Section2.4 discusses the application of a finite element approximation to both estimators. In particular, it considers the computational savings that may be gained and the effect on some of the properties we have been interested in.

Chapter 3 is concerned with particular cases of the density spline. We begin by examining the Roberts estimator. This is the simplest form of a density spline. Nevertheless, it is possible to show that it has the nice property of being everywhere nonnegative. It has been named after Stephen Roberts who realised this. Section3.2 deals with a nonconforming finite element method for the thin plate density spline. Thin plate splines have already been discussed in Chapter 2 and this section aims to make the approximation simpler and more computationally efficient. We demonstrate the convergence of the finite element method to the true (spline) solution and examine the behaviour of the resulting estimates.

Chapter 4 deals with methods of choosing the parameters which we have so far regarded as fixed in the model. Clearly, in some circumstances we do not wish to assume what the value of these parameters should be and do not have the luxury of experimenting until we find a parameter which "feels" right. For the most part there is an explicit parameter controlling the smoothness of the estimator and we will present three possible methods of choosing a value for it. The behaviour of these will be examined in the context of density splines. We will also spend some time considering cases in which smoothing parameters are less obvious and less easy to choose.

In Chapters 5 and 6 the results of some numerical experiments are presented. We begin by examining the behaviour of these estimates in one and two dimensions. Many of these experiments have already been foreshadowed in Chapters 3 and 4. We then present the results of some one dimensional Monte Carlo studies concerning convergence rates. Chapter 5 ends with an examination of an application in the area of classification. We attempt to decide which of two data sets a new data point should belong to having first made estimates of the density of each set. This has been carried out on several sets of distinct data. Chapter 6 continues these experiments on real world data. The CSIRO holds a database of insurance policies for a motor vehicle insurance firm. We have used the data to examine the performance of density splines in one and two dimensions in regards to data visualisation and again in classification problems.

Chapter 7 concludes the analysis. We revisit the problems of demonstrating convergence and choosing smoothing parameters in so far as the development in previous chapters has been unsatisfactory. We also examine ways of avoiding the theoretical problems posed by the Sobolev embedding theorem which require our estimates to become unpleasantly complicated. We end with an outline of the application of these estimates in the problems of clustering and area definition and the ways in which they could produce computational savings.

Acknowledgement of Sources and Claims to Originality

A great deal of the theoretical work in this thesis is due to Grace Wahba. All of the convergence results for existing estimators can be found in her papers [Wah75b] and [Wah75a]. These have mostly been carried out in the univariate case and the results contained in Chapter 1 have been modified to extend these to higher dimensions. Theorem 1.3.2 is due to [Far72] and was originally given for the multivariate case, although [Wah75b] reproduces it only for univariate densities. In fact, the constructions in [Far72] have some serious mathematical flaws which have been corrected in Appendix C. Wahba is also partly responsible for the development of the generalised histosplines found in [DW82]. Much of the theory concerning standard splines and reproducing kernel Hilbert spaces is found in her book [Wah90].

Other results come from [Sco92] (particularly Theorem 1.3.1), [Sil86] and [DG85] (Theorem 1.3.5). The methods developed for choosing parameters in Chapter 4 are a compilation of results from these sources and from [DW82]. The essential mathematics required in §3.2 is found in [RHA99]. [BF91] has been used particularly in the development of the theory and approximation of generalised histosplines.

The material in the thesis is new in so far as it relates to density splines. The original conception for these estimators is due to Markus Hegland, who also contributed the greater part of the development in §2.2. As has already been mentioned, Stephen Roberts is responsible for Theorem 3.1.1 along with Nira Dyn, who collaborated in the proof and encouraged these ideas generally. Steve is also responsible for Theorem 2.2.1 jointly with myself. The remainder of the development of the density splines is original although not extraordinary. This includes transferring the nonconforming method found in [HRA97] to a density estimation context. For the most part, the theorems proved there may be essentially found in [RHA99] although the weak convergence rate provided in §3.2.3 is original as, I believe, are the ideas in §3.2.4. Naturally, the experimental assessment of the behaviour of the models, and the conjectures contained in Chapter 7 are my own.

A Note on Notation

The work in the thesis calls on many areas of mathematics: partial differential equations, statistics, functional analysis and numerical analysis. Necessarily, some compromises have been made on notation and a glossary of the conventions employed here may be found in Appendix A. In particular, $W_p^{(m)}$ is given an unusual definition and, unless necessary, the dimension of variables has been suppressed. The work in this thesis has been designed to be interesting to numerical analysts. Where possible, therefore, the over-use of statistical shorthand has been eschewed in order to make the intention of the notation clear. This shorthand has occasionally been necessary to clarify particularly messy formula and in particular we have followed the convention of suppressing the underlying measure space of random variables. In general, capitalised variables are intended to denote observations in a data set which may also be thought of as random variables from some measure space Ω into $\mathbb{R}^d$. One important exception to this is the proof in §1.3.2 in which the set $A_n = \{X_1, \dots X_n : |\gamma_n(X_1, \dots, X_n) - f(x)| \leq a_n\}$ is intended to denote a space in $\mathbb{R}^{nd}$ defined by n d-dimensional Euclidean variables X_i. This notation has been employed because the X_i's are intended to represent potential independent observations of a random variable $X : \Omega \to \mathbb{R}^d$ with density f. In this case we have the identity

$$P(\{\omega_1, \ldots, \omega_n : X_1(\omega_1), \ldots, X_n(\omega_n) \in A_n\}) = \int_{A_n} \prod_{i=1}^{n} f(x_i)dx_i$$

where P is regarded as a measure on Ω. This avoids the confusion of adding yet another set of vectored variables. Throughout the remainder of the thesis, we will suppress the underlying measure space and the use of Ω will be intended to denote some possibly bounded subset of Euclidean space.

Chapter 1

Introduction

Densities are members of that small class of mathematical objects which may be intuitively explained to a lay person. If we imagine a representation of a data set as points in some space, then we have a visual intuition that some regions have more data points in them than others. It is exactly in this sense that we attempt to define clusters of objects: they are groups of data points which are all close to each other in some way and are separated by a fairly sparsely populated region from other such clusters. This corresponds to our physical notions of density as well: a region which is more dense has a greater proportion of mass to volume than a less dense region. Producing some representation of this phenomenon is the essential task of density estimation. This chapter is intended to provide some framework for the methods we will explore later on. We will begin by providing a brief introduction to the theoretical development of densities and some of the standard methods used. We will then develop some rather dense mathematical theory concerning the approximation of density functions which will serve as a framework for the methods we will concern ourselves with in Chapter 2 and the rest of the thesis.

1.1 Defining the Problem

A natural response to seeing that the (data) points of a data set are not evenly distributed over some area is to attempt to create some quantitative representation for this phenomena. Now, our observations concern the amount of data in any particular region. The quantity we will be interested in is therefore the proportion of data which we can expect in that region. In other words, densities are measures on the domain of the data. For our purposes, we will consider only data given in $\mathbb{R}^d$. We will represent the data in this set as the sequence $\{X_i\}_{i=1}^n$ where each X_i represents some d-dimensional data point. The obvious first choice for such a measure is to say that $\mu(A) =$ proportion of data in $A \subset \mathbb{R}^d$. What we would like is to find some manageable representation of μ. We will develop this briefly. Let us define for each point $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ a set F_x where

$$F_x = \bigcap_{i=1}^{d} \{y : y_i \leq x_i\}.$$

This is the set of points y for which each co-ordinate value is smaller than the corresponding co-ordinate of x. We can evaluate $\mu(F_x)$ by determining, for each data point X_i, if the co-ordinates of x are greater than those of X_i. This is the same as evaluating

$$I_{\tilde{F}_{X_i}}(x)$$

where I_A represents the indicator function of A and

$$\tilde{F}_x = \bigcap_{i=1}^{d} \{y : y_i \geq x_i\}$$

is the set of all points y who's co-ordinates are greater than those of x. Taking $\mu(F_x)$ as a function of x gives us the empirical distribution function:

$$\mu(F_x) = \hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\tilde{F}_{X_i}}(x).$$

What this does is to add $1/n$ to the value of $\hat{F}(x)$ for each of the X_i which is less than x in all co-ordinates. Now $\hat{F}$ is very specific to the data set which we are using. We will usually assume that the data can take any value in $\mathbb{R}^d$ but that the distribution of points in any given region will reflect the probability of finding a new data point in that region. It would, after all, be of no use to examine $\{X\}_{i=1}^{n}$ if it contained the set of all possible data. The way we go about representing this idea is to modify our view of the data and to treat each X_i as realisation of a random variable which takes values in $\mathbb{R}^d$. We will then have a theoretical distribution function, $F(x)$:

$$F(x) = P(X \in F_x).$$

This representation remains very confined in the sets for which we are able to find the value of μ. It is useful, however, if we recognise that if F is differentiable we can derive that

$$\mu(A) = \int_A \frac{\partial^d F(x)}{\partial x_1 \dots \partial x_d} dx$$

and we can therefore use $\rho(x) = \frac{\partial^d F(x)}{\partial x_1 \dots \partial x_d}$ as a representation of μ. In fact, ρ is exactly the density which we are looking for and we will invert the derivation above to obtain the following definition.

Given a random variable, X, from some measurable space to $\mathbb{R}^d$, we define the density ρ of X to be the function which satisfies:

$$P(X \in \Omega) = \int_\Omega \rho(x) dx \tag{1.1}$$

and our task will be to reproduce ρ given some finite data set $\{X_i\}_{i=1}^{n}$.

We can represent the density of a discrete set of data by a sum of δ-functions, or point masses:

$$\rho_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i} \tag{1.2}$$

and this will correspond exactly to the empirical distribution function as a distribution in the partial differential equations sense. We can also give this correspondence a statistical meaning by introducing the likelihood functional

$$L(g|X_1,\dots,X_n) = \prod_{i=1}^{n} g(X_i).$$

This is used as a standard measure of goodness of fit in statistics [TT90b]. Essentially, we want g to take greater values where there is data, requiring it to have small values in sparse regions. Now if we let a sequence g_i converge to ρ_n we observe that the likelihood of g_i increases to infinity.

As before, this is not particularly useful, however, when we assume that the data does not represent exactly the situation we are trying to model. In order to create a theoretical density, we will therefore make the assumption that ρ is continuous and possesses some degree of smoothness. Typically, we will assume that ρ has some bounds on its derivatives and may be considered as an element of a reproducing kernel space.

Likelihood is useful, however, as a description of the estimators which we are interested in studying here. Nonparametric estimators may be defined as those which can be chosen to make their likelihood arbitrarily large. They are labelled δ-sequences [Sco92] in the sense that they can be chosen so as to approach ρ_n. Typically, these estimators will depend on some parameter or set of parameters which can be chosen independently of the data. In many cases, fixing the values of these parameters will lead to estimators which might be regarded as being parametric. We take distinction between parametric and nonparametric models to be the ability to vary the parameters of the nonparametric model, possibly best thought of as a family of models, so as to approach a δ-sequence.

1.1.1 Applications of Density Estimates

There are many uses to which an estimate of a probability density function may be put. The simplest and most common of these are data visualisation and exploration. Clearly, for large data sets, some more compact representation of the data is necessary in order to gain an intuitive feeling for its geometry. A representation of density allows us direct access to the spatial distribution of points in the data, enabling us to determine by sight how multi-modal the density is, and where most of the data will be located. It may well allow us to judge whether or not some parametric model will produce a good fit for the data set. There are, however, limitations to this. Functional visualisation is difficult in more than a few dimensions and good representations for higher dimensional data sets are difficult to achieve. Nevertheless, visualisation of the distribution of points over a few attributes can be very helpful.

Similar to the idea of determining the distribution of data are the two related problems of clustering and classification, or discriminant analysis. Clearly, a density function which has two or more distinct "peaks" can be thought of as being divided into two regions, or clusters, of data. It may be useful to have a sensible definition of clusters - a typology of customers in a store, or disease symptoms and so forth. The intuition of clusters as being regions with greater density provides one possible motivation for producing clustering algorithms.

Related to that is the problem of having some set of classes which a data point might belong to and wishing to decide between them. The use of density estimates comes through an application of Bayes theorem [Fri94]. Alternatively, it may be expressed in terms of the Neyman-Pearson lemma which says that given two populations with densities ρ_1 and ρ_2 respectively and suppose that X has equal probability of falling into either population, then the best way of classifying X is the ratio $\rho_1(X) : \rho_2(X)$.

There are cases in which densities are known to be discontinuous, or the underlying random variable may be known to take values in some discrete set. The first of these may be treated in some ad hoc manner; taking two densities either side of the discontinuity and so forth. The second may be satisfactorily treated as a histogram (provided the discrete set is known). Neither of these cases will be considered here.

1.1.2 Requirements

The definition (1.1) of the probability density function for X has several immediate consequences. To begin with, we will assume that X always takes values in some region Ω and therefore that

$$\int_{\mathbb{R}^n} \rho(x)dx = \int_{\Omega} \rho(x)dx = 1.$$

Further, it is clear that ρ will be everywhere nonnegative. We will also typically assume some smoothness properties of the density in question ($\rho \in W_p^{(m)}$ in §1.3.2 below). Finally, we can express the moment of $s(x)$ about X as

$$\int_{\Omega} s(x)\rho(x)dx$$

and in particular we may wish to maintain integer moments, corresponding to $s(x) = x^m$; particularly mean and variance, given for $m = 1, 2$.

Exactly which properties it is desirable to preserve will depend on the application which is to be made of the density estimate. Typically, the unit integrand will be preserved, as will smoothness and we will follow this convention. It proves very difficult to preserve non-negativity, however, and it seems to be necessary to forgo this property in order to achieve a good rate of convergence for the estimate. Little attention has been paid to moments but we will demonstrate that the estimates described in chapter 2 will preserve the empirical integer moments of the dataset to order one less than the degree of smoothness required.

1.2 Standard Estimates

Suppose we know that the data set we are given has an underlying distribution of some type - normal, Poisson etc. Then we can produce a density estimate simply by taking the usual estimates of mean and variance and fitting the underlying distribution to them. This is a standard parametric approach and provided that the assumptions about the form of the density are accurate can be shown to provide a very good fit. They are not so useful, however, when the distribution of the underlying random variables is unknown and may be multi-polar or otherwise pathological. In this case we need to develop nonparametric methods to produce some form of a density estimate. This is analogous to the case of spline fitting when the underlying functions are unknown.

Nonparametric density estimation, however, differs from regression problems in lacking a response variable as an indicator of the values which the estimating function should take. This, in some sense, makes the problem considerably more difficult. It also precludes the direct use of spline methods in the problem. The result of this is that many of the density estimation methods which have been developed seem philosophically somewhat *ad hoc*; something born out by the plethora of differing methods and the lack of a universal treatment of them. We will here outline a few of the more common methods and their properties in order to gain a historical perspective on solutions to this problem. These may be found, studied in greater detail, in [Sil86] and [TT90a]. They will be important to the rest of the thesis.

1.2.1 The Histogram

The histogram is the simplest and oldest of density estimators. The first known example of a histogram-type representation of data (although not visually presented) occurs in 1661 as a representation of the probability of death grouped into ages [TT90a]. The name itself was not coined until 1891 by Karl Pearson [Sco92]. The naïveté of this estimate provides a nice connection between our intuitive concepts of physical density and probability density functions themselves.

Physical density is given in terms of mass per unit volume. If, then, we assign mass $1/n$ to each of the X_i, a first approximation to the density of the X_i in some region Ω_j would be to count the number of X_i in Ω_j and divide by n and the volume of Ω_j. If we do this for some set $\{\Omega_j\}_{j=1}^n$ and represent it as a function, we have exactly the histogram estimate:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{m} \frac{(\text{no. of } X_i \text{ in } \Omega_j)}{|\Omega_j|} I_{\Omega_j}$$

A standard procedure is to choose all the Ω_j the same size, say h, and to then take the simplification

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{j} (\text{no. of } X_i \text{ in } \Omega_j) I_{\Omega_j}$$

This is simple, readily understandable and remains one of the most common tools of data representation. In fact, data is often aggregated into histograms and a large number of estimators are essentially methods for smoothing the histogram discontinuities. The estimate is a density: both

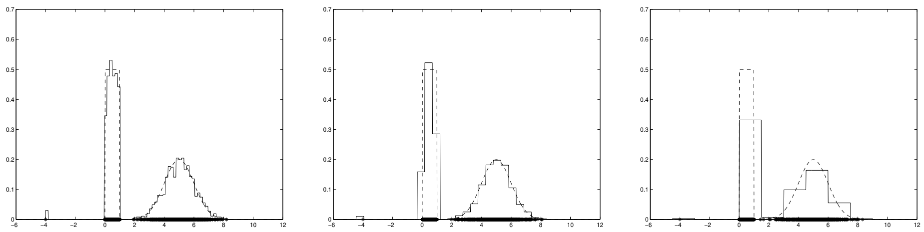


Figure 1.1: Histograms of various bin widths approximating a uniform and a normal distribution. The true (theoretical) density is given by dashed lines.

positive and with integral 1. As we let $h \to 0$, it is also possible to see that this is a δ-sequence estimator. It does, however, violate our smoothness conditions, being not even continuous. This causes problems when it is important that we know the derivatives of the density estimates and can be problematic when the density is being used as an intermediate step; in trying to develop a hill-climbing algorithm for clustering, for example, or in producing a contour map in higher dimensional data. We further see that apart from being dependent on the choice of the volume of Ω_j it is also potentially significantly affected by the placement of the partition. If we take a uniformly distributed set of points on the interval $[0, 1]$ and choose a bin width of 1 and a starting partition at 0 we will match the underlying density estimate perfectly. If we choose the same width and start the partition at $-1/2$, however, we have that $\int |\hat{f}_n(x) - \rho(x)|dx = 1$! These results will, of course, become less problematic as the size of the bins becomes smaller; nevertheless in trying to smooth the data (achieved by choosing larger bins, thereby reducing the number of discontinuities) this phenomenon will introduce unnecessary errors.

1.2.2 Kernel Estimators

It is possible to fairly simply fix the last problem if we alter our perception of what we are doing. When we are creating a histogram we can think of each X_i as adding another 'block' of height $1/n$ to the column in which it appears. The logical transformation, then, is to instead centre the block on X_i and come up with a function of the following form:

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} I_{C_h(X_i)}(x)$$

where C_h is a hypercube centred on X_i with total volume h.

This is typically known as the naive estimator - see Figure 1.2 as an example. The natural generalisation of this, instead of using an indicator function, is to place some smoothed 'bump' at each data point. This will give us the following:

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

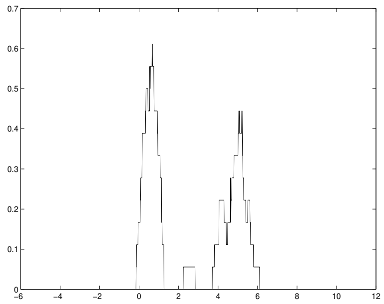


Figure 1.2: A naïve estimator for a distribution as in Figure 1.1

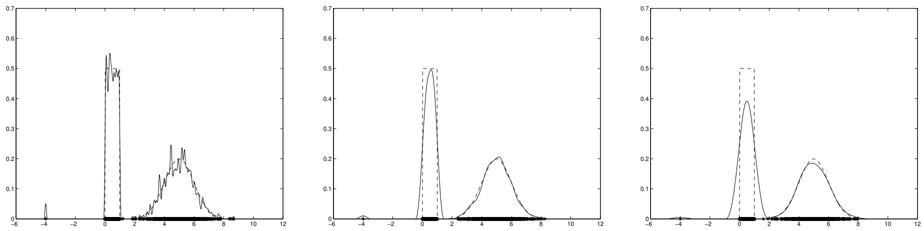


Figure 1.3: A smooth kernel estimate with various choices of h. The theoretical density is given by dashed lines.

Here the kernel, K, is some function around zero, integrating to 1. Typically we will assume that K is symmetric around zero, has some degree of smoothness and is positive, although none of these criteria need be met. [TT90b] lists common univariate kernels as being:

$$K_1(y) = \frac{1}{2},\ |y| \leq 1$$

which gives us the naive estimator;

$$K_2(y) = 1 - |y|,\ |y| \leq 1$$

for continuity;

$$K_3(y) = \frac{15}{16}(1 - y^2)^2,\ |y| \leq 1$$

for smoothness; and

$$K_4(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

the Gaussian kernel.

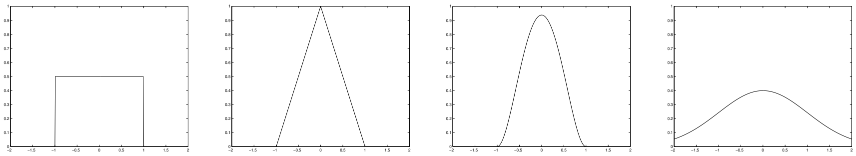


Figure 1.4: Kernels, from left to right, K_1, K_2, K_3 and K_4.

K_4 is intuitively nice in mimicking the ubiquitous normal distribution but the fact that it has infinite support is a drawback computationally. Multivariate kernels may be produced separately, or from tensor products of univariate kernels.

The use of the parameter h allows us to specify the smoothness of the resulting density; letting $h \to 0$ the kernel will converge to a sum of delta functions; increasing h will reduce the size of the derivatives of K and therefore of the resulting density as well. A similar role is played by the size of the bins used in the histogram estimator.

Clearly, this method has some nice properties. The resulting density integrates to 1, may be as smooth as we require it and is nonnegative (assuming we use the correct function for K). This is probably the most common type of estimator used to create densities. Nevertheless, its motivation appears somewhat *ad hoc* and the theory as presented does not allow for an easy control of moments, or for a simple study of convergence properties. In requiring a sum of n functions it is computationally inefficient, although it is possible to reduce this cost by evaluating the estimator on a uniform grid and using fast Fourier transforms.

1.2.3 Orthogonal Series Estimators

It is possible to apply some more sophisticated mathematical techniques to this problem with the hope that we will then be able to make use of some more powerful results in producing an analysis of the density estimate.

Let us suppose that $\{\phi_\nu\}_{\nu=0}^{\infty}$ is an orthogonal series of functions which forms a basis for a space of functions for which ρ is a member, then any function f may be written as the series expansion:

$$f = \sum_{\nu=0}^{\infty} f_\nu \phi_\nu. \tag{1.3}$$

Here the coefficients f_ν are given by:

$$f_\nu = \int \phi_\nu(x) f(x) dx.$$

This is easily seen by considering $\mathrm{Span}\{\phi_1, \phi_2, \dots\}$ to be a Hilbert space with inner product $< f, g >= \int f(x)g(x)dx$. Then (1.3) is just the representation of f in terms of the basis $\{\phi_i\}_{i=1}^{\infty}$. In the case where f is a density, f_ν is just the expectation of ϕ with respect to f and can therefore be approximated by:

$$\hat{f}_\nu = \frac{1}{n}\sum_{i=1}^{n}\phi_\nu(X_i). \tag{1.4}$$

Regarding X_i as a random variable, we have that $E(\phi_\nu(X_i)) = E_f(\phi_\nu)$. Therefore $\hat{f}_\nu$ is an unbiased estimate of f_ν in the sense that $E(\hat{f}_\nu) = f_\nu$.

A standard series to take here is

$$\begin{aligned}\phi_0(x) &= 1\\ \phi_{2r-1}(x) &= \sqrt{2}\cos(2\pi r x)\\ \phi_{2r}(x) &= \sqrt{2}\sin(2\pi r x)\end{aligned}$$

for a density estimated on $[0,1]$. This will produce a Fourier series, with $\hat{f}_\nu$ the sample Fourier co-efficients. More general series will, of course, be equally applicable. One with the nice property of producing a nonnegative estimate is produced in [BB79] based on a Haar system on $[0,1]$:

For any $m \in \mathbb{N}$ we can find a unique $k, j \in \mathbb{N}$, $0 \le j \le 2^k$ such that $m = 2^k + j$. Then set $\phi_0(x) = 1$ and

$$\phi_m(x) = \begin{cases} 2^{k/2} & x \in \left(\frac{j-1}{2^k}, \frac{j-1/2}{2^k}\right)\\ -2^{k/2} & x \in \left(\frac{j-1/2}{2^k}, \frac{j}{2^k}\right)\\ 0 & \text{otherwise.}\end{cases} \tag{1.5}$$

We will see below that this returns us to the histogram.

The specification $\phi_0 = 1$ is typical. This will ensure $\hat{f}_0 = 1$ and, by the orthogonality of the series, that ϕ_ν integrates to zero for $\nu \ge 1$. In particular this assures us that

$$\hat{f}_k(x) = \sum_{\nu=1}^{k}\hat{f}_\nu\phi_\nu(x)$$

will integrate to 1. We have here truncated the series. Observing that as $k \to \infty$ the series tends towards a sum of delta functions (exactly the function we used to calculate our approximate co-efficients in (1.4)) we can regard the k which we choose as a smoothing parameter. Under this assumption, and further assuming that the orthogonal series we choose behaves nicely, we can easily see that

$$E\hat{f}_n(t) = \sum_{\nu=1}^{k} f_\nu\phi_\nu(t) = f(t) + o(1)$$

as $k, n \to \infty$ due to the convergence of the series. It is worth noting at this point that we can turn this into a kernel estimate by considering the function:

$$K_m(t,s) = \sum_{\nu=0}^{m}\phi_\nu(t)\phi_\nu(s)$$

then our estimate becomes:

$$\hat{f}_n(t) = \frac{1}{n}\sum_{i=1}^{n} K_m(t, X_i).$$

We note here that we are giving K_m as a function of two variables, rather than a displacement of a function centred about the origin. In many cases we can reduce this to a standard kernel. We note that in fact this structure is mathematically very convenient. If $g \in \text{Span}\{\phi_1, \dots, \phi_m\}$ then we can examine the moment of g about $\hat{f}_n$:

$$\begin{aligned}
\int_\Omega g(x)\hat{f}_n(x)dx &= \int_\Omega \left(\sum_{\nu=1}^{m} a_\nu \phi_\nu(x)\right) \hat{f}_n(x)dx \\
&= \sum_{\nu=1}^{m} a_\nu \frac{1}{n}\sum_{i=1}^{n} \phi_\nu(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} g(X_i)
\end{aligned}$$

expanding g and using the orthogonality properties of $\{\phi_\nu\}_{\nu=1}^m$. We have therefore managed to preserve the moment of g with respect to the data (alternatively, the generalised sample Fourier co-efficients). Typically we would like to preserve the mean and variance of g, and it will be difficult to achieve this in a truncated Fourier series in general.

If we consider the Haar orthogonal series defined in (1.5) we can get the following expression for the kernel. If $y \in ((j-1)/2^k, j/2^k)$, then

$$K_{2^k} = \sum_{i=1}^{2^k} \phi_i(x)\phi_i(y) = \begin{cases} 2^k & x \in \left(\frac{j-1}{2^k}, \frac{j}{2^k}\right) \\ 0 & \text{otherwise.} \end{cases}$$

This is clearly always positive, and we have produced a histogram of bin size $1/2^k$.

Alternatively, instead of truncating the series at some k, we can taper it: setting $\lambda_\nu \to 0$ we can use

$$\hat{f}_\lambda(x) = \sum_{\nu=1}^{\infty} \lambda_\nu \hat{f}_\nu \phi_\nu(x)$$

or some truncated version of it as an estimator; here the λ_ν represent a series of smoothing parameters.

We have thus constructed a somewhat more sophisticated mathematical context to consider the problem. It allowed us the control of some moments, and it has produced a nice demonstration that we do achieve a form of convergence to the actual density.

1.3 Some Preliminary Results

This section will incorporate a few theorems which will have a bearing on the rest of the thesis. Some will be useful later in the work, others will help set a context in which to analyse and evaluate the methods we develop.

1.3.1 General Kernel Theorem

This result states that a large class of estimators may be thought of as kernel estimators. It was first produced in [TS92] and we reproduce their proof here. This will be done by considering the relationship between the estimator $\hat{f}$ and the empirical distribution function:

$$\hat{F_n} = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j=1}^{d} I_{[x_i^j,\infty)}\right).$$

We are assuming a data set given on $\mathbb{R}^d$ with the j^{th} component of X_i given by x_i^j. Given such a distribution, the original (unordered) data can easily be recovered from the function. A density estimator will typically accept a set of data and return an estimating function; in this case it will be mathematically convenient to suppose that it acts on the empirical distribution function. We can therefore express the estimate at the point x by $T_x\{\hat{F_n}\}$ where T_x is regarded as an operator on $\hat{F_n}$. Under the assumption that this is the case and that T_x is Gâteaux differentiable, we have the following theorem:

Theorem 1.3.1. *Any density estimator that is a continuous and Gâteaux differentiable functional on the empirical distribution function may be written as:*

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} K(y, X_i, \hat{F_n})$$

where K is related to the Gâteaux derivative of $\hat{f}$ under variation of X_i:

$$K(y, X, \hat{F_n}) \equiv \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left[T_y\left\{(1-\epsilon)\hat{F_n} + \epsilon\prod_{j=1}^{d} I_{[x^j,\infty)}\right\} - (1-\epsilon)T_y\{\hat{F_n}\}\right]$$

Proof. We will begin by writing the density estimator as above: $\hat{f}(x) = T_x\{\hat{F_n}\}$. We have defined the kernel as:

$$\begin{aligned}
K(y, X, \hat{F_n}) &\equiv \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left[T_y\left\{(1-\epsilon)\hat{F_n} + \epsilon\prod_{j=1}^{d} I_{[x^j,\infty)}\right\} - (1-\epsilon)T_y\{\hat{F_n}\}\right] \\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left[T_y\left\{\hat{F_n} + \epsilon\left(\prod_{j=1}^{d} I_{[x^j,\infty)} - \hat{F_n}\right)\right\} - T_y\{\hat{F_n}\}\right] + T_y\{\hat{F_n}\} \\
&= DT_y(\hat{F_n})\left[\prod_{j=1}^{d} I_{[x^j,\infty)} - \hat{F_n}\right] + \hat{f}(y)
\end{aligned}$$

Here $DT(z)[w]$ is the Gâteaux derivative of T at z in the direction w. This is linear in the second term (see Appendix B) and we therefore have that:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} K(y, X_i, \hat{F}_n) &= DT_y(\hat{F}_n)\left[\frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j=1}^{d} I_{[x_i^j,\infty)}\right) - \hat{F}_n\right] + \hat{f}(y) \\ &= 0 + \hat{f}(y)\end{aligned}$$

□

Note that we are not enforcing $K(x, y) = K(x - y)$ and no longer have an explicit smoothing parameter. This is a necessary price to pay for the generality of the theorem, which encompasses parametric estimators along with most nonparametric methods. We have already demonstrated that the methods described above can be regarded as kernel methods in this sense. What this result shows is that any reasonable method (in the sense of being stable under small perturbations of the data) can also be regarded as such a method. It is worthwhile noting that the K which is derived here need satisfy none of the conditions necessary to produce a density. Of course, any reasonable functional of the empirical distribution function will force it to do so.

It does, however, provide us with a motivation for an analysis of density estimators - they are methods of choosing kernels. The orthogonal series method is one such motivated by a Hilbert Space structure. In Chapter 2 we will produce a method which makes explicit the trade-off in accuracy and smoothness inherent in smoothing problems. It will also allow us greater flexibility in controlling the moments associated with our density estimate.

1.3.2 Optimal Convergence Rates

Naturally, one of the things we are interested in when developing an approximating function is the rate at which it can be said to converge as we add more data. This attention is somewhat misdirected; infinite data sets are difficult to find in practise, and even harder to work with. Nevertheless, we can readily imagine the possibility of forever adding more data to a collection and this will motivate the expectation that the answers we arrive at for a density estimate will be somewhere near the true density. We will present a result in [Far72] placing a bound on the achievable rates of convergence. It is worthwhile noting that this result is global in the sense that it is true for all estimators. Importantly, it means that, provided we can achieve that rate, comparisons between methods need only be done on the basis of mathematical simplicity and computational efficiency.

We will begin by defining the problem that we want to consider a little more precisely. We let $W_p^{(m)}(\Omega)$ be the Sobolev space of functions whose partial derivatives up to order $m-1$ are absolutely continuous and whose m^{th} partial derivatives are in $\mathcal{L}^p(\Omega)$. We will assume that the density we are approximating is in $B_{W_p^m}(M)$, the subset of $W_p^{(m)}(\Omega)$ for which the estimate $|f|_{W_P^{(m)}(\Omega)} \leq M$ holds. We are suppressing Ω in this notation. This imposes some degree of smoothness on the problem. Our estimate of a pointwise "goodness of fit" will be done using an expectation defined on a probability measure specific to the function. Specifically, P_f is intended to denote the product measure:

$$P_f(A_n) = \int_{A_n} \prod_{i=1}^{n} f(x_i) dx_i$$

where each $x_i \in \mathbb{R}^d$ and $A_n \subset \mathbb{R}^{nd}$[1] and we will find a bound on the rate at which sequences $\{a_n\}_{n=1}^{\infty}$ such that:

$$\liminf_{n\to\infty} \inf_{f\in B_{W_p^m}(M)} P_f(\{X_1, \ldots X_n : |\gamma_n(X_1, \ldots, X_n) - f(x)| \leq a_n\}) = 1 \tag{1.6}$$

can converge to zero. Here γ_n is an estimate of $f(x)$ based on n given observations, $X_1, \ldots, X_n$. In general, we will use the standard shorthand $|\gamma_n - f(x)| \leq a_n$ for the set $\{X_1, \ldots, X_n : |\gamma_n(X_1, \ldots, X_n) - f(x)| \leq a_n\} \subset \mathbb{R}^{nd}$ and assume that it is measurable. This is essentially a form of convergence in measure. Obviously, we would like to be able to choose a_n to converge to zero and at the same time we would like the probability that our estimator is within a_n of $f(x)$ to increase to one. This would mean that in the limit we can be almost certain that $\gamma_n = f(x)$. The rate at which it is possible to choose such a_n to converge to zero might be called the rate of convergence in measure. We assume, of course, that the X_i represent potential observations of independent random variables with density f, so that P_f is a reasonable measure to take on the error.

The proof of the theorem can be performed using the relatively simple process of modifying an appropriate density so that it is significantly different at zero but whose overall distribution is changed relatively little. The theorem produced is due to [Far72]. The proof here follows that given in [Wah75b] only modified to simplify some constructions and to extend the result to the multivariate case.

Theorem 1.3.2. *Suppose that x is fixed and $\{a_n, n \geq 1\}$ is a sequence of nonnegative real numbers such that*

$$\liminf_{n\to\infty} \inf_{f\in B_{W_p^m}(M)} P_f(|\gamma_n - f(x)| \leq a_n) = 1$$

Then,

$$\lim_{n\to\infty} n^{\phi_d(m,p)} a_n^2 = \infty$$

with

[1]The use of this notation suppresses n, it derives from the correspondence

$$P(X \in A) = \int_A f(x)dx$$

where X is a random variable with density X. In fact in this section we have used X_i to denote a variable in d-dimensional Euclidean space. The capitalised variables have been retained because they are intended to be thought of as potential observations of random variables with density f.

$$\phi_d(m,p) = \frac{(2m - 2d/p)}{(2m + d - 2d/p)}$$

where γ_n is an estimate of $f(x)$ and d represents the dimension of the underlying data.

What this theorem states is that if we attempt to estimate f at x given only n data points and the information that $f \in B_{W_p^m}(M)$ then as we increase the amount of data we use, we cannot decrease a_n at a a rate faster than $n^{-\phi_d(m,p)/2}$ and still guarantee convergence in measure.

Proof. Without loss of generality we can suppose we are estimating the function at $x = 0$. Firstly, we will create densities $f, h_\delta \in B_{W_p^m}(M)$ which will be close as measures but dissimilar at 0. This will be done using a modifying function $e_{m\delta} \in W_p^{(m)}(\Omega)$ constructed in Appendix C. In particular, $e_{m\delta}$ possesses the following properties:

$$e_{m\delta}(x) = 0,\ ||x||_\infty \geq \delta \tag{1.7}$$

$$\int_\Omega e_{m\delta}(x)dx = 0 \tag{1.8}$$

$$\max_x |e_{m\delta}(x)| = e_{m\delta}(0) = \delta^{m-2d/s} \tag{1.9}$$

$$\int_\Omega (e_{m\delta}(x))^2 dx \leq c_1 \delta^{2m+d-2d/s} \tag{1.10}$$

where $s > p$ and $c_1 = 2^d$. We also have the estimate:

$$|e_{m\delta}|_{W_p^{(m)}(\Omega)} \leq K\delta^{d(1-p/s)}$$

for some K. We will now choose δ such that

$$c_1\delta^{m+d-d/s} < 1 \tag{1.11}$$

and small enough that $e_{m\delta} \in B_{W_p^m}(M/2)$. This requires $\delta \leq [M^p/(2^pK)]^{1/(d-pd/s)}$. We also take:

$$a = \delta^{m-2d/s} + \varepsilon. \tag{1.12}$$

For ε small enough to ensure unit integrand, we can take f to be a density in $B_{W_p^m}(M/2)$ such that $f(x) = a > 0$ for $||x||_\infty \leq \delta$. We define

$$h_\delta(x) = f(x) + e_{m\delta}(x)$$

and this is a density in $B_{W_p^m}(M)$ from (1.8), (1.12) and the fact:

$$|h_\delta|^p_{W_p^{(m)}(\Omega)} = |f|^p_{W_p^{(m)}(\Omega)} + |e_{m\delta}|^p_{W_p^{(m)}(\Omega)}.$$

because f is flat on the support of $e_{m\delta}$.

Now we will turn to demonstrating a bound on the achievable rate of convergence. This involves the difficulty of distinguishing between a set of points with density f and one with density h_δ. Let us denote by I_n the indicator function of the set $\{X_1, \dots, X_n : |\gamma_n(X_1, X_2, \dots, X_n) - h_\delta(0)| \leq a_n\}$. Then

$$\begin{aligned} P_{h_\delta}(I_n) &= \int \cdots \int I_n(x_1, \dots, x_n) \prod_{i=1}^{n} \left(\frac{h_\delta(x_i)}{f(x_i)^{1/2}} \right) \prod_{i=1}^{n} f(x_i)^{1/2} \prod_{i=1}^{n} dx_i \\ &\leq (P_f(I_n))^{1/2} \left(\int \cdots \int \prod_{i=1}^{n} \frac{h_\delta(x_i)^2}{f(x_i)} \prod_{i=1}^{n} dx_i \right)^{1/2} \\ &\leq (P_f(I_n))^{1/2} \left(\int \frac{h_\delta(x)^2}{f(x)} dx \right)^{n/2}. \end{aligned}$$

We can now re-express h_δ as:

$$h_\delta = \begin{cases} f(x) & \text{if } ||x||_\infty > \delta \\ a + e_{m\delta}(x) & \text{if } ||x||_\infty \leq \delta. \end{cases}$$

Using this expression and (1.8) we can evaluate

$$\begin{aligned} \int \frac{h_\delta(x)^2}{f(x)} dx &= \int_{||x||_\infty > \delta} f(x)dx + \int_{||x||_\infty < \delta} \left(a + 2e_{m\delta}(x) + \frac{e_{m\delta}(x)^2}{a} \right) dx \\ &= 1 + a^{-1} \int (e_{m\delta}(x))^2 \, dx \\ &\leq 1 + c_2 \delta^{2m+d-2d/s} \end{aligned}$$

where $c_2 = c_1/a$. We now make use of this expression to gain the inequality

$$P_{h_\delta}(|\gamma_n - h_\delta(0)| \leq a_n) \leq (P_f(|\gamma_n - h_\delta(0)| \leq a_n))^{1/2} (1 + c_2 \delta^{2m+d-2d/s})^{n/2}. \tag{1.13}$$

In order to bound the left-hand side of (1.13) we make use of the assumption that γ_n provides convergence uniformly across $B_{W_p^m}(M)$ as $a_n \to 0$ and that we therefore have:

$$\inf_{g \in B_{W_p^m}(M)} P_g(|\gamma_n - g(0)| \leq a_n) = 1 - \theta_n$$

where $\theta_n \to 0$. Applying this to (1.13) gives us that:

$$P_f(|\gamma_n - h_\delta(0)| \leq a_n) \geq \frac{(1 - \theta_n)^2}{(1 + c_2 \delta^{2m+d-2d/s})^n}.$$

Using (1.9) we have:

$$|h_\delta(0) - f(0)| = \max_x |e_{m\delta}(x)| = \delta^{m-2d/s}.$$

Now if the sets $\{|\gamma_n - h_\delta(0)| \leq a_n\}$ and $\{|\gamma_n - f(0)| \leq a_n\}$ intersect then we know that for some $X_1, \ldots, X_n$ in their intersection, $2a_n \geq |\gamma_n - f(0)| + |h_\delta(0) - \gamma_n| \geq \delta^{m-2d/s}$. This will be true if the sum of their probabilities is greater than one. This statement can be expressed mathematically as:

$$\frac{(1-\theta_n)^2}{(1 + c_2\delta^{2m+d-2d/s})^n} + (1+\theta_n) \geq 1 \Rightarrow 2a_n \geq \delta^{m-2d/s}.$$

Let $\delta = (d_n/n)^{1/(2m+d-2d/s)}$ be allowed to vary. Some manipulation gives us:

$$\frac{(1-\theta_n)^2}{\theta_n} \geq \left(1 + c_2\frac{d_n}{n}\right)^n \Rightarrow n^{\phi_d(m,s)}a_n^2 \geq \frac{1}{4}d_n^{\phi_d(m,s)}. \tag{1.14}$$

The left-hand side of (1.14) is true whenever

$$d_n \leq \frac{n}{c_2}\left[\left(\frac{(1-\theta_n)^2}{\theta_n}\right)^{1/n} - 1\right] \to \infty, \text{ as } n \to \infty$$

This divergence is clear when we consider that for

$$\left[\left(\frac{(1-\theta_n)^2}{\theta_n}\right)^{1/n} - 1\right] \preceq \frac{1}{n}$$

we require

$$\left(\frac{(1-\theta_n)^2}{\theta_n}\right) \leq \left(1 - \frac{1}{n}\right)^n, \ n \to \infty.$$

But $(1+1/n)^n$ is bounded above which $\left(\frac{(1-\theta_n)^2}{\theta_n}\right)$ is not for $\theta_n \to 0$. We can now always choose d_n such that $d_n \to \infty$. Some calculation show that (1.12) will be satisfied for sufficiently large n as $\delta \to 0$ for $n \to \infty$ and we have that

$$\lim_{n\to\infty} n^{\phi_d(m,s)}a_n^2 = \infty.$$

Letting $s = p+\epsilon$ throughout we have produced a density for which we have explicitly that pointwise convergence is no better than $O(n^{\phi_d(m,p+\epsilon)})$ and we can conclude that this places a bound on the rate of convergence that can take place uniformly across $B_{W_p^m}(M)$. We note that the construction used can be made to hold for $p = \infty = s$ (see Appendix C) and the result is retained in that case. $\square$

The mathematics here is very subtle. We are forced to bound the size of $|f|_{W_p^{(m)}(\Omega)}$ in order to prevent $e_{m\delta}$ from being made too pathological. In fact, without such a bound we are unable to guarantee convergence in this sense at all. Nevertheless, the actual convergence rate does not depend on M and it will be suppressed later. When considering convergence to any particular function f, the size of $|f|_{W_p^{(m)}(\Omega)}$ is a sufficient bound in itself. The exact interpretation of this result is also not entirely clear as it stands. [Wah75b] puts this result in an intuitively nice setting with the following simple corollary:

Corollary 1.3.1. *For a fixed ϵ suppose that:*

$$\sup_{f \in B_{W_p^m}(M)} E_f(\gamma_n - f(0))^2 = b_n n^{-\phi_d(m,p+\epsilon)}$$

Then $\limsup b_n > D_0$ *for some* $D_0 > 0$.

Proof. This is a simple application of Markov's inequality. By the hypothesis:

$$P_f(|\gamma_n - f(0)| \leq a_n) \geq 1 - \frac{E_f(|\gamma_n - f(0)|)^2}{a_n^2} \geq 1 - \frac{b_n n^{-\phi_d(m,p+\epsilon)}}{a_n^2}.$$

Now if we have that $\limsup b_n = 0$ we can choose $a_n = n^{-\phi_d(m,p+\epsilon)/2}$ and we have that

$$\liminf_{n\to\infty} \inf_{f \in B_{W_{pm}}(M)} P_f(|\gamma_n - f(0)|) \leq a_n) = 1.$$

On the other hand $\liminf_{n\to\infty} a_n^2 n^{\phi_d(m,p+\epsilon)} < \infty$ contradicting Theorem 1.3.2 proven above. $\square$

This states that if we have some estimator γ_n of $f(x) \in W_p^{(m)}(\Omega)$ then we can guarantee a convergence rate of no better than $O(n^{-\phi_d(m,p+\epsilon)})$. In other words, there is a bound on how well we can expect our estimators to behave.

Pointwise errors do not, of course, contain information about the global convergence properties that we would like to have. If we assume that the constants associated with the convergence rates at each point actually represent a continuous function, then we can produce a similar bound on the mean integrated squared error (MISE):

$$E \int_\Omega (f(x) - \hat{f}_n(x))^2 dx.$$

This can be seen by simply observing that $(f(x) - \hat{f}_n(x))^2 > 0$ and changing the order of integration. For $\Omega' \subseteq \Omega$ on which the constants are bounded away from zero:

$$\begin{aligned}\int_\Omega E_f(f(x) - \hat{f}_n(x))^2 dx &\succeq \int_{\Omega'} kn^{-\phi_d(m,p+\epsilon)} dx \\ &= |\Omega'| kn^{-\phi_d(m,p+\epsilon)},\end{aligned}$$

which provides a bound under fairly weak assumptions.

1.3.3 Optimal Convergence of the Parzen Kernel Estimate

Having established a bound on the possible rates of convergence over a class of densities we will now proceed to demonstrate that this bound is sharp by showing that a particular type of kernel, known as the Parzen kernel, achieves the best possible rate of convergence. The proof here is developed from [Wah75b] and [Par62], modified in all cases to extend results from the univariate to the multivariate case. We are forced to make the assumption here that $d < p$ and some of the conditions below have also changed. The proof in [Wah75b] also provides the result for $p = d = 1$.

Let us suppose that $f \in W_p^{(m)}$ and $K(y)$ is a real valued function on $\mathbb{R}^d$ that satisfies:

$$\sup_{y\in\mathbb{R}^d} |K(y)| < \infty \tag{1.15}$$

$$\int_{\mathbb{R}^d} |K(y)|dy < \infty \tag{1.16}$$

$$\int_{\mathbb{R}^d} K(y)dy = 1 \tag{1.17}$$

$$\int_{\mathbb{R}^d} y^i K(y)dy = 0, \ |i| = 1, \dots, m-1 \tag{1.18}$$

$$\int_{\mathbb{R}^d} |y^\alpha K(y)|^{1-1/p} dy < \infty, \ |\alpha| = m \tag{1.19}$$

where $y^{|i|} = \prod_{k=1}^d y_k^{i_k}$ and $|i| = \sum_{k=1}^d i_k$. Such a function will be labelled a Parzen kernel and we will produce an estimate

$$\hat{f}_n(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \tag{1.20}$$

with $h > 0$ chosen so that $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$. We can then prove the following theorem.

Theorem 1.3.3. *Let $p > d$, $f \in W_p^{(m)}$ and $\hat{f}_n$ be given by (1.20) with $h = (k_n/n)^{1/d}$ and*

$$k_n = Cn^{(2m-2d/p)/(2m+d-2d/p)}.$$

Then

$$E(f(x) - \hat{f}_n(x))^2 \leq Dn^{-(2m-2d/p)/(2m+d-2d/p)}(1 + o(1)).$$

The proof will depend on finding expressions for the bias and variance terms of

$$E(\hat{f}_n(x) - f(x))^2 = \left[E\hat{f}_n(x) - f(x)\right]^2 + E\hat{f}_n(x)^2$$

and we will examine each in turn in the following lemmata. Let us denote

$$f_n(x) = E\hat{f}_n(x) = \frac{1}{h^d} \int_{\mathbb{R}^d} K\left(\frac{x-\xi}{h}\right) f(\xi)d\xi. \tag{1.21}$$

Then, as a preliminary result, the following appears in [Par62]:

Lemma 1.3.1. *At every continuity point of $f(\cdot)$,*

$$\lim_{n\to\infty} f_n(x) = f(x)\int_{\mathbb{R}^d} K(\xi)d\xi.$$

Proof. We will first note that

$$f_n(x) - f(x)\int_{\mathbb{R}^d} K(\xi)d\xi = \int_{\mathbb{R}^d} [f(x-\xi) - f(x)]\frac{1}{h^d}K\left(\frac{\xi}{h}\right)d\xi.$$

We then split the region of integration in two parts: $|\xi| \le \delta$ and $|\xi| > \delta$ for some $\delta > 0$:

$$\begin{aligned}
\left|f_n(x) - f(x)\int_{\mathbb{R}^d} K(\xi)d\xi\right| &\le \sup_{|\xi|\le\delta|} |f(x-\xi) - f(x)| \int_{|\zeta|\le\delta/h^d} |K(\zeta)|d\zeta \\
&\quad + \int_{|\xi|\ge\delta} \frac{|f(x-\xi)|}{|\xi|}\frac{|\xi|}{h^d}K\left(\frac{\xi}{h}\right)d\xi + |f(x)|\int_{|\xi|\ge\delta}\frac{1}{h}K\left(\frac{\xi}{h}\right)d\xi \\
&\le \sup_{|\xi|\le\delta|} |f(x-\xi) - f(x)| \int_{\mathbb{R}^d} |K(\zeta)|d\zeta + \frac{1}{\delta}\sup_{|\zeta|\ge\delta/h^d} |\zeta||K(\zeta)| \int_{\mathbb{R}^d} |f(\xi)|d\xi \\
&\quad + |f(x)|\int_{\zeta\ge\delta/h^d} |K(\zeta)|d\zeta.
\end{aligned}$$

Now, (1.16) implies

$$\lim_{|y|\to\infty} |y||K(y)| = 0.$$

Making use of these properties, assume that $f \in \mathcal{L}^1(\mathbb{R}^d)$ and recalling that $h \overset{n\to\infty}{\to} 0$, we can let $n \to \infty$ and choose δ_n so that the right hand side converges to zero. □

We note that in fact (1.17) gives us the general convergence of such estimators. We will now make use of this result in estimating the variance of $\hat{f}_n$.

Lemma 1.3.2. *The variance of $\hat{f}_n(x)$ can be expressed as:*

$$\begin{aligned}
E(\hat{f}_n(x) - f_n(x))^2 &= \frac{f(x)}{nh^d}\int_{\mathbb{R}^d} K^2(y)dy\left(1 + O\left(\frac{1}{nh^d}\right)\right) \\
&\le B\frac{1}{nh^d}.
\end{aligned}$$

This appears in [Par62] as a consequence of Lemma 1.3.1.

Proof. Some simple manipulation gives us the variance of $\hat{f}_n(x)$ as:

$$\mathrm{Var}(\hat{f}_n(x)) = \frac{1}{n}\mathrm{Var}\left[\frac{1}{h^d}K\left(\frac{x-X}{h}\right)\right]$$

Now, by Lemma 1.3.1 we have

$$h^d E\left[\left\{\frac{1}{h^d}K\left(\frac{x-X}{h}\right)\right\}^2\right] = \frac{1}{h^d}\int_{\mathbb{R}^d} K^2\left(\frac{x-\xi}{h}\right) f(\xi)d\xi \to f(x)\int_{\mathbb{R}^d} K^2(\xi)d\xi$$

since (1.15) and (1.16) combine to give us that:

$$\int_{\mathbb{R}^d} K^2(\xi)d\xi < \infty$$

and K^2 satisfies all other conditions required in Lemma 1.3.1. We can now assert:

$$nh^d \mathrm{Var}(\hat{f}_n(x)) = f(x)\int_{\mathbb{R}^d} K^2(\xi)d\xi + o(1)$$

which gives us the required result. □

Lemma 1.3.3. *The bias term is bounded by:*

$$[E(f_n(x) - f(x))]^2 \leq Ah^{2m-2d/p}.$$

Proof. We can establish this as follows. First, by a change of variables we have:

$$E(f_n(x) - f(x)) = \int_{\mathbb{R}^d} K(-\xi)f(x+\xi h)d\xi - f(x). \tag{1.22}$$

Now we will take a Taylor expansion of f:

$$f(x+\xi h) = f(x) + \sum_{j=1}^{m-1} \frac{(\xi h\cdot\nabla)^j}{j!} f(x) + R_m(f;x,x+\xi h)$$

where $R_m(f;x,x+\xi h)$ is the m^{th} Taylor remainder term. Appendix A contains an explanation of the shorthand used here. Now, if we substitute this into (1.22) we can use (1.17) to cancel out $f(x)$. Further, (1.18) will allow us to ignore all terms apart from the Taylor remainder which gives us:

$$E(f_n(x) - f(x)) = \int_{\mathbb{R}^d} K(-\xi)R_m(f;x,x+\xi h)d\xi. \tag{1.23}$$

Let us examine the Taylor remainder, taking the identity given in Appendix A and changing the order of integration and twice making use of a Hőlder inequality with $\frac{1}{p}+\frac{1}{q}=1$ gives:

$$\begin{aligned}
\int_{\mathbb{R}^d} K(-\xi)R_m(f;x,x+\xi h)d\xi &= \int_0^1 \int_{\mathbb{R}^d} K(-\xi)\sum_{|\alpha|=m}\frac{\xi^\alpha h^m}{(m-1)!}(1-t)^{m-1}D^\alpha f(x+t\xi h)d\xi dt \\
&\leq \int_0^1 \left(\sum_{|\alpha|=m}\int_{\mathbb{R}^d}\left|\frac{\xi^\alpha h^m}{(m-1)!}K(-\xi)\right|^q d\xi\right)^{1/q}\left(\sum_{|\alpha|=m}\int_{\mathbb{R}^d}\left|(1-t)^{m-1}D^\alpha f(x+t\xi h)\right|^p d\xi\right)^{1/p} dt \\
&= Ch^m\int_0^1 (1-t)^{m-1}\left(\sum_{|\alpha|=m}\int_{\mathbb{R}^d}|D^\alpha f(x+t\xi h)|^p d\xi\right)^{1/p} dt
\end{aligned}$$

using (1.19). Now making a change of variables $\xi \to t\xi h$ on the inner integral gives us:

$$
\begin{aligned}
&\int_{\mathbb{R}^d} K(-\xi) R_m(f; x, x+\xi h) d\xi \\
&\quad \le C h^{m-d/p} \int_0^1 \frac{(1-t)^{m-1}}{t^{d/p}} \left(\sum_{|\alpha|=m} \int_{\mathbb{R}^d} |D^\alpha f(x+\xi)|^p \, d\xi \right)^{1/p} dt \\
&\quad = C |f|_{W_p^{(m)}} h^{m-d/p} \int_0^1 \frac{(1-t)^{m-1}}{t^{d/p}} dt \\
&\quad \le A^{1/2} h^{m-d/p}
\end{aligned}
$$

provided that $d > p$. Squaring provides the result. □

Proof of Theorem 1.3.3. Adding the bias and variance terms gives us:

$$
E(f(x) - \hat{f}_n(x))^2 \le A h^{2m-2d/p} + B \frac{1}{nh^d}. \tag{1.24}
$$

We will now define $k_n = nh^d$. Substituting this in (1.24) and minimising, we obtain that k_n has an optimal value of:

$$
k_n = \frac{B}{(2m - 2d/p)A} n^{(2m-2d/p)/(2m+d-2d/p)}
$$

which gives us:

$$
E(f(x) - \hat{f}_n(x))^2 \le D n^{(2m-2d/p)/(2m+d-2d/p)}.
$$

□

This is useful in showing that the bound on convergence is sharp. Moreover, we suspect that where the methods below can be regarded as kernel estimates, they satisfy the necessary conditions for Lemma 1.3.1 and therefore converge to the true density in expectation. We will note also that some justification is given here for ignoring the non-negativity condition on density estimates; for $m > 2$ the condition (1.18) will force Parzen kernels to take negative values, but this is necessary to ensure optimal convergence.

1.3.4 Optimal Convergence of the Histogram in W_p^1

[Wah75b] also demonstrates that particular classes of orthogonal series estimates and polynomial interpolations of histograms converge optimally. Additionally, it can be shown that for $m = 1$ the histogram achieves this convergence rate. The proof is relatively simple, and remarkably similar to the result for kernels above. The result will also be useful later and has therefore been included here. Once again the proof follows [Wah75b], with modifications to include multivariate densities. We again make the assumption that $d < p$. We can represent the value of a histogram at a point x in the hypercube $I_j = \prod_{k=1}^d [j_k h, (j_k+1)h]$ as:

$$\hat{f}_n(x) = \frac{Y_j}{nh^d}$$

where $Y_j =$ (no. of $X_1, \ldots, X_n$ in I_j). We can think of Y_j as having binomial distribution $\mathrm{Bi}(n, p_j)$ where $p_j = \int_{I_j} f(\xi) d\xi$. Using standard results about this distribution, we get:

$$E\hat{f}_n(x) = \frac{1}{h^d} p_j$$

$$\mathrm{Var}\hat{f}_n(x) = \frac{p_j(1-p_j)}{nh^{2d}} \leq \frac{\Lambda_j}{nh^d}.$$

Now it is easy to see that:

$$E\left(f(x) - \hat{f}_n(x)\right) = \frac{1}{h^d}\int_{I_j} \left(f(x) - f(\xi)\right) d\xi$$

Now we will take a slightly different form of the Taylor remainder to that used in Lemma 1.3.3 and obtain the following identity:

$$\begin{aligned}
\frac{1}{h^d}\int_{I_j} \left(f(x) - f(\xi)\right) d\xi &= \frac{1}{h^d}\int_{I_j}\int_0^1 (x-\xi)\cdot\nabla f(x + t(\xi - x)) dt d\xi \\
&\leq \frac{1}{h^d}\int_0^1 \left(\sum_{k=1}^d \int_{I_j} |(x_k - \xi_k)|^q \, d\xi\right)^{1/q} \left(\sum_{k=1}^d \int_{I_j} \left|\frac{\partial f}{\partial \xi_k}(x + t(x-\xi))\right|^p d\xi\right)^{1/p} dt \\
&\leq \; d|f|_{W_p^1} h^{(1+d)/q - d}\int_0^1 t^{-d/p} dt \\
&\leq M_j h^{1-d/p}
\end{aligned}$$

by essentially the same calculations for $\frac{1}{p} + \frac{1}{q} = 1$. Once again we assume that $d < p$. Then we have:

$$E\left(f(x) - \hat{f}_n(x)\right)^2 \leq M_j h^{(2-2d/p)} + \frac{\Lambda_j}{nh^d} \tag{1.25}$$

Now if we set $k_n = nh$ in (1.25) and minimise for k_n we arrive at:

$$k_n = \left[\frac{1}{(2-2d/p)}\frac{\Lambda_j}{M_j}\right] n^{(2-2d/p)/(2+d-2d/p)}.$$

Substituting we get:

$$E(f(x) - \hat{f}_n(x))^2 \leq D n^{-\phi_d(1,p)}$$

which completes our proof. We have used strict hypercubes here. It is possible to replace these with hyper-rectangles using different $\{h_i\}_{i=1}^d$ for each side. In cases where the density is scaled differently in different dimensions, this may well improve the rate of convergence.

We note that in fact the proof above has essentially provided the convergence of the mean integrated squared error. This result will be useful later on and we develop it here. We have actually used $\Lambda_j = \int_{I_j} p_j(1-p_j)$ above, so that

$$\frac{\Lambda}{nh^d} = \sum_j \frac{\Lambda_j}{nh^d}$$

is the integrated variance. Examining the bias term, we can also subsume an integral into M_j in (1.25). If we sum these as well, we achieve the following result.

Theorem 1.3.4. *Let a histogram $\hat{f}_n$ be constructed in d dimensions on a discretisation of hypercubes of side h using n independent identically distributed samples with common density f. Then it is possible to choose h so that:*

$$E\int_\Omega (\hat{f}_n(x) - f(x))^2 dx \le Dn^{-\phi_d(1,p)}.$$

1.3.5 Nonnegative Projections of Density Estimates

Many of the estimators which we have considered and the ones which we will develop fail the positivity condition - that is, they can take negative values. Indeed, we have already seen that the proofs that exist of the optimal convergence of Kernel estimates require the use of Parzen kernels which must take negative values. We present a result stating that, in an $\mathcal{L}^1$ sense, the nonnegative projection (defined below) of a density estimate is always better than one that is allowed to take negative values.

In particular, the following theorem is proved in [DG85]:

Theorem 1.3.5. *Let*

$$g^*(x) = \frac{g(x)I_A(x)}{\int_A g(x)dx},\ A = \{x : g(x) > 0\}$$

be the nonnegative projection of g, then:

$$\int |f(x) - g^*(x)|dx \le \int |f(x) - g(x)|dx.$$

Proof. Since $\int g(x)dx = 1$, $\int_A g(x)dx \ge 1$ and therefore $g \ge g*$ on A. Using the fact $\int f(x)dx = \int g(x)dx$, we have

$$\begin{aligned}
\int |f(x) - g^*(x)|dx &= 2\int (g^*(x) - f(x))_+ dx \\
&= 2\int_A (g^*(x) - f(x))_+ dx \\
&\leq 2\int_A (g(x) - f(x))_+ dx \\
&= 2\int (g(x) - f(x))_+ dx \\
&= \int |g(x) - f(x)|dx.
\end{aligned}$$

□

The use of the $\mathcal{L}^1$ norm has the advantage of being invariant under scaling $\mathbb{R}^1$ data. More explicitly if we have data $\{X_i\}_{i=1}^n$ with underlying density f that we estimate by $\hat{f}_n$, then $\{aX_i\}_{i=1}^n$ has underlying density $(1/a)f(x/a)$ and

$$\int |f(x) - \hat{f}_n(x)|dx = \int \left| \frac{1}{a} f\left(\frac{x}{a}\right) - \frac{1}{a}\hat{f}_n\left(\frac{x}{a}\right) \right| dx.$$

This allows us to compare how difficult it is to estimate different densities. The result holds only on $\mathbb{R}^1$ however and we shall not consider this error further. Nevertheless, the theorem does point to one possible method of dealing with negativity in density estimates which might be useful. There are disadvantages associated with this technique, however, in that it will destroy any differentiability which we may have in the estimate as well as any attempts which we might make in controlling moments.

Chapter 2

Density Splines

The methods discussed in Chapter 1 have the disadvantage of appearing to be somewhat *ad hoc* in formulation and reliant on some choice of function - orthogonal series, kernel, and so forth. What we will do here is introduce an attempt to produce a density estimate motivated by the use of splines. These are well established in the context of functional regression when a response variable is present and we hope to make use of the same techniques in producing density estimates. In doing this, we remove ourselves one step further from direct judgements about the underlying density. We now concern ourselves with picking a functional characterisation of the density that intuitively makes sense. The chapter will begin by providing a brief overview of spline methods as a background to the methods we will choose. We will then describe the density splines which are the main focus of the thesis. We will also relate these to an existing density estimate known as the histospline. Finally, we will make use of the finite element method, which will provide us with a scalable approximation.

2.1 Regression Spline Smoothing

We here provide a brief overview of the methods of standard spline interpolation. The reader is directed to [Wah90] for a full account. Let us suppose that we are given a set of data points x_i on which we wish to approximate a function with a known response $f(x_i) = y_i + \epsilon_i$ where the ϵ_i are a set of error terms with zero mean and variance σ^2. It is clear that straightforward interpolation of the data (creating some function f for which $f(x_i) = y_i$, by piecewise linear interpolants, for example) will lead to overfitting, and a solution that is too "rough". In order to counter this overfitting, we introduce a penalty functional to estimate the "roughness" of the approximation - usually some integral involving a differential operator. Our task then becomes the minimisation of both the infidelity to the data and of the penalty functional of the final estimator. Typically we would consider a functional along the lines of

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda J(f). \tag{2.1}$$

Here $J(f)$ is some measure of the smoothness of f and λ is a parameter which trades off the bias of the estimator versus its smoothness. The associated interpolation problem would be to match $f(x_i) = y_i$ while minimising $J(f)$. A general formulation for this penalty term is given in [DW82] as $J(u) = C(u, u)$ where:

$$C(u, v) = \sum_{|\alpha|,|\beta|=m} \int_\Omega a_{\alpha\beta}(x) D^\alpha u(x) D^\beta v(x) dx, \ a_{\alpha\beta} \in \mathcal{L}^\infty(\Omega)$$

is a bilinear form. On $H^m(\Omega)$ ($W_2^{(m)}(\Omega)$ in the notation of Chapter 1), $C(u, v)$ is continuous. If we further require that:

$$\sum_{|\alpha|,|\beta|=m} a_{\alpha\beta}(x) y_\alpha y_\beta > C \sum_{|\alpha|=m} y_\alpha^2 \tag{2.2}$$

for all $y = (y_1, \ldots, y_k)$ and $k =$ no. of $\{\alpha, |\alpha| = m\}$ we have that $C(u, u)^{1/2}$ is a semi-norm on $H^m(\Omega)$. This last condition is used to ensure the positivity of $C(u, u)$. At this point it becomes helpful to make use of the techniques of reproducing kernel Hilbert spaces. The reader is directed to Appendix D for an overview of the subject. In general, we will work in some reproducing kernel Hilbert space H_R on which $C(\cdot, \cdot)$ defines a semi-norm and assume that there is a decomposition $H_R = H_0 \oplus H_1$ where $C(\cdot, \cdot)$ defines a norm on H_1 and has null-space H_0. Explicitly, $||f + g||_{H_R} = ||f||_{H_0} + ||g||_{H_1}$ for $f \in H_0$, $g \in H_1$. If we let L_i be the functional $L_i : f \to f(t_i)$ we can re-write (2.1) as:

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - L_i f)^2 + ||Pf||_{H_R}^2$$

where P is the orthogonal projection of f onto H_1. The technique at this point is to represent the solution as:

$$\hat{f}_n = \sum_{\nu=1}^{m} c_\nu \phi_\nu + g$$

where $\{\phi_\nu\}_{\nu=1}^m$ spans H_0 and $g \in H_1$. We require that $< \phi_\nu, g > = 0$ for all $1 \leq \nu \leq m$ so that $g = Pg$. If we let η_i be the representor of L_i, the evaluation functional at the point t_i, and consider $\xi_i = P\eta_i$ we have that if ρ is perpendicular to $\{\xi_i\}_{i=1}^n$ and $\{\phi_\nu\}_{\nu=1}^m$ then $L_i\rho = 0$. If $g = g_1 + \rho$, $g_1 \in \mathrm{Span}\{\xi_1, \ldots, \xi_n\}$, then $||g||_{H_1}^2 = ||g_1||_{H_1}^2 + ||\rho||_{H_1}^2$ with $\sum_{i=1}^n (y_i - L_i(g_1 + \rho))^2 = \sum_{i=1}^n (y_i - L_i g)^2$. Then ρ only adds to the penalty term while leaving the bias unaffected and we can conclude that $\rho = 0$ from which we have:

$$\hat{f}_n = \sum_{\nu=1}^{m} c_\nu \phi_\nu + \sum_{i=1}^{n} d_i \xi_i. \tag{2.3}$$

The problem can now be reduced to a linear equation as follows. We note that P is self-adjoint [Wah90] and therefore that $< P\eta_i, u > = < \eta_i, Pu >$. This allows us to see that for $\hat{f}_n$ defined as in (2.3), $C(\hat{f}_n, \hat{f}_n) = d^T C d$ where $[C]_{ij} = C(\xi_i, \xi_j) =< \xi_i, \xi_j >$ and $d = (d_1, \ldots, d_n)$. We will also have that $L_i \xi_j =< \xi_i, \xi_j >$ from the fact that P is self-adjoint. The minimisation can now be restricted to finding vectors c and d to minimise:

$$\frac{1}{n} ||y - (Tc + Cd)||^2_{H_R} + \lambda d^T C d$$

where $[T]_{i\nu} = L_i \phi_\nu$. Some manipulation now allows us to reduce this to the set of linear equations

$$\begin{bmatrix} M & T \\ T & 0 \end{bmatrix} \begin{bmatrix} d \\ c \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

where $M = C + n\lambda I$. This can now be solved with standard matrix methods.

We have not yet made use of the fact that we are working in a reproducing kernel Hilbert space, except in requiring functional evaluation to be a bounded linear operator. We can make further use of the reproducing kernel, however, in evaluating the matrix C. Let H_R have reproducing kernel $Q(s,t) = Q_1(s,t) + R(s,t)$. We observe that $\xi_i(t) =< \xi_i, Q_t >$. Using the self-adjointness of the projection P, we have:

$$\begin{aligned} < \xi_i, Q_t > &= < \eta_i, PQ_t > \\ &= L_i R_t \\ &= R(t, t_i) \end{aligned}$$

Now let us look at the entries of C:

$$\begin{aligned} < \xi_i, \xi_j > &= L_i \xi_j \\ &= \xi_j(t_i) \\ &= R(t_i, t_j) \end{aligned}$$

and if we know R it is a straightforward calculation to evaluate the entries of C. In the case where we are working in $H^m(\Omega)$ we let

$$H_1 = \{u \in H^m(\Omega) : \int_\Omega D^\alpha u(x) dx = 0, |\alpha| \leq m\}$$

and observe that for any $u \in H^m(\Omega)$ we can find a unique q in the null-space of $C(u,u)$ such that $u_1 = u - q \in H_1$. This gives us the desired decomposition. We will revisit this idea in §2.3.

It is worthwhile noting here that this method, although exact, is very technical and the matrices that result are not sparse. It is attractive, therefore, to apply a finite element method to the problem. This approach is pursued in [HRA97], for example. Some variational calculus gives us that $\hat{f}_n$ satisfies:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{f}_n(t_i)s(t_i)+\lambda\sum_{|\alpha|,|\beta|=m}\int_{\Omega}a_{\alpha\beta}D^{\alpha}\hat{f}_n(x)D^{\beta}s(x)dx=\frac{1}{n}\sum_{i=1}^{n}s(t_i)y_i\ \forall s\in H^m$$

When we take a finite element basis, the resulting matrices are sparse and the system can easily be solved. In general, we expect to find a discretisation error of order h^{m+1} in the $\mathcal{L}^2$ norm where h is the size of the discretisation.

Of particular interest are thin plate splines, which use the penalty term

$$J_2(f)=\int\int(f_{x_1^2}^2+2f_{x_1x_2}^2+f_{x_2^2}^2)dx_1dx_2. \tag{2.4}$$

This is derived from the bending energy of metal sheets. In this case we have a null space spanned by $1, x_1, x_2$ and the reproducing kernel associated with J_2 is

$$Q(s,t)=\int G(s,u)G(t,u)du$$

where G is the Greens function for the partial differential equation $J_2(f)=\phi$. [Duc76] demonstrates that this comes out as:

$$\xi_i(t)=Q(t,t_i)=\frac{1}{2^3\pi}|t-t_i|^2\ln|t-t_i| \tag{2.5}$$

and we can proceed to solve the system. [HRA97] have examined a nonconforming finite element approximation to this problem with promising results; we will return to this in Chapter 3 and make use of the methods they develop in the context of the density splines defined below.

2.2 Density Splines

We have seen above that regression splines have a well-motivated existence as the minimiser of a sensible functional. The density splines that will be defined below are motivated by an attempt to apply similar principles in the context of density estimation. In particular, these will be shown to have some nice properties.

2.2.1 Producing an Estimator

What we wish to do now is to place a density estimation functional in the same context as those we have just outlined for the more standard regression problem. The idea is to begin with a highly undersmoothed estimate and apply the same motivation as we have produced above to the solution. Take $\hat{\rho_\epsilon}$ to be an initial $\mathcal{L}^2$ estimate; a naïve estimator with $h=\epsilon/2$ (see §1.2.2), say. We wish to minimise:

$$\int(\hat{f}_n(x)-\rho_\epsilon(x))^2dx+\lambda J_m(\hat{f}_n).$$

Translating this into a variational approach we have:

$$\int \hat{f}_n(x)s(x)dx + \lambda \sum_{|\alpha|,|\beta|=m} \int a_{\alpha\beta}(x)D^\alpha \hat{f}_n(x)D^\alpha s(x)dx = \int s(x)\rho_\epsilon(x)dx \;\forall s \in H^m. \tag{2.6}$$

Now the right-hand side of this equation can be re-interpreted as an approximation to $E_\rho(s)$. Making the assumption that we are working in a reproducing kernel Hilbert space, it can be replaced by:

$$\frac{1}{n}\sum_{i=1}^{n} s(X_i).$$

This is equivalent to letting $\epsilon \to 0$ and removing the dependence of $\hat{f}_n$ on the form of ρ_ϵ. The result will be equivalent to integrating s against a sum of δ-functions. This is plainly not $\mathcal{L}^2$ and we will therefore regard (2.6) as the defining equation for $\hat{f}_n$. Note that using this approximation we are assuming that the x_i are given exactly as a set of data points. It is worthwhile observing here that the method has an advantage over kernel methods in allowing us to consider other approximations of $E_\rho(s)$ when we are given aggregated data or there is some known degree of error in the placement of the x_i. In such a case, we could use a histogram or naive estimator for ρ_ϵ in the minimisation problem, although the latter may have consequences for the scalability of the algorithm later.

Within the formulation that we have, we consider the equation:

$$\int u(x)s(x)dx + \lambda C(u,s) = \frac{1}{n}\sum_{i=1}^{n} s(x_i), \;\; \forall s \in H^m(\Omega). \tag{2.7}$$

C is continuous, $C(u,u)$ is a topologically equivalent semi-norm to $|\cdot|_{H^m(\Omega)}$ by (2.2) and so

$$\int u(x)s(x)dx + C(u,s)$$

is equivalent to the usual norm on $H^m(\Omega)$ [Tri78]. If we are working in a reproducing kernel Hilbert space - using $2m \geq d$ where d is the dimension of Ω will be sufficient - then functional evaluation, and therefore the right-hand side, is also a bounded operator. We can use therefore the Lax-Milgram Theorem to conclude that this problem has a unique solution $\hat{f}_n \in H^m(\Omega)$. In cases where we do not have a reproducing kernel Hilbert space, the right-hand side of (2.7) will not be bounded. There may be times when we do wish to use penalty terms that are less strong and in this case we will return to the right-hand side of (2.6) and first aggregate data into a histogram against which we may integrate s.

If we take our domain to be $\mathbb{R}^n$, and assume that all derivatives tend to zero at infinity (which is reasonable for a density estimate), we can integrate by parts in (2.7) to derive:

$$\int u(x)s(x)dx + \lambda \sum_{|\alpha|,|\beta|=m} \int (-1)^m D^\beta(a_{\alpha\beta}(x)D^\alpha u(x))s(x)dx = \frac{1}{n}\sum_{i=1}^{n}\int \delta_{x_i}(x)s(x)dx$$

from whence we derive the Euler equations:

$$u(x) + \lambda \sum_{|\alpha|,|\beta|=m} (-1)^m D^\beta(a_{\alpha\beta}(x) D^\alpha u(x)) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(x)$$

which is an average of shifted Green's functions satisfying:

$$g(x) + \lambda \sum_{|\alpha|,|\beta|=m} (-1)^m D^\beta(a_{\alpha\beta}(x) D^\alpha g(x)) = \delta(x).$$

We can regard this as a kernel estimate with smoothing parameter λ. More explicitly, if

$$K(x) + \sum_{|\alpha|,|\beta|=m} (-1)^m D^\beta(a_{\alpha\beta}(x) D^\alpha K(x)) = \delta(x)$$

we can give an explicit formula for the kernel estimate:

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda^{1/(2m)}}\right).$$

We can state that, at least in unbounded domains, the approach above can be regarded as a method of choosing a kernel. We will see below that we can attach some worthwhile properties to this choice. In fact, a similar analysis could be made on bounded domains, with some rather complicated boundary conditions. Doing this, however, we will no longer be able to produce kernels in the form above, but will instead find that the kernel equation incorporates some dependence on the placement of the X_i.

2.2.2 Examples and Consequences

Here we will provide a demonstration of the above method in practical terms. Consider the univariate thin plate spline, minimising:

$$J(f) = \int (f(x) - \rho_\epsilon)(x))^2 dx + \lambda \int s''(x)^2 dx$$

from which we derive the variational equations:

$$\int f(x)s(x)dx + \lambda \int f''(x)s''(x)dx = \frac{1}{n}\sum_{i=1}^{n} s(x_i)\ \forall s \in W_2^2(\mathbb{R})$$

and the Euler equation:

$$f(x) + \lambda f^{(4)}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)$$

If we have:

$$g(x) + g^{(4)}(x) = \delta(x) \tag{2.8}$$

then

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} g\left(\frac{x - x_i}{\lambda^{1/4}}\right).$$

We can give the Green's function satisfying (2.8) as:

$$g(x) = \frac{1}{\sqrt{2}} e^{-|x|/\sqrt{2}} \left(\sin(|x|/\sqrt{2}) + \cos(|x|/\sqrt{2})\right)$$

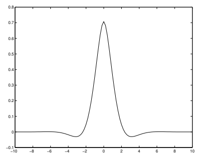


Figure 2.1: The univariate thin plate spline density kernel

It is immediately apparent that this estimator has at least failed to meet the requirements of non-negativity. This is a failing common to many of the more sophisticated estimators; it was implied in the definitions of Parzen kernels and will be true of most orthogonal series estimators.

It is worthwhile commenting on some consequences of the variational form we have used. It is immediate, to begin with, that the sample moment of s is preserved for all s annihilated by $a_{\alpha\beta}D^{\alpha}$ for all $|\alpha| = m$ in the sense that $\frac{1}{n}\sum_{i=1}^{n} s(x_i) = E_{\hat{f}_n}(s)$. This is simply demonstrated by substituting s into (2.7). Substituting $s = 1$ also ensures that the estimate that we get is a density at least in so far as it integrates to one.

This is a simple method of obtaining kernels with a generalisation of the moment preserving condition provided in [DG85]. In the case of the thin plate spline above, we have preserved first moments. If we wanted also to preserve variance, we could use a penalty term involving third derivatives. In fact, so long as we can find a differential operator annihilating some set of functions, then we can construct an estimate which will preserve that set of moments.

For the case $a_{\alpha\beta}$ non-zero constants, $\Omega = \mathbb{R}^d$, it is easy to show that we can better this by integrating by parts within the definition of the penalty term in (2.7) and demonstrate that we have preserved up to the $2m^{\text{th}}$ moments. More selective choices of $a_{\alpha\beta}$ will allow us to be more selective in the moments we preserve.

2.2.3 Convergence Results

Considering the thin plate spline kernel derived above, some fairly simple calculations will show that it satisfies the conditions for optimal convergence in §1.3.3 with $m = 2$, $1 \leq p \leq \infty$. It is clear, in fact, that any density spline kernel produced on the entire domain will have the properties (1.18), (1.15) and (1.17). This can be seen by substituting $s(x) = x^{|i|}$, $s(x) = \hat{f}_n(x)$ and $s(x) = 1$ respectively into (2.7) with $X_i = X_1 = 0$ and using the fact that we know $\hat{f}_n \in H^m(\mathbb{R}^d)$. We have no proof of (1.16) or (1.19) but it seems likely that such a result could be achieved.

This, of course, will not prove the result for the case of bounded domains on which the resulting kernels will no longer be translationally invariant and will therefore require a somewhat more sophisticated treatment. In the following section we will prove a convergence result in which data is first aggregated into a histogram which is then smoothed. It is reasonable to expect that density splines produced without such aggregation will perform at least as well.

Convergence of the Density Spline with Aggregated Data

There are many cases in which we will deal with aggregated data: when it is presented as such, for example, but also in cases when the boundedness of pointwise evaluation prevents us from using the standard right-hand side. In this less satisfactory case we are able to demonstrate that we do achieve convergence. Here the defining equation for $\hat{f}$ will be:

$$\int_\Omega \hat{f}(x)s(x)dx + \lambda \int_\Omega L\hat{f}(x)Ls(x)dx = \int_\Omega h(x)s(x)dx, \ \forall s \in H^m(\Omega) \tag{2.9}$$

where h is the histogram of aggregate data, and L is an m^{th} order differential operator such that

$$\int_\Omega (Ls(x))^2 dx \preceq |s|_{H^m(\Omega)}$$

where $a_k \preceq b_k$ is intended to signify $a_k \leq Cb_k$ for some constant C independent of k. We will also define a function f_ρ by the equation

$$\int_\Omega f_\rho(x)s(x)dx + \lambda \int_\Omega Lf_\rho(x)L(x)dx = \int_\Omega \rho(x)s(x)dx, \ \forall s \in H^m(\Omega) \tag{2.10}$$

.

We can now prove the following theorem:

Theorem 2.2.1. *For $\hat{f}$ defined as above, the following estimate holds:*

$$||\hat{f} - \rho||^2_{\mathcal{L}^2(\Omega)} \preceq ||h - \rho||^2_{\mathcal{L}^2(\Omega)} + \lambda|\rho|^2_{H^m(\Omega)}$$

Proof. Subtracting (2.10) from (2.9) and substituting $s = \hat{f} - \rho$ yields:

$$\int_\Omega (\hat{f}(x) - f_\rho(x))^2 dx + \lambda \int_\Omega (L\hat{f}(x) - Lf_\rho(x))^2 dx = \int_\Omega (h(x) - \rho(x))(\hat{f}(x) - f_\rho(x))dx$$

and using the Cauchy-Schwartz inequality and the positivity of the penalty term we get:

$$||\hat{f} - f_\rho||_{\mathcal{L}^2(\Omega)} \preceq ||h - \rho||_{\mathcal{L}^2(\Omega)}. \tag{2.11}$$

We will now turn to equation (2.10). Re-arranging and subtracting $\lambda \int_\Omega L\rho(x)s(x)$ from both sides, we can substitute $s = f_\rho - \rho$ to get:

$$\int_\Omega (f_\rho(x) - \rho(x))^2 dx + \lambda \int_\Omega (Lf_\rho(x) - L\rho(x))^2 dx = \lambda \int_\Omega L\rho(x)(L\rho(x) - Lf_\rho(x))dx. \tag{2.12}$$

Ignoring the first term, we have that

$$\lambda ||Lf_\rho - L\rho||^2_{\mathcal{L}^2(\Omega)} \leq \lambda |\rho|_{H^m(\Omega)} ||Lf_\rho - L\rho||_{\mathcal{L}^2(\Omega)},$$

which gives us

$$||Lf_\rho - L\rho||_{\mathcal{L}^2(\Omega)} \leq |\rho|_{H^m(\Omega)}.$$

Substituting this into (2.12) and ignoring the second term this time, we arrive at

$$||f_\rho - \rho||_{\mathcal{L}^2(\Omega)} \preceq \lambda^{1/2} |\rho|_{H^m(\Omega)} \tag{2.13}$$

and a triangle inequality on (2.11) and (2.13) gives us the result. □

Now, from §1.3.4 we know that if ρ is the underlying density function for a sample of independent observations $\{X_i\}_{i=1}^n$, then it is possible to choose the histogram h, constructed from the X_i so that

$$E \int_\Omega (h(x) - \rho(x))^2 dx \leq Dn^{-\phi_d(1,p)}$$

for ρ in $W_p^{(1)}(\Omega)$, which we certainly have. A suitable choice of λ will then give us:

$$\begin{aligned} E \int_\Omega (\hat{f} - \rho)^2 dx &\leq E \int_\Omega (h(x) - \rho(x))^2 dx + \lambda |\rho|^2_{H^m(\Omega)} \\ &\leq Dn^{-\phi_d(1,p)}. \end{aligned}$$

In fact, in smoothing the histogram, we would expect to gain a rate of convergence significantly better than this. We would also expect that the density splines using a pointwise estimation of expectation would converge at least as well. Proofs of these conjectures, however, are not yet known.

2.3 Generalised Histosplines

A related estimator to that described above was developed in the early 1970's by Boneva, Kendall and Stefanov [BKS71], which they labelled the histospline. As the name suggests, this involves a spline smoothing of the histogram and can be considered useful when dealing with aggregated data. These techniques were employed in [Wah81], for example, in producing a graphical representation of cancer mortality rates for the United States that were given on a state-by-state basis. [BKS71] defined the problem on univariate data, minimising first derivatives only, a general description of the problem is given in [DW82]:

Problem 2.3.1. *Let Ω be a smooth bounded subset of $\mathbb{R}^d$. Find $u \in W_m^2(\Omega)$ to minimise $J(u) = C(u,u)$ where*

$$C(u,v) = \sum_{|\alpha|,|\beta|=m} \int_\Omega a_{\alpha\beta}(x) D^\alpha u(x) D^\beta v(x) dx, \ \ a_{\alpha\beta} \in \mathcal{L}^\infty(\Omega)$$

subject to:

$$\int_\Omega \psi_i(x) u(x) = s_i, \ \ i = 1, \dots, N.$$

Where $a_{\alpha\beta}$ are polynomials satisfying the regularity conditions (2.2) and $\{\psi_i\}_{i=1}^N$ are a linearly independent set of functions in $\mathcal{L}^2(\Omega)$.

In essence, this might be considered to be an interpolation variant on the regression problem described above. The application considered here will be to a somewhat more specific problem. As with the cases above, our bilinear forms will typically be symmetric and integrals of differential operators, and the ψ_i will be taken as indicator functions of bins in a histogram. This, of course, immediately leads us to infer that the function we have is a density, at least in so far as it integrates to 1. It may also be useful to consider the preservation of higher moments in a technique similar to that of the density splines developed above by requiring:

$$\psi_j^s = s(x) I_{\Omega_j}(x)$$

where Ω_j represents a histogram bin. This would then allow us to explicitly preserve the sample moment about any $\mathcal{L}^2(\Omega)$-function without requiring a particular form of the smoothing penalty term.

2.3.1 Producing a Solution

An exact solution to the generalised histospline problem can be produced using techniques reminiscent of those for standard regression splines. We are concerned to give more than a simple overview here, however, and will treat the subject a little more formally. In particular we will begin by demonstrating the existence and uniqueness of solutions to Problem 2.3.1. The variational form of this problem will be employed later in using the finite element method and we will introduce it here:

Problem 2.3.2. *Find $u \in H^m(\Omega)$, $q \in \mathrm{Span}\{\psi_1, \dots, \psi_N\}$ that satisfy:*

$$C(u,v) + \int_\Omega v(x)q(x)dx = 0 \tag{2.14}$$

with the matching conditions:

$$\int_\Omega u(x)\psi_i(x) = s_i. \tag{2.15}$$

We can now proceed to demonstrate the existence and uniqueness of solutions to this problem. The following construction may be found in [DW82].

Lemma 2.3.1. *On the subspace*

$$H_0 = \{u \in H^m(\Omega) : \int_\Omega D^\alpha u(x)dx = 0, |\alpha| \leq m\} \tag{2.16}$$

$\sqrt{C(u,u)}$ *is an equivalent norm to* $||u||_{H^m(\Omega)}$.

Proof. The continuity of $C(u,u)$ in $H^m(\Omega)$ has already been shown. To prove coercivity, we will iterate the Poincaré inequality, for $0 \leq s \leq m$:

$$\sum_{|\alpha|=s} \int_\Omega D^\alpha u(x)^2 dx \preceq \sum_{|\alpha|=m} \int_\Omega D^\alpha u(x)^2 dx + \sum_{s \leq |\alpha| < m} \left[\int_\Omega D^\alpha u(x)dx\right]^2$$

and using (2.16) and (2.2) we have that

$$||u||^2_{H^m(\Omega)} \preceq \sum_{|\alpha|=m} \int_\Omega D^\alpha u(x)^2 dx \preceq C(u,u).$$

□

The idea now is to transform the problem so that we may use a space on which $C(u,u)$ is coercive. We note that the null space of $C(u,u)$, which we will denote by Q, is finite dimensional, and let us suppose that it is spanned by $\{q_\nu\}_{\nu=1}^M$ for $M < N$. We will further suppose that the $\{\psi_i\}_{i=1}^N$ are linearly independent over this null space and we can therefore choose the q_ν so that

$$\int_\Omega q_\nu \psi_{N-M+i} = \delta_{ij},\ \nu, i = 1, \dots, M \tag{2.17}$$

where δ_{ij} represents the Kronecker delta.

Lemma 2.3.2. *In the subspace*

$$H_1(\Omega) = \{u \in H^m(\Omega) : \int_\Omega u(x)\psi_i(x)dx = 0,\ i = N-M+1, \dots, N\} \tag{2.18}$$

$\sqrt{C(u,u)}$ *is an equivalent norm to* $||u||_{H^m(\Omega)}$.

Proof. This rests on the fact that for any $u \in H^1(\Omega)$, it is possible to find a q in Q such that $u_0 = u - q \in H_0(\Omega)$. Further, if we let $q = \sum_{\nu=1}^{M} a_\nu q_\nu$ then using (2.17) we get

$$-\int_\Omega u_0(x)\psi_{N-M+i}(x)dx = \int_\Omega (u(x) - u_0(x))\psi_{N-M+i}(x)dx = a_i,\ i = N-M+1, \dots N$$

and therefore

$$u(x) = u_0(x) + \sum_{\nu=1}^{M} q_\nu(x) \int_\Omega u_0(x)\psi_{N-M+\nu}(x)dx.$$

In view of Lemma 2.3.1 we have:

$$\begin{aligned} ||u||_{H^m(\Omega)} &\preceq ||u_0||_{H^m(\Omega)} + \sum_{\nu=1}^{M} ||q_\nu||_{H^m(\Omega)} ||\psi_{N-M+\nu}||_{\mathcal{L}^2(\Omega)} ||u_0||_{H^m(\Omega)} \\ &\preceq \sqrt{C(u_0, u_0)} = \sqrt{C(u,u)} \end{aligned}$$

□

Now we note that this satisfies the assumptions of [BF91][Proposition 1.1] as H_1 contains the null space of $\int_\Omega u(x)\phi(x)dx$ for each $q \in \text{Span}\{\phi_1, \dots, \phi_N\}$. This provides the existence and uniqueness of a solution.

2.3.2 An Explicit Solution

Producing a solution explicitly, we can transform the problem by the observation that for u satisfying (2.15),

$$\tilde{u}(x) = u(x) - \sum_{\nu=1}^{M} s_{N-M+\nu} q_\nu(x) \in H_1(\Omega)$$

and satisfies:

$$\int_\Omega \tilde{u}(x)\psi_j(x)dx = s_j - \sum_{\nu=1}^{M} s_{N-M+\nu} \int_\Omega q_\nu(x)\psi_j(x)dx = \tilde{s}_j.$$

Solving Problem 2.3.2 is now equivalent to minimising $C(\tilde{u}, \tilde{u})$ subject to:

$$\int_\Omega \tilde{u}(x)\tilde{\psi}_j(x)dx = \tilde{s_j},\ j = 1, \dots, N-M$$

with

$$\tilde{\psi_j}(x) = \psi_j(x) - \sum_{i=1}^{M} \alpha_{ij}\psi_{N-M+i}(x).$$

(2.17) allows us to choose α_{ij} so that

$$\int \tilde{\psi}_j(x)q(x)dx = 0 \ \forall q \in Q.$$

Now, considering $A(u,v)$ to be the inner product on this space we can use the Riesz representation theorem to find ξ_i such that:

$$A(u,\xi_i) = \int u(x)\tilde{\psi}_i(x)dx \ \ i = 1,\ldots,N-M$$

and given that the $\tilde{\psi}_i$ are linearly independent, we have that the ξ_i are also. At this point we observe that

$$A(q,\xi_i) = \int_\Omega q\tilde{\psi}(x) = 0$$

by the choice above. We can use the same arguments as we saw in §2.1 to show that

$$\tilde{u}(x) = \sum_{i=1}^{N-M} c_i\xi_i(x)$$

and retrieve $u(x)$. In order to produce the c_i it is enough to solve the linear equation:

$$C\mathbf{c} = \mathbf{s}$$

where $[C]_{ij} = C(\xi_i,\xi_j)$, $\mathbf{c}_i = c_i$ and $\mathbf{s}_j = \tilde{s_j}$.

In order to recover the full system, however, it is necessary to first solve the linear equations necessary to get $\tilde{u}$ and $\{\tilde{\psi}_i\}_{i=1}^{N-M}$. In general, these equations will be dense.

[DW82] demonstrates that for $\psi_j(x) = I_{\Omega_j}(x)$ the solution will be of a form

$$\sum_{j=1}^{N} c_j\phi_j(x) + \sum_{\nu=1}^{M} d_\nu q_\nu(x)$$

where ϕ_j solves:

$$\begin{array}{l} \sum_{|\alpha|,|\beta|=m}(-1)^m D^\beta\left(a_{\alpha\beta}(x)D^\alpha\phi_j(x)\right) = 1, \ x \in \Omega_j \\ \sum_{|\alpha|,|\beta|=m}(-1)^m D^\beta\left(a_{\alpha\beta}(x)D^\alpha\phi_j(x)\right) = 0, \ x \notin \Omega_j \end{array} \tag{2.19}$$

and the linear system will become

$$\begin{aligned} K\mathbf{c} + T\mathbf{d} &= \mathbf{s} \\ T^T\mathbf{c} &= 0 \end{aligned} \tag{2.20}$$

where $[K]_{ij} = \int_{\Omega_i}\phi_j(x)dx$ and $T_{j\nu} = \int_{\Omega_j} q_\nu(x)dx$.

2.3.3 Producing an Approximation

Clearly, producing a solution to the equations (2.19) will not be easy, and [DW82] therefore reduce the penalty term to the thin plate spline (2.4) and make an approximation to the system on a fine pointwise grid. Essentially the idea is to approximate the integral by the average of a large number of pointwise evaluations:

$$\int_{\Omega_j} \phi_i(x)dx \approx \frac{|\Omega_j|}{n_j} \sum_{k=1}^{n_i} \phi_i(x_k).$$

Let $a_i = |\Omega_i|/n_i$ where n_i is the number of mesh points on Ω_i. We now minimise $J_2(u)$ subject to:

$$a_i \sum_{t_l \in \Omega_i} u(t_l) = s_i.$$

This produces a system similar to (2.20) with

$$[K]_{ij} = a_i a_j \sum_{t_k \in \Omega_i} \sum_{t_l \in \Omega_j} E(t_k - t_l)$$

and

$$T_{j\nu} = a_i \sum_{tk \in \Omega_j} q_\nu(t_k).$$

The solution is then of the form:

$$u(x) = \sum_{i=1}^{N} c_i \eta_i(x) + \sum_{\nu=1}^{M} q_\nu(x)$$

where

$$\eta_i(x) = a_i \sum_{t_l \in \Omega_i} E(x - t_l)$$

and E is the fundamental solution to the thin plate spline equation (2.5). The approximation saves the mathematical effort of solving (2.19) but does little for the computational aspects of solving the problem. It also makes the resulting density considerably more computationally expensive to estimate. Moreover, the functions (2.5) will increase quadratically away from the origin, violating the unit integrand requirement outside the area we are considering. This behaviour is apparent in the computational experiments performed in [Wah81]. A more efficient approximation using the finite element method is shown below.

2.3.4 Optimal Convergence of the Ordinary Histospline

The results presented here provide the convergence of the original univariate histospline produced by [BKS71] on a bounded domain with a regular histogram. In particular, it solves the following:

Problem 2.3.3. *Minimise:*

$$\int_o^1 (g'(x))^2 dx$$

subject to:

$$\int_{jh}^{(j+1)h} g(x)dx = h_j, \ j = 0, 1, \ldots, N.$$

This is probably the simplest example of the histospline that we have, and the first developed. [Wah75a] provides the following theorem:

Theorem 2.3.1. *Suppose that $f \in W_p^{(m)}([0,1])$. Then*

$$E(f(x) - \hat{f}_n(x))^2 \leq Dn^{-\phi_1(m,p)}$$

for the following cases:

$$\begin{aligned} m = 1, &\quad p = 2, \quad p = \infty \\ m = 2, 3, &\quad 1 \leq p \leq e, \quad p = \infty. \end{aligned}$$

The proof is long and rather tedious in this case, so we will only provide a sketch here. It is interesting primarily in the transformation that [Wah75a] makes in the problem.

Proof. The essential idea here is to transform the problem into a standard spline-fitting problem on the distribution function. Suppose that $\hat{F}$ minimises

$$\int_0^1 (\hat{F}''(x))^2 dx$$

subject to:

$$\hat{F}(ih) = \sum_{j=0}^{i-1} h_j, \ i = 1, \ldots, N+1.$$

Then $\frac{d}{dx}\hat{F}(x)$ solves Problem 2.3.3. This fact actually rests on fixing the end points of $\hat{f}_n$. Typically, we would require $\hat{F}'(0) = \hat{F}'(1) = 0$. [Wah75a] extends the proof to cover a given choice of end-points which can be proved to converge to the real values of $f(0)$ and $f(1)$ at an appropriate rate.

We now introduce the spline interpolant of the true distribution function. $\tilde{F}(x)$ minimises:

$$\int_0^1 (\tilde{F}''(x))^2 dx$$

subject to:

$$\tilde{F}(ih) = F(ih),\ i = 1, \dots, N$$

where F is the true distribution function corresponding to the density in question. It can be seen that $\tilde{F}$ is the natural cubic spline interpolant of F. We can now write:

$$\begin{aligned} f(x) - \hat{f}_n(x) &= \frac{d}{dx}\left(F(x) - \hat{F}(x)\right) \\ &= \frac{d}{dx}\left(F(x) - \tilde{F}(x)\right) + \frac{d}{dx}\left(\tilde{F}(x) - \hat{F}(x)\right) \\ &= \frac{d}{dx}\left(F(x) - \tilde{F}(x)\right) + \frac{d}{dx}H_n(x). \end{aligned}$$

In fact this is closely analogous to the standard bias and variance terms that we would normally use. We can think of $\hat{F}$ as interpolating a set of points which can be regarded as random variables with mean at the true value of F at those points. In this case, the expectation of $\hat{F}$ should approach $\tilde{F}$. Heuristic motivation aside, we can still express:

$$E\left(f(x) - \hat{f}_n(x)\right)^2 \leq 2\left(\frac{d}{dx}(F(x) - \tilde{F}(x))\right)^2 + 2E\left(H_n(x)\right)^2$$

as $\tilde{F}$ depends only on the true distribution F. At this point we will confine ourselves to citing the relevant lemmata from [Wah75a] to arrive at the conclusion. To begin with, we find a familiar bound on what might be described as the bias.

Lemma 2.3.3. *Let $f \in W_p^{(m)}([0,1])$, for $m = 1$, $p = 2, \infty$ or $m = 2, 3$, $1 \leq p \leq 2$, $p = \infty$ and $\tilde{F}$ defined as above. Then*

$$\left(\frac{d}{dx}(F(x) - \tilde{F}(x))\right)^2 \preceq ||f^m||_p^2 h^{2m-2/p}.$$

We can also provide a bound on the variance:

Lemma 2.3.4.

$$E\left(H_n(x)\right)^2 \preceq \frac{1}{nh} + ||f^m||_p^2 h^{2m-2/p}.$$

Putting these together we find

$$E\left(f(x) - \hat{f}_n(x)\right)^2 \leq 2\left(\frac{d}{dx}(F(x) - \tilde{f}(x))\right) \preceq A||f^m||_p^2 h^{2m-2/p} + \frac{B}{nh}. \tag{2.21}$$

Setting $k_n = nh$ and minimising (2.21) as in §1.3.3 and §1.3.4 we achieve the result. □

Intuitively, it should be possible to extend this to higher penalty terms and to multivariate densities. The development of such a theory would be amply sufficient for a thesis in itself, and we will therefore leave it here. It should be noted that this approach will not obviously work in proving the convergence of the density splines we describe. This is essentially due to the difficulty of finding a functional on distribution functions whose solution gives the density splines as we describe them.

2.3.5 Comparisons with Density Splines

We can view the relationship between the histosplines developed here and the density splines above as somewhat akin to that between the interpolation and regression problems encountered in ordinary spline smoothing. It is less clear in this case, however, that the density splines proposed above are the correct formulation of the regression problem. We will argue here that they represent a more accurate representation than the alternatives. Consider the formulation of the problem given in [DW82] for inexact data:

Problem 2.3.4. *Minimise:*

$$\sum_{i=1}^{n} \omega_i \left[\int_{\Omega_i} g(x)dx - Z_i \right]^2 + \lambda J_m(g).$$

This represents the main alternative to the density spline formulation using a histogram estimate for ρ_ϵ. The main difference here being that instead of taking the integrated squared error with the histogram as the bias term, we have taken the error in the volume estimates over each bin. In some sense this is closer to the spline regression problem, treating each bin as a pointwise value. It also has the nice property of being able to control the local accuracy of the estimate by the use of the parameters ω_i. A natural choice would be $\omega_i = \mathrm{Var} Z_i$. Let us consider the variational form of Problem 2.3.4, however:

$$\sum_{i=1}^{N} \omega_i \int_{\Omega_i} g(x)dx \int_{\Omega_i} s(x)dx + \lambda C(g,s) = \sum_{\Omega_i} \omega_i Z_i \int_{\Omega_i} S(x)dx, \ \forall s \in H^m\Omega.$$

This takes on something like the form of the estimates given for the density splines:

$$\tilde{E}_g(s) + \lambda C(g,s) = \tilde{E}_Z(s), \ \forall s \in H^m(\Omega). \tag{2.22}$$

What is different in this formulation is that we have replaced the exact moment of s about g with an approximation, first projecting g onto a histogram. This has allowed us to control the local bias of g, but it has come at the expense of relinquishing some of the nicer properties of the density spline. To begin with, introducing the approximation $\tilde{E}$ is likely to create another set of errors in the estimate, which may well lead to slower convergence. We are not able to make use of pointwise data without aggregation in the estimate, either. It is reasonable to expect that the use of pointwise data should improve the order of accuracy of the estimate, and this cannot be done easily while controlling the local bias of the solution. Finally, while it is possible to substitute $s = 1$ in (2.22) and see that the resulting density has unit integrand, this will not be the case for higher moments; we will have preserved the moment against the density projected onto a histogram instead of against the density itself. Aggregated data will particularly be used when no other is available, and often this will entail relatively large bins. The projection of g onto a histogram may therefore introduce a significant error in those moments which are not specifically preserved.

Despite this, there are a number of practical advantages that the generalised histospline possesses. To begin with, we will not face the problem of choosing a smoothing parameter, discussed in Chapter 4, and we already have some analysis of rates of convergence. The formulation also incorporates greater flexibility in removing from us the problems due to the Sobolev embedding theorem and the need for higher-order penalty terms. Less positively, the resulting problems are saddle point problems and the resulting linear algebra will be less efficient.

2.4 Approximating a Solution

Mathematically, the nice aspect of the formulation above is the existence of a functional definition of the estimate. We have already been able to use this to produce a relatively simple analysis of some of the properties of these estimates. We will now make further use of it in providing a finite element approximation. While we have shown that it is possible to produce an exact solution for both the density spline and histospline problems, and that we can do so for fairly simple cases, this will not necessarily be possible on bounded domains or for more complicated penalty terms. Additionally, it will turn out to be computationally cheaper to deal with these approximations than with the exact solutions.

2.4.1 Finite Element Density Splines

It is clear from the working above that in particular cases we are able to get an exact solution in terms of kernel estimators. When we use bounded regions or more complicated penalty terms, however, the density spline kernels will be more difficult to discover. There may well not be analytical solutions and they will certainly no longer be translationally invariant. This task is also made complicated when our data is aggregated or not given exactly, in which case we can no longer represent it as a sum of δ-functions. What we propose as an alternative is to employ a finite element method to approximate the solution. The variational formulation given in (2.7) readily admits a finite element application and it is hoped that this may yield a more computationally efficient solution to the problem.

The computational problem associated with kernel estimators involves functional evaluation. For a data set of size n, the evaluation of the corresponding kernel estimate at each point potentially requires n computations. This may be reduced in a simplistic manner if we use kernels with compact support but this will not be the case for the Green's functions defined above. In the univariate case, [Sil86] has suggested an evaluation technique using a fast Fourier transform on a regular grid, some spline interpolation could then be used to evaluate the function between points.

This does not represent a solution to the problem of evaluating the density spline, however, as it is only applicable to known, translationally invariant kernels. In any case the finite element method will achieve the same order of error.

Turning to the application of the finite element method to the problem, it is clear that the uniqueness of the solution to the variational principle (2.7) implies existence and uniqueness in the approximating finite element subspace spanned by $\{\phi_i\}_{i=1}^k$ defined on a discretisation T_h of the domain. We obtain a matrix equation:

$$(A + \lambda C)\mathbf{x} = \mathbf{b}$$

where $[A]_{ij} = \int \phi_i(x)\phi_j(x)dx$, $[C]_{ij} = \sum_{|\alpha|,|\beta|=m} \int a_{\alpha\beta} D^\alpha \phi_i(x) D^\beta \phi_j(x)dx$ and $\mathbf{b}_j = n^{-1} \sum_{i=1}^n \phi_j(X_i)$.

Clearly, we require $O(n)$ operations to evaluate $\mathbf{b}$. Regarding $A + \lambda C$, it is fairly easy to see that the norm:

$$||u||_{\mathcal{L}^2} + C(u, u)$$

is equivalent to $\mathbf{x}^T(A + \lambda C)\mathbf{x}$ for

$$u(x) = \sum_{i=1}^k \mathbf{x}_i \phi_i(x)$$

and therefore that $A + \lambda C$ is positive definite. When $\{\phi_i\}_{i=1}^k$ is chosen in the usual manner, $A + \lambda C$ is sparse and may therefore be inverted in at worst $O(k^3)$ operations. We hope that sparse methods may allow us to produce a method requiring only $O(k)$ operations. If we have $k \ll n$ we can compute the general formula:

$$\hat{f}_n(x) = \mathbf{x}^T \phi(x)$$

in (hopefully) $O(n + k)$ operations and evaluate $\hat{f}_n$ in just $O(k)$ operations. We will see later that the discrete form of the problem that we have here will provide similar computational savings when we discuss automated choices of smoothing parameters in Chapter 3. We also have an alternative motivation of the density spline. We can regard the finite element density spline as the projection of the true density into $\mathrm{Span}\{\phi_1, \ldots \phi_n\}$. In fact, the projection that we are able to produce is that of a sum of δ-functions. This then requires the smoothing penalty term as a correction factor.

If we wish to examine the properties of the resulting estimate, it is immediately apparent that the same empirical moments are preserved provided we use a finite element basis capable of reproducing polynomials of order up to $m - 1$. An order m tensor product B-spline basis will do this nicely. It may be the case that a particular region Ω has been chosen with the knowledge that the support of the true density is strictly contained in that region. It would then make sense to enforce Dirichlet boundary conditions. This will introduce potential errors into the resulting moments, but only to the extent that polynomials may not be approximated at the boundary.

These errors will be $0(h)$. If we examine the error estimate, we know by assumption that the fundamental solution is in H^m and we can therefore obtain the standard finite element error estimates: $||u - u_h||_{H^s(\Omega)} \preceq h^{m-s}|u|_{H^m(\Omega)}$. In particular, the Céa lemma guarantees that the error will be equivalent to a polynomial spline interpolation. This is significantly better than the error associated with the regression spline and the true density and we will return to this in §7.1.3.

2.4.2 Alternative Approximations for Histosplines

We can equally consider a finite element approximation to the variational equations in (2.3.2). Once again this allows us to save some effort involved in calculating fundamental solutions and we also expect to obtain a computational saving. Taking a finite element basis as above we derive the matrix equations:

$$\begin{bmatrix} C & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \omega \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{s} \end{bmatrix}$$

where $[B]_{ij} = \int \phi_j(x)\psi_i(x)dx$ (with ψ_i given in terms of the moment-matching constraints), ω is a vector of Lagrange multipliers, C is as above and $\mathbf{s}$ is the vector of local moments s_i which we are matching. This is again for

$$\hat{f_N}(x) = \mathbf{x}^T\phi(x).$$

It is clear from this formulation that in order to obtain a unique solution we require that $k > N$ where N is the number of ψ_i. We therefore do not have the same computational advantage in evaluating the function. Some advantage can be gained using the support properties of the finite elements and this may be sufficient to offset the increased number of elements we have to deal with. It is not likely, however, to reduce the order (in terms of N) of the calculation.

What does make the finite element approximation attractive here is the fact that the linear system we are now trying to solve for is sparse. If we have $k = O(N)$ (and it should not be necessary to provide a particularly fine discretisation to achieve a reasonable error) then we can hope to require $O(k)$ operations to solve the discrete system, but potentially $O(N^3)$ for the exact case. If we wish to verify that this conjecture is correct, we will observe that the error estimate for the finite element approximation may be expressed in exactly the same form (by [BF91][Proposition 2.4]). In other words, we can expect an error estimate of $O(h^{m-s})$ in the $H^s(\Omega)$ norm. Now the histospline will converge at best at a rate of $n^{-\phi_d(m,p)/2}$ (taking the root of the mean integrated squared error - as in an $\mathcal{L}^2$ norm) and an optimally chosen histogram for the problem will have bin widths on the order of $n^{-1/(2m+d-2/p)}$. Choosing a finite element scale on the same order, the convergence achieved is still as good as the MISE.

It is worthwhile noting that it is possible to regard the finite element approximation to the problem as being, in a sense, a generalisation of the orthogonal series estimator. If we replace the standard finite elements with an orthogonal basis, we can consider the method to be as follows:

begin with an already truncated orthogonal series and then choose the tapering parameters according to the penalty function. If, on the other hand, we take $\lambda = 0$ and use non-overlapping piecewise constant elements, we retrieve the histogram. Using a finite element histospline approach, we could equally regard ourselves as placing some particular control on the moments that we are trying to match.

Chapter 3

Specific Estimators

In this chapter we present two specific cases that are of interest in the estimates that we wish to produce. The first of these is simply the density spline as defined above with $m = 1$ and $a_{\alpha\beta} = 1$. It is important as it has a fundamental solution which is always positive and has been labelled the Roberts Estimator after Stephen Roberts' realisation of this fact. The second of these relates to the thin plate spline penalty term; in this section we detail the application of a nonconforming finite element method derived from the techniques used in [HRA97] for estimating the solution to the thin plate spline problem in regression smoothing. It is useful here in removing the need to use a H^2 approximating subspace and allowing us to deal only with H^1 elements. These are considerably easier to work with and evaluate, with smaller matrices containing fewer entries.

3.1 The Roberts Estimator

The estimates that we have produced plainly fail the positivity condition for densities. This is often not a problem for the applications intended and we may well deal with such estimates by using a positive projection as in §1.3.5 if it is. Nevertheless, it may be useful to produce an estimate that is positive and H^1.

Theorem 3.1.1. *The fundamental solution to:*

$$\int_\Omega u(x)s(x)dx + \lambda \int_\Omega (\nabla u(x) \cdot \nabla v(x))dx = \hat{E}(s) \ \forall s \in H^1(\Omega) \tag{3.1}$$

is always positive.

It should be noted that the usual estimate we take of $E(s)$ will not be applicable here due to the fact that H^1 is not a reproducing kernel Hilbert space for dimensions greater than one. Typically, an integral against an initial histogram estimate may be an appropriate, potentially scalable, approximation. The proof of this is a straightforward application of a maximal principle and was developed by Stephen Roberts and Nira Dyn.

Proof. We will begin by reducing the variational form to a differential equation. Integration by parts gives us that

$$\int_\Omega u(x)s(x)dx - \lambda \int_\Omega \Delta u(x)s(x)dx = \int_\Omega \eta(x)s(x)dx,\ \forall s \in H^1(\Omega)$$

with the von Neumann boundary conditions

$$\frac{\partial u}{\partial n}(x) = 0 \text{ on } \partial\Omega$$

and $\eta(x)$ is the $\mathcal{L}^2$ Riesz representor of $\hat{E}(s)$ in Ω. We can now reduce to an Euler equation. u satisfies:

$$u(x) - \lambda\Delta u(x) = \eta(x),\ x \in \Omega \tag{3.2}$$

subject to the boundary conditions above. Any reasonable empirical approximation of $E(s)$ will have a density as its representor. In particular, we can expect η to be nonnegative.

Let us suppose that $u(x) < 0$ in some maximal subdomain $\Omega-$ with boundary $\partial\Omega-$. It is clear that

$$\frac{\partial u}{\partial n}(x) \geq 0 \tag{3.3}$$

for $x \in \partial\Omega-$. We can obtain

$$\int_{\Omega-} \Delta u = \int_{\partial\Omega-} \frac{\partial u}{\partial n} \tag{3.4}$$

by substituting $s = 1$ in the general Greens formula

$$\int_D \nabla u \cdot \nabla s = -\int_D s\Delta u + \int_{\partial D} s\frac{\partial u}{\partial s}.$$

Using (3.3) we see that:

$$\int_{\Omega-} \Delta u \geq 0.$$

On the other hand, making use of (3.2) and the non-negativity of η, we have:

$$\int_{\Omega-} \Delta u \leq \frac{1}{\lambda}\int_{\Omega-} u - \eta < 0$$

by assumption, giving us a contradiction. $\square$

There are clearly drawbacks to the use of this estimator - the resulting functions are at best H^1 and we have produced no control over moments. Placing constraints on the estimator will, naturally, invalidate the proof above. In fact, we find that the finite element approximation is also not guaranteed to have positivity and empirical results indicate that relatively fine meshes are necessary to produce it. An alternative approximation would be to produce a finite difference approximation to (3.2).

If this approach is taken, we have that on a uniform grid with the usual node numbering - using $m_1, \ldots, m_d$ nodes in dimensions $1, \ldots, d$ respectively - the following equation obtains in the interior:

$$\left(1 + \sum_{k=1}^{d} \frac{2}{h_k}\right) u(x_i) - \sum_{i=1}^{d} \frac{1}{h_k} \left(u(x_{i-1_k}) + u(x_{i+1_k})\right) = \eta(x_i)$$

where $i = (i_1, \ldots, i_k)$ and 1_k is the k^{th} basis vector in this space. Now, suppose that

$$u(x_i) \leq u(x_{i \pm 1_k}), \; k = 1, \ldots, d.$$

In other words, u represents a local minimum. From this we have that:

$$u(x_i) = \eta(x_i) + \sum_{k=1}^{d} \frac{1}{h_k} \left(u(x_{i-1_k}) - 2u(x_i) + u(x_{i+1_k})\right) > 0$$

and in fact u must be positive in the interior. Enforcing the von Neumann boundary conditions, we have that at a boundary hyperplane perpendicular to the k^{th} co-ordinate direction

$$u(x_i) = u(x_{i-1_k}) > 0.$$

Finally, at the corners we will merely stipulate that the approximation takes on the average value of the surrounding nodes and the estimator remains positive. Figure 3.1 provides an illustration of the difference between a finite element and a finite difference approximation. Although the finite difference approximation appears to be better, we would in fact expect to achieve the same rate of convergence.

3.2 A Nonconforming Method for the Thin Plate Density Spline

In this section we will be concerned with what might be termed the thin plate spline density estimator given in the bivariate case by the usual equation with penalty term:

$$C(u, s) = \int_{\Omega} (u_{x_1,x_1} s_{x_1,x_1} + 2u_{x_1,x_2} s_{x_1,x_2} + u_{x_2,x_2} s_{x_2,x_2}) dx$$

which will preserve a unit integrand and first-order moments. This choice of penalty term is intuitively pleasing - effectively we are minimising curvature - corresponding to our visual intuition about “roughness”, and has been used extensively in spline smoothing techniques.

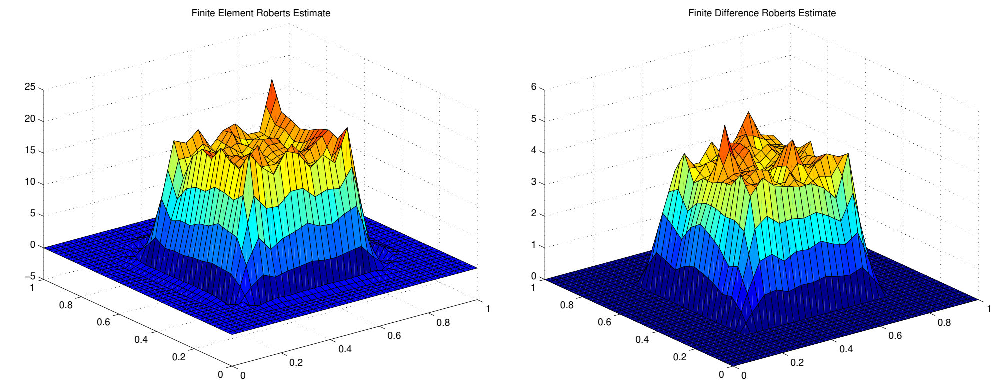


Figure 3.1: Finite element and finite difference Roberts estimators for a uniform distribution using 10,000 data points with a mesh size of 0.05 and $\lambda = 10^{-12}$.

In dealing with this penalty term we clearly require H^2 elements in a finite element approximation. These can be reasonably easily produced in two dimensions, but become steadily more difficult to create for higher-dimensional cases. One simple alternative is to take tensor products of univariate H^2 elements - cubic B-splines, say, for higher dimensions. These are simple to work with but have the disadvantage that they significantly increase the number of diagonals present in the matrices that must be solved, adding to the computational cost. Indeed, any d-dimensional H^2 set of elements will have on the order of 4^d diagonals in the matrices resulting from it. The increased number of elements which overlap the boundary will also require a larger system to be solved for the same finite element mesh. Experimentally, it was found that evaluating H^2 elements to determine their empirical moments substantially increased the cost of performing the algorithm. What we present here is a method to reduce the complexity of these elements to H^1, although we retain better behaviour in the over-all estimator. In so doing, we also reduce the density of the linear systems we need to calculate. The method below can be regarded as an adaptation of the finite element thin plate splines developed in [HRA97] to the density spline functionals listed above. The proofs of existence and uniqueness below follow those found in [RHA99].

In deriving this nonconforming method, we will make use of the following motivation. Let $\mathbf{u} = (u_1, u_2) = \nabla u$. Then we can express

$$C(u, s) = \int_\Omega (\nabla u_1(x) \cdot \nabla s_1(x) + \nabla u_2(x) \cdot \nabla s_2(x))dx = A(\mathbf{u}, \mathbf{s}).$$

Now we only require that $\mathbf{u} \in H^1$ and we would like to use this to reduce to using H^1 elements generally. We do this as follows: let $P\mathbf{u}$ be the solution to:

$$\begin{gathered}(\nabla v, \nabla P\mathbf{u}) = (\nabla v, \mathbf{u}), \ \forall v \in H^1(\Omega) \\ \textstyle\int_\Omega P\mathbf{u} = 0\end{gathered} \tag{3.5}$$

where $(\cdot, \cdot)$ denotes the usual $\mathcal{L}^2$ inner product. Alternatively, we can express this as:

$$\Delta P\mathbf{u} = \nabla \cdot \mathbf{u} \text{ in } \mathcal{L}^2(\Omega) \tag{3.6}$$
$$\frac{\partial P\mathbf{u}}{\partial \mathbf{n}} = \mathbf{u} \cdot \mathbf{n} \text{ in } H^{1/2}(\partial\Omega). \tag{3.7}$$

If in addition we know that curl$\mathbf{u} = 0$ then $\mathbf{u} = \nabla P\mathbf{u}$. We will ignore this condition here, but will return it later.

The requirement that $\int_\Omega P\mathbf{u} = 0$ is necessary to ensure that $P\mathbf{u}$ is well defined in (3.5). In fact, we will use $P\mathbf{u}$ as our density estimate but it will turn out to be computationally easier to require zero integrand and add on an appropriate constant afterwards. We will now re-express our variational equations (2.6) as:

$$(P\mathbf{u}, P\mathbf{s}) + \lambda((u_1, s_1)_{H^1(\Omega)} + (u_2, s_2)_{H^1(\Omega)}) = \frac{1}{n}\sum_{i=1}^{n} P\mathbf{s}(X_i), \ \forall \mathbf{s} \in H^1(\Omega)^2. \tag{3.8}$$

Clearly, enforcing the condition curl$\mathbf{u} = 0$ will give us the standard thin plate spline penalty term.

Theorem 3.2.1. *There is a unique solution* $\mathbf{u}$ *to equation (3.8).*

It is necessary here only to demonstrate the continuity of the functional defined by the linear form

$$F(\mathbf{s}) = \frac{1}{n}\sum_{i=1}^{n} P\mathbf{s}(X_i)$$

and the ellipticity in $H^1(\Omega)^2$ of the norm defined by the bilinear form:

$$c(\mathbf{u}, \mathbf{s}) = (P\mathbf{u}, P\mathbf{s}) + \lambda((u_1, s_1)^2_{H^1(\Omega)} + (u_2, s_2)^2_{H^1(\Omega)})$$

where $|u|_{H^1(\Omega)} = \int_\Omega (\nabla u(x))^2 dx$ is the standard semi-norm with associated inner product $(\cdot,\cdot)_{H^1(\Omega)}$. While doing this, we will also consider the discrete versions of these functionals, F_h and c_h defined on a finite element space $V_h \subset H^1(\Omega)$ and V_h^2 respectively by substituting $P_h\mathbf{u}$ for $P\mathbf{u}$ where $P_h\mathbf{u}$ solves the analogue of (3.5) in V_h. This proof shall be carried out in the following propositions.

Proposition 3.2.1. *The bilinear form* $c(\cdot,\cdot)$*, the linear form* $F(\cdot)$ *and their corresponding discrete forms,* c_h *and* F_h*, are continuous on* $H^1(\Omega)^2$*.*

Proof. This result rests on the fact that we can assume that $P\mathbf{u} \in H^2(\Omega)$ implicitly. Recall equation (3.6): $\mathbf{u} \in H^1(\Omega)^2$ and so $\nabla \cdot \mathbf{u} \in \mathcal{L}^2(\Omega)$ and $\mathbf{u} \cdot \mathbf{n} \in H^{1/2}(\partial\Omega)$. We now have regularity theorems that will give us [Gri85]:

$$|P\mathbf{u}|_{H^1(\Omega)} + |P\mathbf{u}|_{H^2(\Omega)} \preceq ||\mathbf{u}||_{H^1(\Omega)^2} \tag{3.9}$$

When we restrict $P\mathbf{u}$ to having zero mean, a Poincaré inequality holds from [GT77, Lemma 7.16]:

$$||P\mathbf{u}||_{\mathcal{L}^2(\Omega)} \preceq |P\mathbf{u}|_{H^1(\Omega)}$$

and so we have that $a(\mathbf{u},\mathbf{u})$ is continuous from (3.9). This also give us

$$||P\mathbf{u}||_{H^2(\Omega)} \preceq ||\mathbf{u}||_{H^1(\Omega)^2} \tag{3.10}$$

and in particular the Sobolev embedding theorem implies that pointwise evaluation is bounded and therefore that $F(\mathbf{u})$ is also continuous.

We need to demonstrate the truth of this claim for an H^1 finite element approximation $P_h\mathbf{u}$. We cannot assume that $P_h\mathbf{u} \in H^2(\Omega)$ but we will assume that this is true of the restriction of $P_h\mathbf{u}$ to each subdomain K in our discretisation T_h. In this case we can conclude that

$$||P_h\mathbf{u}||_n^2 \preceq \sum_{K\in T_h} ||P_h\mathbf{u}||_{H^2(K)}^2 \preceq ||P\mathbf{u}||_{H^2(\Omega)}^2$$

where $||\cdot||_n$ is the pointwise norm at X_i. This will give us both the boundedness of $||P_h\mathbf{u}||_{H^2(\Omega)}$, which is enough to prove the continuity of c_h, and of pointwise evaluation of $P_h\mathbf{u}$. In particular,

$$F_h(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} P_h\mathbf{u}(X_i)$$

will be bounded. $\square$

It remains to show that c is elliptic in the $H^1(\Omega)^2$ norm.

Proposition 3.2.2. *$c(\cdot,\cdot)$ and $c_h(\cdot,\cdot)$ are $H^1(\Omega)^2$-elliptic.*

Proof. We will begin by denoting $a_1 = \int_\Omega u_1 dx$, $a_2 = \int_\Omega u_2 dx$ and observing that the Poincaré inequality give us that:

$$||\mathbf{u}||_{H^1(\Omega)^2}^2 \preceq (a_1^2 + a_2^2) + |\mathbf{u}|_{H^1(\Omega)^2}^2.$$

This can be seen by observing that $\tilde{u_i} = u_i - c_i$ has zero mean, substituting the expression in the $\mathcal{L}^2$ norm and using Poincaré. We now have that:

$$a_1^2 + a_2^2 \preceq [a_1\ a_2]\, X\, [a_1\ a_2]^T = ||a_1x_1 + a_2x_2||_{\mathcal{L}^2(\Omega)}^2$$

where $[X]_{ij} = \int_\Omega x_i x_j dx_i dx_j$ is a positive definite matrix. Now $P[a_1\ a_2] = P_h[a_1\ a_2] = a_1x_1 + a_2x_2$. Hence:

$$||\mathbf{u}||_{H^1(\Omega)^2}^2 \preceq \left(||P[a_1\ a_2]||_{\mathcal{L}^2(\Omega)}^2 + |\mathbf{u}|_{H^1(\Omega)^2}^2 \right)$$

Finally we note that:

$$\begin{aligned} ||P[a_1\ a_2]||^2_{\mathcal{L}^2(\Omega)} &\leq ||P\mathbf{u}||^2_{\mathcal{L}^2(\Omega)} + ||P(\mathbf{u} - [a_1\ a_2])||^2_{\mathcal{L}^2(\Omega)} \\ &\leq ||P\mathbf{u}||^2_{\mathcal{L}^2(\Omega)} + ||\mathbf{u} - [a_1\ a_2]||^2_{H^1(\Omega)^2} \\ &\leq ||P\mathbf{u}||^2_{\mathcal{L}^2(\Omega)} + |\mathbf{u}|^2_{H^1(\Omega)^2}. \end{aligned}$$

Consequently:

$$||\mathbf{u}||^2_{H^1(\Omega)^2} \preceq \left(||P\mathbf{u}||^2_{\mathcal{L}^2(\Omega)} + |\mathbf{u}|^2_{H^1(\Omega)^2}\right) \preceq a(\mathbf{u},\mathbf{u})$$

and the k_i depend only on Ω and λ. It follows that there is a unique solution to this problem and that this result will hold in a nonconforming H^1 approximation. □

3.2.1 Finite Element Calculations

Applying a finite element approximation on a basis $\{\phi_i\}_{i=1}^k$, the problem (3.5) may be approximated as a discrete problem in H^1 by:

$$C\mathbf{u}_0^h = B_1\mathbf{u}_1^h + B_2\mathbf{u}_2^h$$

where $[C]_{ij} = (\nabla\phi_i, \nabla\phi_j)$, $[B_k]_{ij} = (\partial\phi_i/\partial x_k, \phi_j)$ and $P_h\mathbf{u} = \sum_{i=1}^k [\mathbf{u}_0^h]_i\phi_i$. Writing

$$\vec{C} = \begin{bmatrix} C & 0 \\ 0 & C \end{bmatrix},\ B = [B_1\ B_2],\ \mathbf{u}^h = \begin{bmatrix} \mathbf{u}_1^h \\ \mathbf{u}_2^h \end{bmatrix}.$$

We can express the discretised (3.8) in matrix form as:

$$(B^TC^{-1}AC^{-1}B + \lambda\vec{C})\mathbf{u}^h = B^TC^{-1}\mathbf{n}$$

with $[A]_{ij} = (\phi_i, \phi_j)$, $[\mathbf{n}]_j = \sum_{i=1}^n \phi_j(X_i)$. This will now be a $2k$ set of linear equations which will no longer be sparse. It can, however, be solved implicitly by iterative methods requiring only multiplication by sparse matrices. In practise, it has been found that the matrix $B^TC^{-1}AC^{-1}B+\lambda\vec{C}$ is poorly conditioned, leading to very slow convergence of iterative methods. A good preconditioner has yet to be developed.

3.2.2 Zero Curl Condition

We presently address the issues involved in leaving out the condition that $\text{curl}\mathbf{u} = 0$. In [RHA99] it can be shown that the solution to the spline problem using the functional c_h will converge to the exact solution when the response variable has been measured exactly. Such theorems are not yet known for the nonconforming density spline we have just outlined. However, we will provide some exploration of the consequences of the neglect of the zero curl condition with the following numerical experiments.

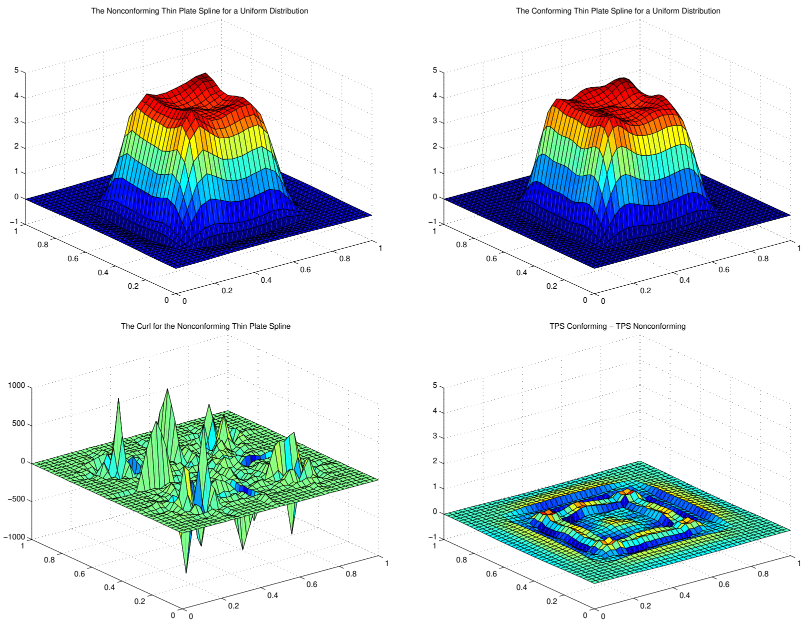


Figure 3.2: Nonconforming Thin Plate Spline (top left), Conforming Thin Plate Spline (top right), curl**u** for the Nonconforming Thin Plate Spline (bottom left) and the Conforming-Nonconforming difference (bottom right), all plotted for a uniform distribution.

These experiments will use essentially the same format on different data structures. We will use three densities: a uniform distribution in the box $[0.25, 0.75] \times [0.25, 0.75]$, a normal distribution centred on $(0, 0)$ and a multi-modal distribution using three displaced normal distributions with differing weights. Both Figure 3.3 and Figure 3.4 indicate that, over all, the curl will remain small and the difference between conforming and nonconforming solutions to the problem will be similarly small. In contrast to the findings in [HRA97] the discrepancy appears to occur, for the most part, away from the edges of the regions in the areas where the density is larger. Similarly, the curl becomes non-zero in the region of data points contrasting with the opposite case found for the regression problem. Figure 3.2 suggests that the curl may become significantly non-zero particularly for densities which have discontinuities. Further experiments suggest that small values of λ tend to lead to the curl becoming large. Nevertheless, the small difference between the densities even in this case suggests that this behaviour is not necessarily problematic.

3.2.3 Finite Element Convergence

The method we present here is nonconforming and as such the convergence of the finite element method must therefore be checked. We will assume that we are working with a function space V_h with a projection $Q_h : H^1(\Omega) \to V_h$ such that

$$||v - Q_h v||^2_{H^k(\Omega)} \leq h^{2(1-k)} |v|^2_{H^1(\Omega)}$$

for $k = 0, 1$. For $\mathbf{u}$ solving (3.8) and $\mathbf{u}_h$ its discrete analogue, the densities we produce will be

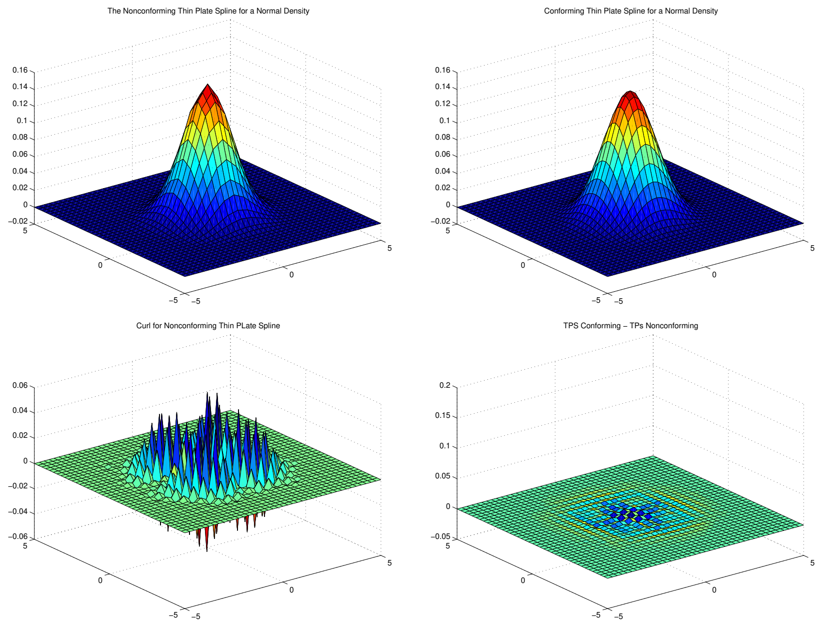


Figure 3.3: Nonconforming Thin Plate Spline (top left), Conforming Thin Plate Spline (top right), curl**u** for the Nonconforming Thin Plate Spline (bottom left) and the Conforming-Nonconforming difference (bottom right), all plotted for a normal distribution.

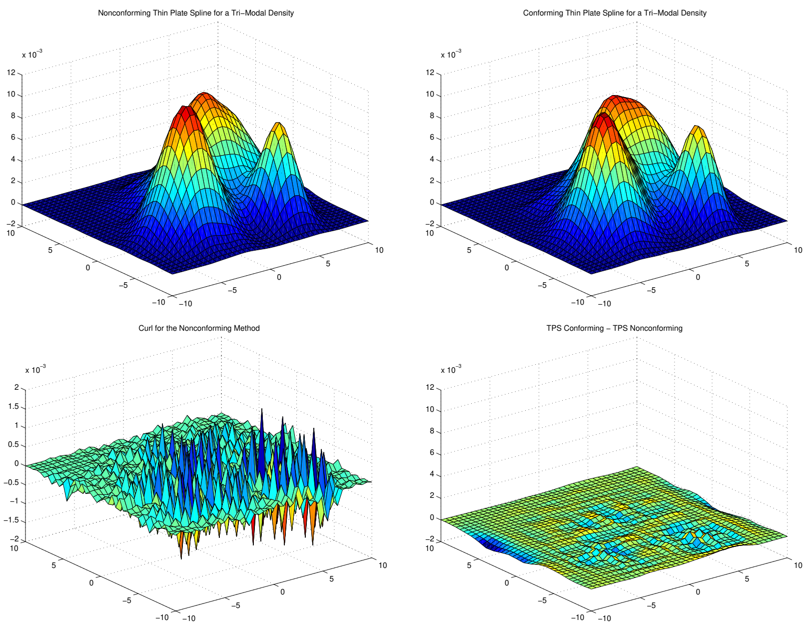


Figure 3.4: Nonconforming Thin Plate Spline (top left), Conforming Thin Plate Spline (top right), curl**u** for the Nonconforming Thin Plate Spline (bottom left), and the Conforming-Nonconforming difference (bottom right), all plotted for a multimodal distribution.

given by $P\mathbf{u}$ and $P_h\mathbf{u}_h$, with the addition of a common constant, respectively. In the context of the finite element method, the error we are concerned with is:

$$||P_h\mathbf{u}_h - P\mathbf{u}||^2_{H^k(\Omega)} \leq ||P_h(\mathbf{u}_h - \mathbf{u})||^2_{H^k(\Omega)} + ||P_h\mathbf{u} - P\mathbf{u}||^2_{H^k(\Omega)}.$$

Now, we know that P_h is a bounded operator on $H^1(\Omega)^2$ and therefore that

$$\begin{aligned} ||P_h(\mathbf{u}_h - \mathbf{u})||^2_{H^k(\Omega)} &\preceq ||\mathbf{u}_h - \mathbf{u}||^2_{H^k(\Omega)} \\ &\preceq h^{2(1-k)}|\mathbf{u}|^2_{H^k(\Omega)^2} \end{aligned} \tag{3.11}$$

by the usual use of the Céa Lemma [Cia75][pp. 29]. Now we also have that (3.5) represents an elliptic system for any given $\mathbf{u}$ and we can again employ the Céa Lemma to obtain:

$$||P_h\mathbf{u} - P\mathbf{u}||^2_{H^k(\Omega)} \preceq h^{2(1-k)}|P\mathbf{u}|^2_{H^k(\Omega)}. \tag{3.12}$$

Combining (3.11) and (3.12) will give us

$$||P_h(\mathbf{u}_h - \mathbf{u})||^2_{H^k(\Omega)} \preceq h^{2(1-k)}\left(|P\mathbf{u}|^2_{H^k(\Omega)} + |\mathbf{u}|^2_{H^k(\Omega)^2}\right)$$

and the case $k = 0$

$$\begin{aligned} ||P_h(\mathbf{u}_h - \mathbf{u})||^2_{\mathcal{L}^2(\Omega)} &\preceq h^2\left(|P\mathbf{u}|^2_{\mathcal{L}^2(\Omega)} + |\mathbf{u}|^2_{\mathcal{L}^2(\Omega)^2}\right) \\ &\approx h^2||P\mathbf{u}||^2_{H^1(\Omega)}. \end{aligned}$$

We note here that first order convergence, which is essentially what we achieve, is a fairly weak result. It is likely that this can be extended to second order, but as yet there is no proof of such a result.

3.2.4 Generalisations

To what extent is the method used here generalisable? Clearly, the proof we provide here will also work in higher dimensional cases. This will not get us very far, however, as the Sobolev embedding theorem will prevent us from using this method in greater than three dimensions. As with the Roberts estimator, replacing pointwise evaluation with integration against a histogram would allow us to continue to higher dimensions. It may be preferable to get an analogous method for higher-order penalty terms.

The general m^{th} order bivariate thin plate spline penalty term in variational form is given in [Wah90] as:

$$C_m(u, s) = \int_{\mathbb{R}^2}\left(\sum_{\nu=1}^{m}\frac{\partial^m u(x)}{\partial x_1^\nu \partial x_2^{m-\nu}}\frac{\partial^m s(x)}{\partial x_1^\nu \partial x_2^{m-\nu}}\right)dx. \tag{3.13}$$

It can be seen that if we let $\nabla u = (u_1, u_2)$, $\nabla x = (s_1, s_2)$ that

$$C_m(u, s) = C_{m-1}(u_1, s_1) + C_{m-2}(u_2, s_2)$$

and we can employ the same techniques as we have for the ordinary thin plate spline to reduce by one the order of the Sobolev space in which we need to find basis functions. This is, of course, ignoring the zero curl condition. It would be nice at this point to be able to bootstrap this idea and reduce an m^{th} order penalty term to requiring $H^1(\Omega)$ functions. In fact, we cannot do this directly; $P\mathbf{u}$ has been required to have zero mean. We will develop the method as follows, define $P^i\mathbf{u}$ recursively by:

$$\left(\nabla P^i\mathbf{u}, \nabla\mathbf{s}\right) = \left(P^{i+1}\mathbf{u}, \nabla\mathbf{s}\right), \ \forall\mathbf{s} \in H^1(\Omega)^{2(i-1)} \tag{3.14}$$

starting with the vector $\mathbf{u} \in H^1(\Omega)^{2(m-1)}$. Essentially, we are continually taking gradients of each term. The problem now arises that (3.14) does not adequately define $P\mathbf{u}^i$. This did not present a problem for the standard thin plate spline functionals as we had the final requirement $\int_\Omega P\mathbf{u} = 0$ and added a constant. In the generalised setting, however, we will need to compensate for this by adding in an element q of the null space of $|\cdot|_{H^{m-1}(\Omega)}$. This is equivalent to stating that we have $\mathbf{u}^i = \mathbf{v}^i + \mathbf{c}^i$ where $\mathbf{v}$ is the solution to (3.14) with zero mean enforced and $\mathbf{c}^i$ a constant vector. Alternatively, $P[p(x, y), q(x, y)] = xp(x, y) + yq(x, y) - K$ where K enforces a zero mean. We then get a variational definition of the generalised nonconforming thin plate spline:

$$\left(P^{m-1}\mathbf{u} + p, P^{m-1}\mathbf{s} + q\right) + \lambda(\mathbf{u}, \mathbf{s})_{H^1(\Omega)^{2(m-1)}} = \frac{1}{n}\sum_{i=1}^{n}(P^{m-1}\mathbf{s} + q)(X_i), \ \forall\mathbf{s} \in H^1(\Omega)^{2(m-1)}.$$

Naturally, this can also be generalised to higher dimensional splines. In $\mathbb{R}^d$ the generalised thin plate spline term may be given as [Wah90]:

$$\int_\Omega \left(\sum_{|\alpha|=m} \frac{m!}{\alpha_1! \dots \alpha_d!} \frac{\partial^m u(x)}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \frac{\partial^m s(x)}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \right) dx$$

and in fact the generalised nonconforming method presented above can be used in these functionals. We note that thin plate spline functionals are uniquely suited to this bootstrapping method: it relies explicitly on the generalised binomial coefficients contained in the penalty term. Generalisations of other penalty terms may be given by this type of procedure, but will not maintain the same form in passing to higher-order derivatives. We notice here that the same moment preserving conditions will apply here as in other splines (this has not yet been noted for the standard nonconforming method above) by remarking that $P^{m-1}\mathbf{s}$ in the null space of C_m will be given by $\mathbf{s} = \textit{constant}$ and therefore $(\mathbf{u}, \mathbf{s})_{H^1(\Omega)^{2(m-1)}} = 0$.

Despite the promising nature of this bootstrap, it is not necessarily computationally efficient. We have replaced one system of equations here with d^{m-1} systems which we must solve. The elements we need to use, of course, are now H^1 and the matrices will turn out to have 3^d diagonals. In comparison, matrices produced using a tensor product of H^m elements will typically have $(4m-1)^d$ elements. This gives a crude approximation to the cost of solving the system. Typically, we will have $2m \geq d$ to conform with the Sobolev embedding theorem. The advantage that this would give to the nonconforming method, however, will be tempered by the fact that the relationship between $P^{m-1}\mathbf{u}$ and $\mathbf{u}$ is governed by $\sum_{i=1}^{m-1} 2^{i-1}$ inverse Laplacian matrices. It may also be the case that ignoring the zero curl condition will increasingly distort the splines as they are iterated.

Chapter 4

Choosing Parameters

All δ-sequence estimators rely on at least one parameter which governs the amount of smoothing which the estimator makes and produces a δ-sequence in an asymptotic limit. In the case of kernels and density splines this parameter is given explicitly as h and λ respectively; it is given more implicitly in orthogonal series as the order of the truncation of the series, and in histograms (and associated smoothers such as histosplines) in the size of the histogram bins. This parameter will govern the trade-off between the need to represent the behaviour of the actual data that we are given and the smoothness that we assume the underlying density possesses. We can think of this as trading off the opposing tendencies to become either a δ-sequence or a constant function. As such it would be useful to come up with some criteria by which we could choose this parameter. For low-dimensional data sets which are being investigated manually, it may be sufficient to choose smoothing parameters visually - whichever parameter gives the best-looking results. However, when a density estimate is produced as part of some automatic process, or when we are dealing with higher-dimensional data (in practise it is difficult to coherently visualise densities in more than two or three dimensions) it becomes necessary to automate the choice of smoothing parameter. Making such a choice will necessarily add to the computational time taken to produce an estimated density and one of our central concerns in this chapter will be with the scalability of this estimation. We will also require some gauge of the "goodness" of a particular value of the parameter. There are a number of measures available here, but we will take a minimisation of the mean integrated squared error (MISE):

$$E \int_{\Omega} (f(x) - \hat{f}_{n;\lambda}(x))^2 dx \tag{4.1}$$

to be a desirable choice of λ in the case of density splines. Here $\hat{f}_{n;\lambda}$ indicates the explicit dependence of $\hat{f}_n$ on λ,

We outline some standard methods for making parameter choices in the case of density splines. It will also be necessary to consider the problems associated with aggregated data; when data is already aggregated or rounded, or when the Sobolev embedding theorem prevents us from using a pointwise estimation of expectation. In this case, we are faced with two problems; how to create a histogram and how to then choose an appropriate smoothing parameter for the density spline.

In fact, if we have exact data, the latter problem can be reduced to our calculations above, but we will outline a method that can be employed with data that is already aggregated.

4.1 An Initial Approach - Using Test Data

Writing (4.1) as

$$E\int_\Omega \hat{f}_{n;\lambda}(x)^2dx - 2E\int_\Omega \hat{f}_{n;\lambda}(x)f(x)dx + E\int_\Omega f(x)^2dx,$$

it is clear that only the first two terms in this expansion will vary with λ and we therefore need only minimise

$$M(\hat{f}_{n;\lambda}) = E\int_\Omega \hat{f}_{n;\lambda}(x)^2dx - 2E\int_\Omega \hat{f}_{n;\lambda}(x)f(x)dx.$$

Now we can calculate

$$\int_\Omega \hat{f}_{n;\lambda}(x)^2dx$$

directly from the function we have obtained and it is only necessary to find some means of calculating

$$\int_\Omega \hat{f}_{n;\lambda}(x)f(x)dx = E_f(\hat{f}_{n;\lambda}(x)).$$

In this case, taking the approximation to $E_f(\hat{f}_{n;\lambda}(x)$ that we used in formulating a finite element scheme will not be appropriate; we would expect that $M(\hat{f}_{n;\lambda}) \to -\infty$ as $\lambda \to 0$. This is easily seen by observing from (2.6) that

$$\int_\Omega \hat{f}_{n;\lambda}(x)^2dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{n;\lambda}(X_i) = \lambda C(\hat{f}_{n;\lambda},\hat{f}_{n;\lambda}) - \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{n;\lambda}(X_i)$$

and noting that this is minimised for $\lambda = 0$. A simple alternative would be to split the data set into $\{X_{1,i}\}_{i=1}^{n_1}$ and $\{X_{2,i}\}_{i=1}^{n_2}$ and use part of the data in the formation of density splines for each λ and the rest to evaluate $E_f(\hat{f}_{n;\lambda}(x)$ in the MISE functional. We will then want to minimise:

$$\hat{T}(\lambda) = \int_\Omega \hat{f}_{n;\lambda}(x)^2dx - \frac{2}{n_2}\sum_{i=1}^{n_2}\hat{f}_{n;\lambda}(X_{2,i})$$

where $\{X_{2,i}\}_{i=1}^{n_2}$ is our reserved data set. This is common practise in many contexts when a large amount of data is available. If we have a response variable this would be a confirmation of the accuracy of the resulting function at each test data point. The computation is potentially expensive but we are able to make use of the finite structure of the functions we produce. We need only evaluate $\frac{1}{n_1}\sum_{i=1}^{n_1}\phi_j(X_{1,i})$ and $\frac{1}{n_2}\sum_{i=1}^{n_2}\phi_j(X_{2,i})$ for $j = 1, \ldots, k$ once. Here the first subscript in the data points denotes to which sub-dataset it belongs. Let $\mathbf{u}_0$ represent the vector of co-efficients of the ϕ_i in our approximation and

$$\mathbf{b}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\phi_j(X_{1,i}),\ \mathbf{b}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\phi_j(X_{2,i}).$$

Then we can express the MISE function as

$$\hat{T}(\lambda) = \mathbf{u}_0^T M\mathbf{u}_0 - \mathbf{b}_2^T\mathbf{u}_0,$$

which makes the evaluation of each function, including obtaining $\mathbf{u}_0$, an $O(k)$ operation. In particular we would expect that this should be very fast. In general, for $n_1 \approx n_2$ we would expect the evaluation of $E_f(\hat{f}_{n;\lambda}(x))$ to take $O(n^2)$ time for a general kernel method. This therefore represents a significant saving.

4.2 Least-Squares Cross-Validation

A reasonable criticism of this approach is that it wastes information that might otherwise be used in producing the density function, if we already knew λ by some other means. It is also necessary to take care when sampling different sets of data in this case in order to ensure that both the function-building and parameter-choosing samples do contain data with the same distribution. If the records in the data set are ordered with respect to the time they were collected, for example, it may be that the underlying density has changed over time. This may be interesting in itself, but simply dividing the data set half-way through will create a tendency to oversmooth in order to compensate for the resulting differences in the distribution of the two sets of data. We therefore propose an alternative approach, labelled least-squares cross-validation, which is similar to cross-validatory techniques in standard regression problems. Developed in the early 1980's and presented in [Sil86], the idea is that the method should be good at estimating the value at a point when that point is not used in forming the function - a very similar idea to the test data approach. We can therefore choose the parameter that minimises the sum of empirical errors at each data point as it is left out of the function calculation. In the case of density estimation, the idea is not so intuitive, but we can make use of it readily. It is clear that the difficulty lies in the estimation of

$$\int_\Omega \hat{f}_{n;\lambda}(x)f(x)dx = E(\hat{f}_{n;\lambda}(X)). \tag{4.2}$$

We will define $\hat{f}_{-i;\lambda}$ to be the function generated by all the data except X_i with smoothing parameter λ. We can now approximate (4.2) by

$$\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i,\lambda}(X_i) \tag{4.3}$$

and we can see that this has the same expectation as (4.2):

$$\begin{aligned} E\left(\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i;\lambda}(X_i)\right) &= E\hat{f}_{-n;\lambda}(X_n) \\ &= E\int_{\Omega}\hat{f}_{n;\lambda}(x)f(x)dx \end{aligned}$$

because $E_f(\hat{f}_n)$ does not depend on the sample size. We can now create a least-squares cross-validation functional to minimise

$$\hat{M}(\lambda) = \int_{\Omega}\hat{f}_{n;\lambda}(x)^2dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i;\lambda}(X_i).$$

Once again, we clearly require $O(n^2)$ calculations to evaluate the function. We can again make use of the linearity of the finite element approximation to reduce this to $O(k)$. Let us express by y_i the vector $[y_i]_j = \phi_j(X_i)$. Then we can rewrite (4.3) as follows. Let $A(\lambda) = M + \lambda C$ be the defining matrix for the system of equations we are required to solve. Then we can express:

$$\hat{f}_{-i;\lambda}(X_i) = y_i^T A(\lambda)^{-1}\left(\frac{1}{n-1}\sum_{j\neq i} y_j\right)$$

and therefore

$$\frac{1}{n}\sum_{i=1}^{n}\hat{f_{-i,\lambda}}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i^T A(\lambda)^{-1}\left(\frac{1}{n-1}\sum_{j\neq i} y_j\right)\right]$$

Some re-arranging gives:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{f_{-i,\lambda}}(X_i) = \left(\frac{1}{n}\sum_{i=1}^{n}y_i^T\right)A(\lambda)^{-1}\left(\frac{1}{n-1}\sum_{i=1}^{n}y_i^T\right) - \frac{1}{n-1}\sum_{i=1}^{n}y_i^T A(\lambda)^{-1}y_i.$$

Collecting terms, we get:

$$\hat{M}(\lambda) = \mathbf{u}_0^T M\mathbf{u}_0 - \frac{2n}{n-1}(\mathbf{b}^T A(\lambda)^{-1}\mathbf{b} - \mathrm{Tr}A(\lambda)^{-1}Y)$$

where $[Y]_{ij} = \frac{1}{n^2}\sum_{k=1}^{n}\phi_i(X_k)\phi_j(X_k)$. In practise, this can be calculated using $O(k)$ operations and $O(n)$ storage using the identity $Y = N^T N$ with $[N]_{ij} = \phi_i(X_j)$.

4.3 Choosing Bin Widths for Histograms

It is apparent that the Sobolev embedding theorem will require us to take higher-derivative penalty terms as the dimension of the data set increases if we are to continue using

$$\hat{E}(s) = \frac{1}{n}\sum_{i=1}^{n} s(X_i)$$

as an estimate of expectation. This will correspondingly require elements in $H^{d/2}(\Omega)$ to be used in the finite element approximation. As already stated in §3.2 these spaces can become quite complicated and will increasingly add diagonals to the matrices being inverted. A reasonable alternative is to accept the need to aggregate the data into a histogram before continuing to evaluate the load vector. At this point we are faced with the problem of deciding on the size of the grid on which we are forming the histogram.

A number of procedures exist for resolving this problem, falling mainly into two categories. Firstly, there are rules based solely on the number of points, motivated by a comparison to the normal density or an analysis of error. Sturge's rule, often used as a default in the univariate case, is given by

$$k = 1 + \log_2 n$$

where k is the number of histogram bins and n is the size of the data set. This is developed by choosing the number of data points in a theoretical histogram with k equal bins so as to give the best approximation to a normal density. The result then follows by examining the relationship between k and n. See [Sco92] for details.

As an alternative, [Sco92] suggests choosing the bin width to minimise the mean squared error. If we examine the results obtained in §1.3.4 we have that the mean square error is minimised for

$$k_n = nh^d \approx n^{(2-2d/p)/(2+d-2d/p)},$$

which gives us

$$h \approx n^{(2-2d/p)/d(2+d-2d/p)-1/d}.$$

In the usual case $p = \infty$ (assumed in [Sco92]),

$$h \approx n^{-1/(2+d)},$$

which will produce a considerably less smooth estimate than the bin width implied by Sturge's rule: $h \approx (\log_2 n)^{-1}$ for $d = 1$.

It is also relatively cheap to calculate the least-squares cross-validation score for a histogram. If we have equal bins, $\{I_j\}_{j=1}^{k}$ of volume h in which $\{\nu_j\}_{j=1}^{k}$ data points fall then minimising the MISE functional will be equivalent to minimising

$$\frac{1}{n^2 h}\sum_{j=1}^{n}\nu_j^2 - \frac{2}{n(n-1)h}\sum_{j=1}^{n}(\nu_j - 1)\nu_j = \frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h}\sum_{j=1}^{k}\nu_j^2$$

which is readily computable. Some disadvantages do exist, however, in that the MISE function is rough, which makes finding minima difficult. [Sco92] also demonstrates that we face problems when data is repeated or rounded. Nevertheless, we will re-consider this in a discussion of the relationship between the size of the data set to be considered and the finite element mesh that is required (see §7.1.3).

While these considerations may point to a method of first choosing an optimal histogram and then attempting to smooth it, we have not yet examined the relationship between the histogram and the finite element mesh that we wish to employ in smoothing it. Some elementary calculation, using the results of §1.3.4, gives us that the error in the histogram for $h \approx n^{-1/(2+d)}$ will be at least $O(h^2)$. We now note that the convergence of the finite element method in the squared $\mathcal{L}^2$ norm will be $O(h^2)$ at worst. The calculations in §2.2.3 indicate that we can bound the error of the exact density spline by the error of the histogram. This would suggest that we are able to choose a finite element mesh which is more coarse than the histogram. However, the calculation of histogram values is already an $O(k)$ operation. As it takes no more than $O(k)$ after that to compute the finite element density spline with the same mesh, we lose nothing in doing so and make some substantial gains in simplifying the calculation of the load vector. We will return to these ideas in §7.1.3 when we discuss the relationship between finite element meshes and data size.

4.4 Smoothing Parameters for Aggregated Data

Clearly, there are problems present in the cross-validation approach to the choice of smoothing parameter when we are given data, not as a collection of points, but already aggregated; i.e. in a histogram. An obvious approach is to take a histospline of some degree for the area. If we suspect that there is some error attached to the aggregation, however, we may wish to use a density spline to correct for this. In this case we will again need to choose a smoothing parameter λ, but without recourse to pointwise data.

Instead, we will propose a new minimisation functional, used in [DW82] for inexact data. Let Z be the vector giving the volume of the histogram at each bin. If we have a large sample size with a small proportion of data in each bin, we can regard each Z_i as a random variable having independent Poisson distribution with mean and variance μ_i. We can then think of $\epsilon_i = Z_i - \mu_i$ as a vector of independent random variables with zero mean. Let the histogram bins be given as Ω_i, then a reasonable criteria for choosing λ would be to minimise

$$R(\lambda) = E\sum_{i=1}^{N} \theta_i \left(\mu_i - \int_{\Omega_i} \hat{f}_{N;\lambda}(x)dx\right)^2 \tag{4.4}$$

where the θ_i are a series of weights. We can now use the linearity properties of the finite element calculations to see that there is a matrix $A(\lambda)$ such that

$$A(\lambda)Z = \begin{pmatrix} \int_{\Omega_1} \hat{f}_{N;\lambda} \\ \int_{\Omega_2} \hat{f}_{N;\lambda} \\ \vdots \\ \int_{\Omega_1} \hat{f}_{N;\lambda} \end{pmatrix}.$$

Then we can express (4.4) as

$$\begin{aligned} R(\lambda) &= E||D_\theta^{1/2}\left(\mu - A(\lambda)Z\right)||^2 \\ &= ||D_\theta^{1/2}\left[(I - A(\lambda))\mu - A(\lambda)\epsilon\right]||^2 \\ &= ||D_\theta^{1/2}(I - A(\lambda))\mu||^2 + \mathrm{Tr} D_\theta A(\lambda)\Sigma A(\lambda)^T \end{aligned}$$

where $D_\theta = \mathrm{diag}\{\theta_1, \dots, \theta_N\}$ and $\Sigma = \mathrm{diag}\{\sigma^2(Z_1), \dots, \sigma^2(Z_N)\}$. This is not a particularly helpful reformulation, but it allows us to find an unbiased estimator $\hat{R}(\lambda)$ of $R(\lambda)$ which we may employ.

Theorem 4.4.1. *Define*

$$\begin{aligned} \hat{R}(\lambda) = ||D_\theta^{1/2}(I - A(\lambda))Z||^2 &- \mathrm{Tr} D_\theta^{1/2}(I - A(\lambda)\hat{\Sigma}(I - A(\lambda)^T D_\theta^{1/2} \\ &+ \mathrm{Tr} D_\theta^{1/2} A(\lambda)\hat{\Sigma} A(\lambda)^T D_\theta^{1/2} \end{aligned} \tag{4.5}$$

where $\hat{\Sigma} = \{Z_1, \dots, Z_2\}$. *Then* $E\hat{R}(\lambda) = ER(\lambda)$.

Proof. We prove this by observing that

$$E||D_\theta^{1/2}(I - A(\lambda))Z||^2 = ||D_\theta^{1/2}(I - A(\lambda))\mu||^2 + \mathrm{Tr} D_\theta^{1/2}(I - A(\lambda)\hat{\Sigma}(I - A(\lambda)^T D_\theta^{1/2}$$

and that

$$E\hat{\Sigma} = \Sigma$$

and substituting this into (4.5). □

We remark that (4.5) simplifies to

$$\hat{R}(\lambda) = ||D_\theta^{1/2}(I - A(\lambda))Z||^2 + \sum_{i=1}^{N} \theta_i Z_i - 2\mathrm{Tr} D_\theta^{1/2} A(\lambda)\hat{\Sigma} A(\lambda)^T D_\theta^{1/2}$$

which is more readily calculated. This functional can be thought of as a generalisation of the unbiased risk estimate for the spline regression problem given by

$$\hat{T}(\lambda) = \frac{1}{n}||(I - A(\lambda))y||^2 - \frac{\sigma^2}{n}\mathrm{Tr}(I - A(\lambda))^2 + \frac{\sigma^2}{n}\mathrm{Tr} A(\lambda)^2,$$

which may be used for pointwise data to approximate the minimiser of

$$T(\lambda) = \frac{1}{n}||A(\lambda)(y+\epsilon) - y||^2.$$

where y is a response variable whose (known) measurements have independent error with zero mean and known variance σ^2.

These problems will not be encountered when we are forced by the Sobolev inequality to use an integral against a histogram to estimate expectations instead of a pointwise estimate. The existence of pointwise data still allows us to evaluate the resulting functional in a pointwise fashion. In this case the least-squares cross-validation functional becomes a little more complicated.

4.5 Searching for Minima

It is worth commenting on the searching procedures required to produce a value for λ minimising the error functionals, $\hat{T}$, $\hat{M}$ and $\hat{R}$ which we have developed. Empirical studies (see Figures 4.1 through 4.5) show that the value of λ has a threshold close to the minimum but does not significantly alter the error or the visual appearance of the resulting functional away from that threshold. It is therefore reasonable to search for a minimum on a fairly discrete set of values. In particular, searching along orders of magnitude is appropriate. Similarly, the graph of the MISE function against λ is close to monotone, and does not appear to contain more than one local minimum. It is therefore reasonable to terminate the search once a local minimum has been reached. This will significantly speed up the automated choice of an optimal λ.

Running a few experiments for each of the generalised histospline, nonconforming and conforming thin plate spline estimators and the Roberts estimator achieved similar results to Figure 4.1. Fairly clearly, both visually and in terms of the cross-validation score, there is a dramatic change in the fit of the function around $\lambda = 10^{-3}$ and considerably less effect away from that value. In fact, the cross-validation score will choose a value for λ smaller than this, largely due to the monotonicity of the graph.

The same experiment may be repeated with other estimates. We will here provide three examples of a similar experiment to that above. Figure 4.2 uses a least-squares cross-validation when no aggregation has been applied. Figure 4.3 demonstrates the Roberts estimator using (4.5) as a method of choosing the parameter with aggregated data and Figure 4.4 presents a conforming thin plate spline method using test data.

Clearly, Figure 4.3 indicates that $\hat{R}(\lambda)$ will tend to undersmooth somewhat. In that sense it may be undesirable. However this may well not be the case for a more coarse histogram mesh. Nevertheless, it seems to be reasonable to reject this principle as a general grounds for parameter choice in the cases we are interested in here.

Finally, some comparison of the choices made by these estimators would be worthwhile. Figure 4.5 presents a comparison of the graphs of $\hat{T}$, $\hat{R}$ and $\hat{M}$ for the Roberts estimator using a normal distribution with 10,000 data points. These plots indicate that all three methods provide reasonably similar patterns for their estimates. We do face a problem here, however, in that below

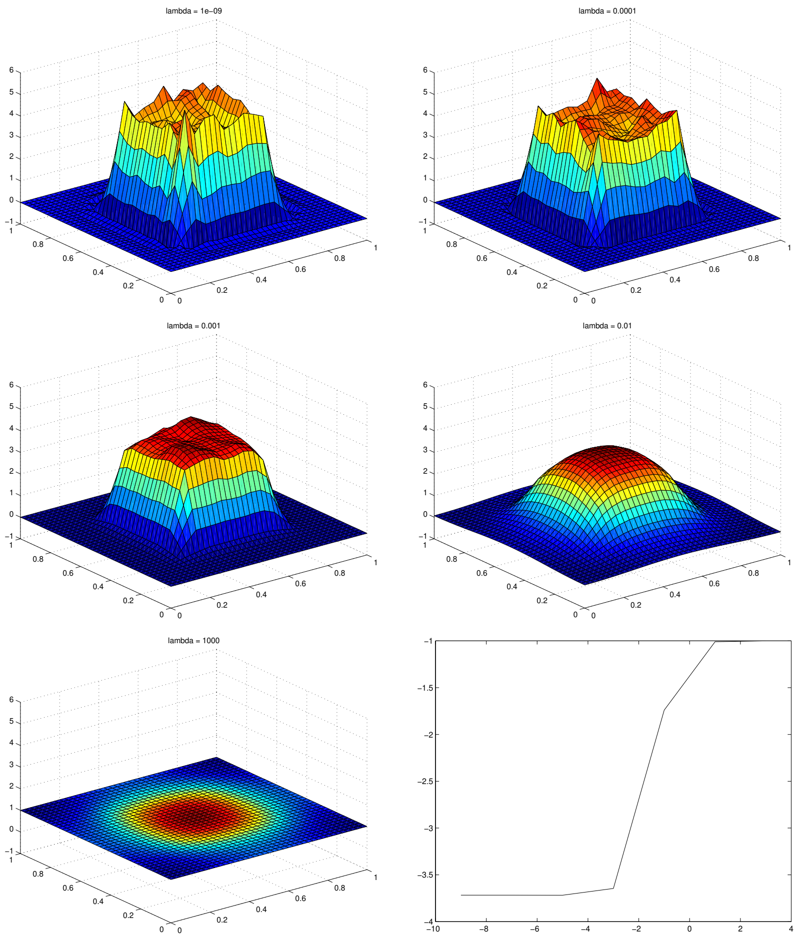


Figure 4.1: Resulting density functions for various values of λ using the finite element Roberts estimator for a 10,000-point data set with a uniform distribution on $[0.25, 0.75] \times [0.25, 0.75]$ and the graph of the values of the least-squares cross-validation functional against $\log \lambda$ with $\lambda \in [10^{-9}, 10^{3}]$

the critical value of $\lambda = 10^{-2}$, the functions are essentially flat. This will lead to a parameter choice considerably smaller than intended. An alternative search that takes account of this feature would provide a reasonable solution to this problem. The theoretical implications of this behaviour are not clear, but we have seen that smaller values of λ in fact have relatively little impact on the resulting estimator and it is reasonable to suspect that below the critical value the choice of λ is not particularly important.

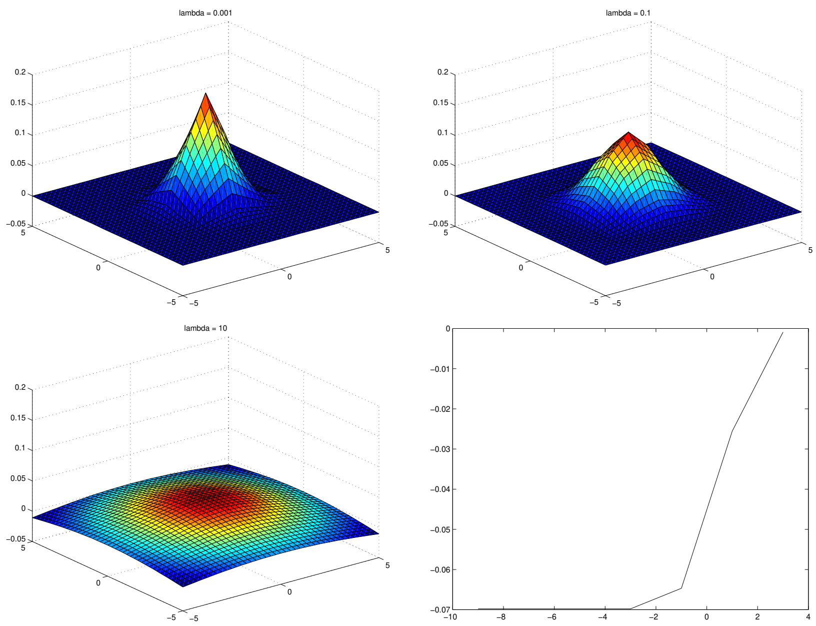


Figure 4.2: Resulting density functions for various values of λ using the nonconforming thin plate spline estimate for a 10,000-point data set with a normal distribution centred on zero and the graph of the values of the least-squares cross-validation-functional against $\log \lambda$ with $\lambda \in [10^{-9}, 10^{3}]$. The plots shown are for values of λ near the critical point.

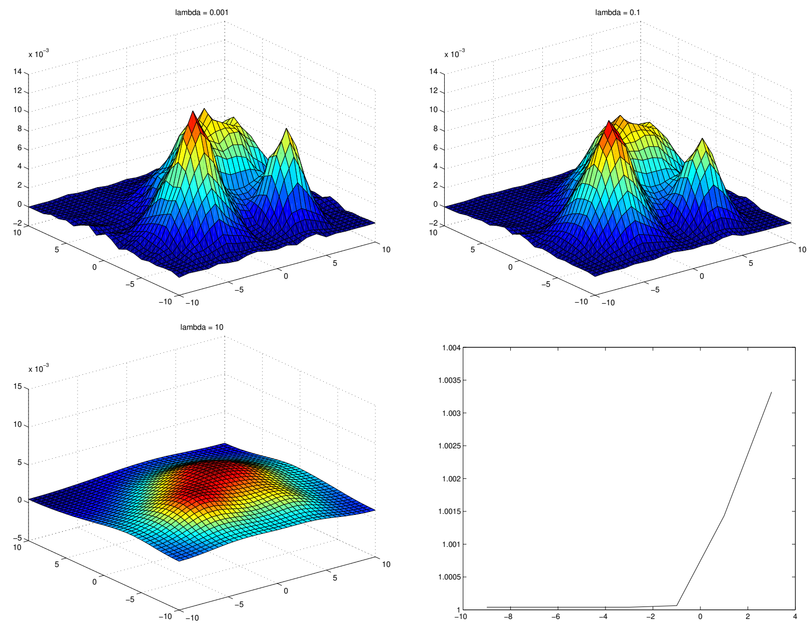


Figure 4.3: Resulting density functions for various values of λ using the Roberts estimator for a 10,000-point data set with a tri-modal distribution and the graph of the values of the functional $\hat{R}(\lambda)$ against $\log \lambda$ with $\lambda \in [10^{-9}, 10^{3}]$. The plots shown are for values of λ near the critical point.

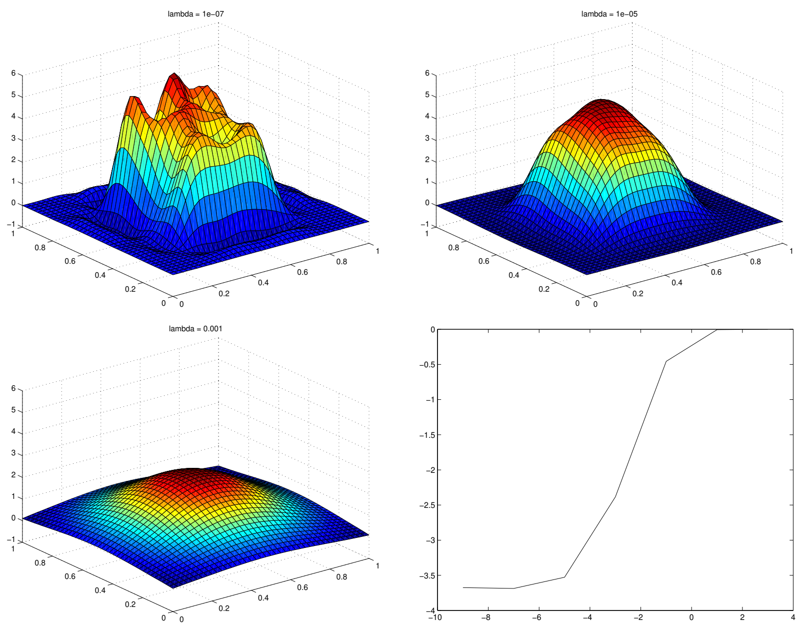


Figure 4.4: Resulting density functions for various values of λ using the conforming thin plate spline estimate using 10,000 initial and 10,000 training data points with a uniform distribution on $[0.25, 0.75] \times [0.25, 0.75]$ and the graph of the values of the functional $\hat{T}(\lambda)$ against $\log \lambda$ with $\lambda \in [10^{-9}, 10^{3}]$. The plots shown are for values of λ near the critical point.

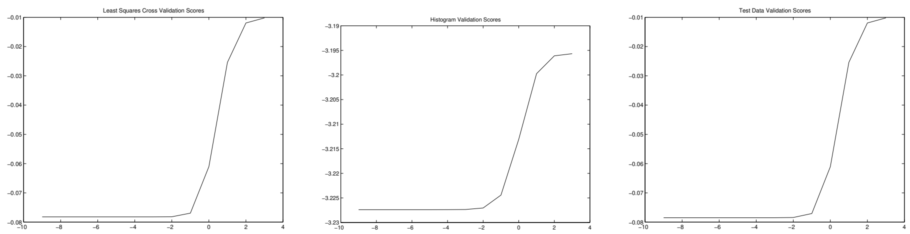


Figure 4.5: Comparative scores for parameter choice in the Roberts estimator using 10,000 data points with a normal distribution. Values plotted against $\log(\lambda)$, from left to right $\hat{M}(\lambda)$, $\hat{R}(\lambda)$ and $\hat{T}(\lambda)$.

Chapter 5

Numerical Experiments

In this chapter we present the results of some numerical experiments using the estimators developed in Chapter 2. The aims here are to verify that the theoretical results we have been able to achieve are borne out in practise and to provide some support for the conjectures that we have been unable to prove so far. Most of the demonstrations will be visual due to the difficulty of producing reliable quantitative results. We will examine the behaviour of the univariate and bivariate nonconforming method, Roberts estimator and histospline visually, in experiments which continue on from those in Chapters 3 and 4. We will also produce some brief Monte Carlo studies concerning the rates of convergence of these estimates. One of the applications of densities foreshadowed in §1.1.1 was in classifying points belonging to one of two populations and it will be worth examining the success of this method in some simulated examples.

5.1 Preliminary Studies

What we will do here is focus on simulations similar to those already undertaken in exploring the zero curl condition in §3.2.2 and searching for minima in Chapter 4. We will compare these estimates on uniform and normal distributions, considering the effect of finite element meshes, histogram meshes and smoothing parameters on the resulting densities. We will start off with univariate densities and proceed to examine the behaviour of densities in the bivariate case.

5.1.1 Univariate Densities

We will begin by investigating the behaviour of density splines and histosplines in the univariate case. Density estimates have traditionally been developed in the univariate case. However, we are more interested in producing multivariate densities and will therefore treat this exploration in a fairly cursory manner. The pictures in Figure 5.1 have been produced using a data set which is evenly distributed between a uniform density on $[0, 1]$ and a normal density centred at five. We have also included 5% of the data as outliers at -4 (intended not to affect the overall distribution in order to investigate the behaviour of our estimates for those points).

Here all of the finite element estimates have been produced using 10,000 data points. The

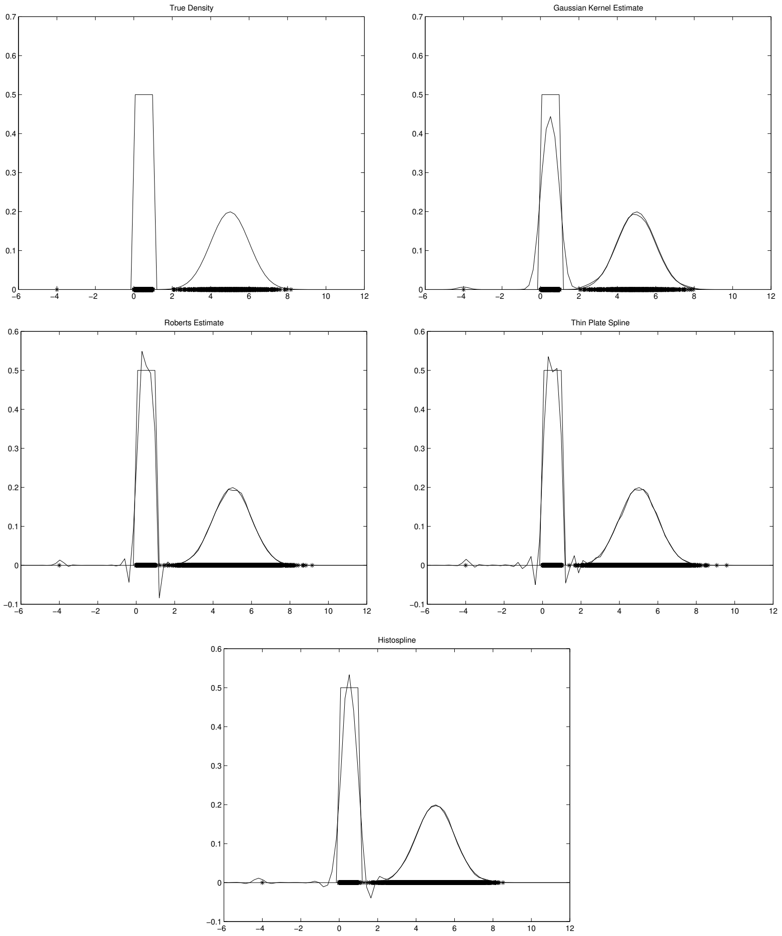
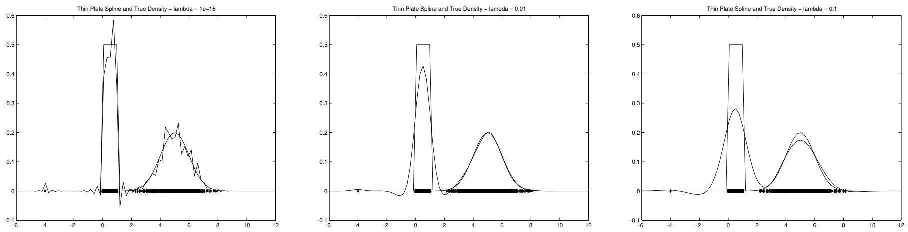


Figure 5.1: Univariate density estimates. Top row: the exact density and Gaussian kernel estimate. Middle: the Roberts estimator and thin plate spline. Bottom: the histospline.

kernel has used only 1,000 due to the computational time required to evaluate it. We can see that all the estimators produce reasonably good results with automatic choices of smoothing parameter. We note that the kernel estimate has had to be considerably smoothed in order to produce a good result for the normal density and this has compromised its fit of the uniformly distributed points.

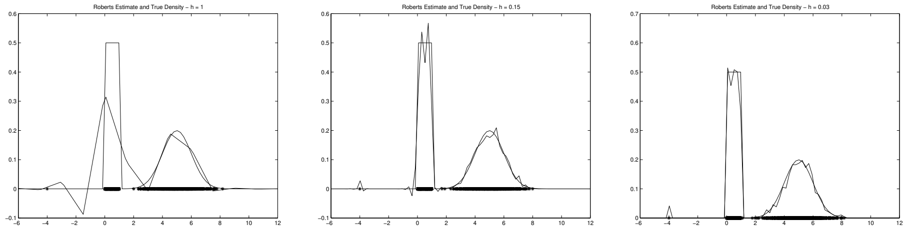
In Chapter 4 we demonstrated the effect of varying the smoothing parameter on the resulting estimate in the bivariate case. Similar results hold in the univariate case. Figure 5.2 presents a similar study using the thin plate spline. It is easy to see that, once again, there is a threshold value at which increasing λ begins to significantly affect the behaviour of the estimator.

What is more interesting here is a study of the effect of the finite element grid on the resulting estimator. As would be expected, for standard pointwise estimation refining the size of the mesh used will produce convergence to a function given by the underlying kernel. Figure 5.3 demonstrates this for a data set using 1000 points making use of the Roberts estimator. We note that while coarse grids do allow the estimator to become negative, this is not the case for very fine grids.



A B C

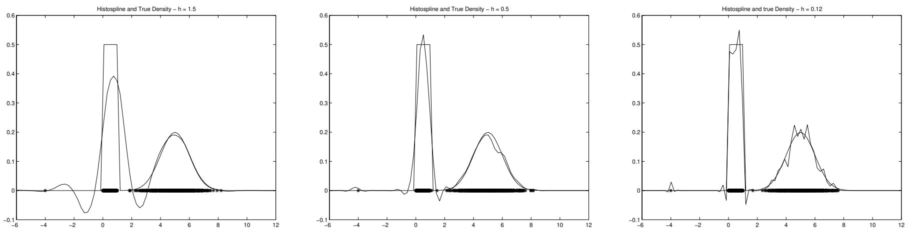
Figure 5.2: Thin plate spline estimates varying the value of λ: 10^{-16} (A), 0.01 (B) and 0.1 (C).



A B C

Figure 5.3: Roberts estimator using 1000 data points and varying finite element mesh sizes. $h = 1$ (A), 0.15 (B) and 0.03 (C).

The same behaviour will not be apparent, however, when we reduce the size of the histogram mesh in the histospline. This has been done in Figure 5.4. We can see that coarse and fine grids both produce errors through either too great a discretisation or through overfitting a rough histogram. Once again, the estimators have been produced using 1,000 data points.



A B C

Figure 5.4: Histosplines for different histograms. Constructed using bin sizes $h = 1.5$ (A), 0.5 (B) and 0.12 (C).

5.1.2 Bivariate Densities

Throughout this thesis, we have been particularly interested in bivariate densities. They represent a considerable move into generalisations to higher dimensions while maintaining a degree of visualisability which is not generally available. We have already explored particular aspects of the behaviour of these methods in §3.2.2 and Chapter 4. Nevertheless, there are a couple of further questions worth considering, similar to those in the univariate case above. Figure 5.5 presents the estimators based on a simulated tri-modal density. Clearly, all the estimators perform well, visually.

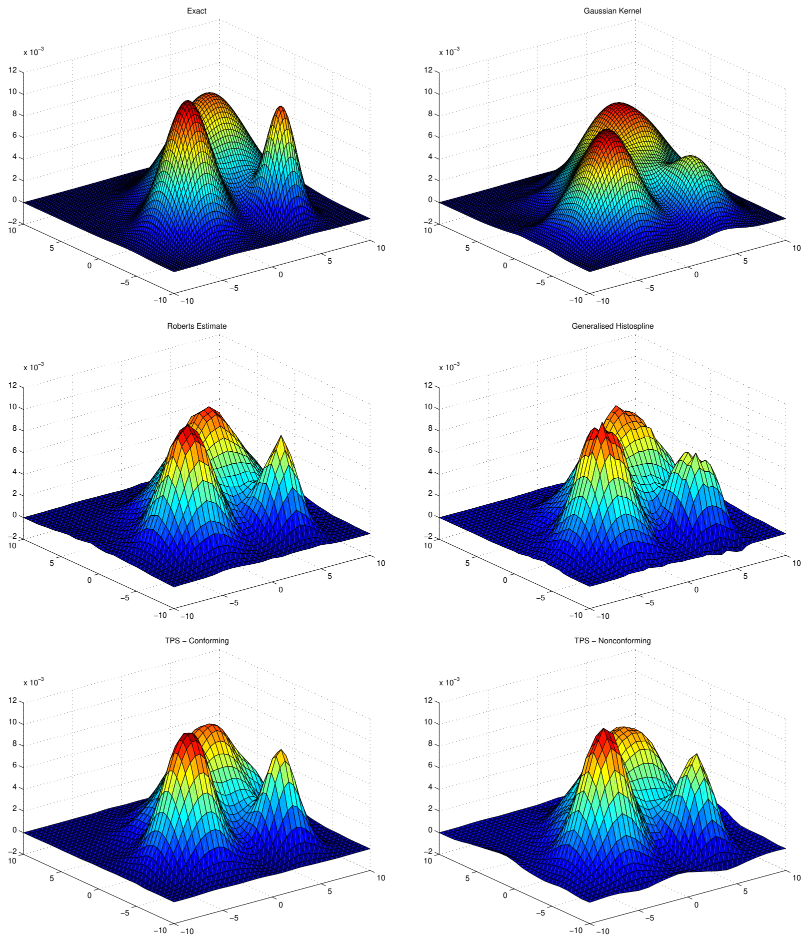
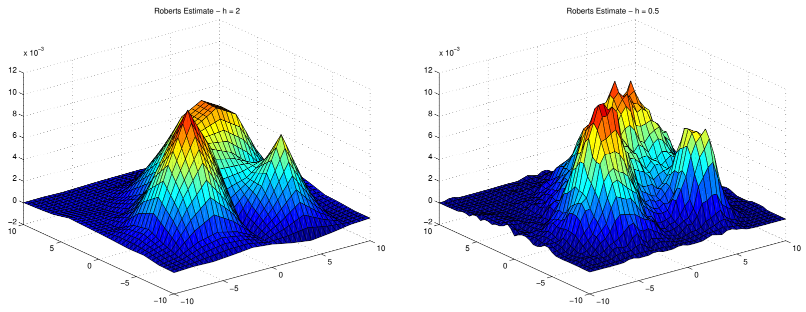


Figure 5.5: The exact density function, with a Gaussian kernel estimate on 1,000 points and finite element density splines on 10,000 points.

We are interested in considering the effect of the use of aggregation in our estimators. In the univariate case, we could make use of pointwise evaluation in the Roberts estimator and this is not available to us here. We have already seen the effect of a poor choice of histogram on the univariate histospline; Figure 5.6 presents a similar study for the above distribution produced using the Roberts estimator. We can see once again that a poor choice of histogram will adversely affect the overall estimate.

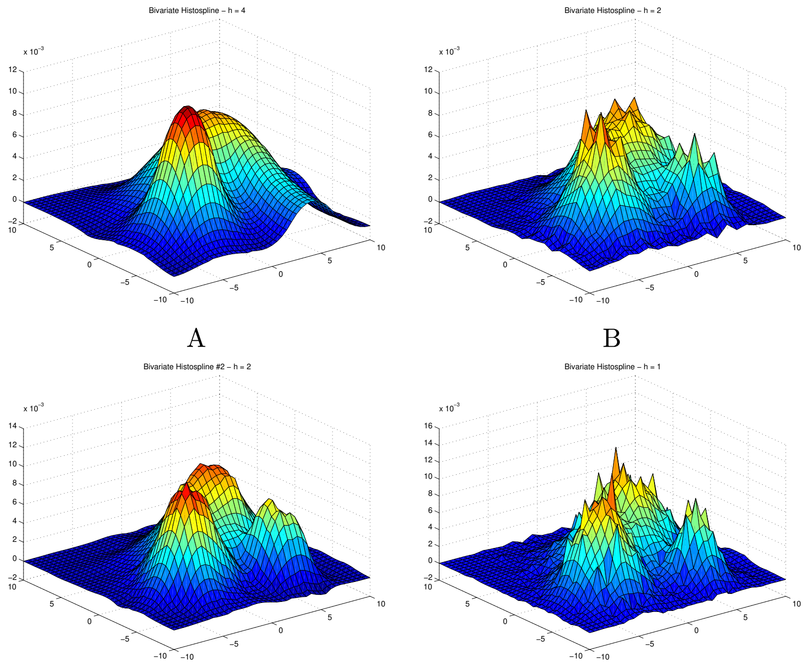
In Figure 5.7 we continue our investigation into the effect of the histogram that we select. In particular, we can observe the usual overfitting and underfitting behaviour of the estimator with poorly chosen histograms. This gives credence to our concern with the Sobolev embedding theorem; using pointwise evaluation removes one potential source of error from the approximation.



A B

Figure 5.6: Roberts estimator varying the size of the histogram mesh. $h = 2$ (A) and $h = 0.5$ (B)

This will be discussed further in §7.1.4. We have also included some variation of the size of the finite element mesh relative to the size of the histogram discretisation. Clearly, a discretisation close to the size of the finite element mesh will, itself, produce large errors. However, we have only needed to refine the histogram mesh by a factor of four in the finite element approximation to produce reasonable results. This supports the claims made in §4.3.



C D

Figure 5.7: Bivariate histosplines varying finite element mesh size and histogram mesh size. Top: $h = 4$ (A), $h = 2$ (B) for finite element mesh of size 1. Bottom: $h = 2$ (C) and $h = 1$ (D) for finite element mesh size 0.5.

5.2 Monte Carlo Results

So far we have made a visual demonstration of the behaviour of our methods and this suggests that they do, in fact, perform well. We will now undertake an investigation of the quantitative convergence of the methods. The following experimental results have been gained by using the Matlab "rand" and "randn" functions to generate uniformly and normally distributed points. For each value of n that is considered, the estimators have been run on 10 independent simulations of a data set of size n and the average squared pointwise error has been calculated. The results are given in the following table for a normal distribution with unit variance. The squared error has been calculated at zero and 1.6 (one standard deviation). Here λ has been chosen using least-squares cross-validation. We have also given the average rate of convergence over the simulation.

$\log(n/500)$	At 0			At 1.6		
	Roberts	H-spline	TPS	Roberts	H-spline	TPS
1	0.31×10^{-2}	0.67×10^{-4}	0.98×10^{-4}	0.0109	0.0058	0.0064
2	0.94×10^{-3}	0.82×10^{-4}	0.35×10^{-4}	0.0046	0.0076	0.0065
3	0.31×10^{-3}	0.38×10^{-4}	0.63×10^{-4}	0.0075	0.0065	0.0064
4	0.40×10^{-4}	0.17×10^{-4}	0.15×10^{-4}	0.0069	0.0068	0.0064
5	0.45×10^{-4}	0.48×10^{-4}	0.19×10^{-4}	0.0067	0.0076	0.0064
6	0.14×10^{-4}	0.95×10^{-5}	0.46×10^{-5}	0.0066	0.0066	0.0064
Avg. Rate	-1.1	-0.40	-0.61	-0.1031	0.0026	-0.0016

Figure 5.8: Errors for the Roberts estimator, bivariate histospline and thin plate spline density estimator as the number of data points is increased. Measurements are against a normal distribution at its mean and standard deviation.

Clearly the results are mixed and somewhat unexpected. The large amount of variation in the results indicates that the number and size of simulations was insufficient to achieve consistent results. It is surprising that the thin plate spline estimate has performed so poorly. In fact, its errors in absolute terms are consistently better than those of the other estimators, but the rate of convergence is substantially slower. In part, this may be due to the effect of the finite element approximation. The Roberts estimator performs particularly well; however, in the simulations run, this estimator consistently chose the minimum value set for λ. This suggests that in fact the finite element grid is too coarse for the number of data points and what we are really seeing is the effect of a finite element convergence (which we would expect to be $O(h^2) = O(n^{-1})$) rather than the convergence of the fundamental solutions themselves. The thin plate spline approximation, on the other hand, did choose progressively smaller values of λ. In this case we are able to assume that the estimator is in H^2. We also know that the kernel for the estimator behaves optimally and we would therefore expect a convergence rate of $4/5$. The rate we achieve is about half that, leading us to suspect that there may be some errors in the algorithm. We note that the histospline converges at approximately the expected rate.

We do achieve at least some convergence close to the mean of the data. If we look at convergence at one standard deviation, however, we see that all the methods stop converging with an error of 0.0065. This may partly be due to the difficulty of estimating densities in regions with sparse data. It may also be due to inaccuracies in Matlab's generators.

The next table displays the results for a simulated uniform distribution. In this case we have again run ten algorithms for every sample size. In this case we have also taken the average error at ten evenly distributed points across the domain. In this case we observe that convergence is achieved across the range. The thin plate spline has a rate of convergence that is closer to what we might expect, even though the true density function is discontinuous. In the case of the histospline, we have started off with a very good estimate, which may have reduced the later rate of convergence. The lower rate may also be due to a poor choice of histogram.

$\log n$	Roberts	H-spline	TPS
1	0.0706	0.0086	0.0360
2	0.0281	0.0084	0.0073
3	0.0133	0.0138	0.0173
4	0.0046	0.0082	0.0021
5	0.0012	0.0019	0.0010
6	0.0009	0.0009	0.0009
Avg. Rate	-0.8334	-0.4620	-0.7449

Figure 5.9: Averaged pointwise errors and average rate of convergence on a uniform distribution using the Roberts estimator, bivariate histospline and thin plate spline estimator.

5.3 Classification Experiments

One of the foreshadowed possible applications of these density estimates is in the area of classification. There are a number of ways of approaching this, of which density estimation is one. We will first of all give a brief introduction to the theory behind these algorithms, largely derived from [Fri94], and proceed to some experimental demonstrations of their performance. Suppose that we wish to assign an object $x \in \mathbb{R}^d$ to one of J groups, $\{G_i\}_{i=1}^J$, where the correct group is unknown. Let us suppose that we have some measure of loss L_{jk} which assigns a numerical value to the loss incurred by assigning x to group G_j when it is really in G_k. We might think about the loss incurred by a company for losing a customer's business, for example. What we wish to do, then, is to minimise the risk of misclassification defined for each G_j as:

$$R_j = \sum_{i=1}^{J} L_{jk} P(G_k|x)$$

where $P(G_j|x)$ is the probability of being in class G_j given the position x. This is not particularly enlightening, but we may transform the problem by making use of Bayes theorem:

$$P(G_j|x) = \frac{P(x|G_j)P(G_j)}{\sum_{i=1}^{J} P(x|G_i)P(G_i)} \tag{5.1}$$

where $P(G_j)$ is the unconditional probability of being in group G_j (in our terms, the theoretical proportion of the data in G_j) and $P(x|G_j)$ is the probability of observing an object from class G_j at the point x. In other words, $P(x|G_j)$ is the value of the density function for G_j at x. Now we have reduced the classification problem to already known values. If we have sample groups from each G_j we can proceed to estimate their probability density function and then to calculate $P(G_j|x)$.

In what follows we will make a few simplifications to the method. To begin with, we will be primarily interested in predicting the correct group, rather than the costs of getting it wrong, and we will therefore assign $L_{ij} = 1 - \delta_{ij}$ where δ_{ij} is the Kronecker delta. We will also consider only two groups and notice that at this point we can ignore the denominator in (5.1) and merely compare $P(x|G_1)P(G_1)$ and $P(x|G_2)P(G_2)$. This is a relatively simple exercise and we will briefly present the results of a few studies.

To begin with, it would be nice to demonstrate that for spatially distinct groups the method will work well. In all the simulations, each of the two groups used to create the density functions consist of 10,000 data points produced using Matlab's "rand" and "randn" functions. The Roberts estimator has been used here due to its computational efficiency. Figure 5.10 demonstrates the method on spatially separated data sets and in this case there is no classification error. The graphs presented in these figures should all be interpreted in the same manner. Surface plots will come in pairs, indicating the density spline surfaces for each data set. Graphs of points will also indicate the real or predicted labels of points in a test data set, indicated by $*$ and $\diamond$. A more difficult problem is posed in Figure 5.11 by two uniform distributions which, while not overlapping, do actually touch. Once again, the method proves able to cope with this - there is an error here of 0.6%.

At this point we will consider a more difficult classification problem and allow the groups to overlap. Figure 5.12 presents the results when both groups are weighted evenly and for weightings $P(G_1) = 0.75$ and $P(G_2) = 0.25$. The resulting errors are 18% and 19% and we can see that the density estimates effectively make a division along a hyperplane separating the two nodes as might be expected. The weighting here affects the position of the hyperplane. A larger error for an uneven weighting is understandable - the data set on which we are testing the classification contains an equal proportion of points from G_1 and G_2. A better realisation of a classification problem for groups with unequal probabilities is given in Figure 5.13. The test data set now contains 375 (75%) points from group 1 and 125 (25%) points from group 2, and the classification algorithm has been weighted similarly. The results here are much more encouraging. We achieve a 12.5 % error rate for the overlapping densities suggesting that the methods are reasonably effective.

So far, our studies have relied on data generated by Matlab in order to test the behaviour of our methods. Here we will make use of some synthetic data from the German National Research Centre for Information Technology [MRW+99]. These sets have been labelled "banana data" for reasons which are obvious in Figure 5.14. One hundred versions of the data are available along

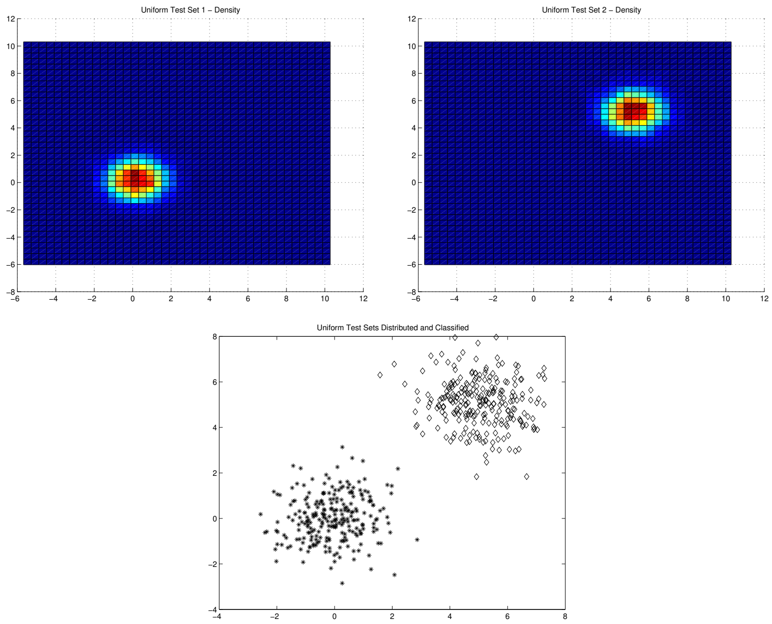


Figure 5.10: Classifying points in separated normal distributions - classified with zero error.

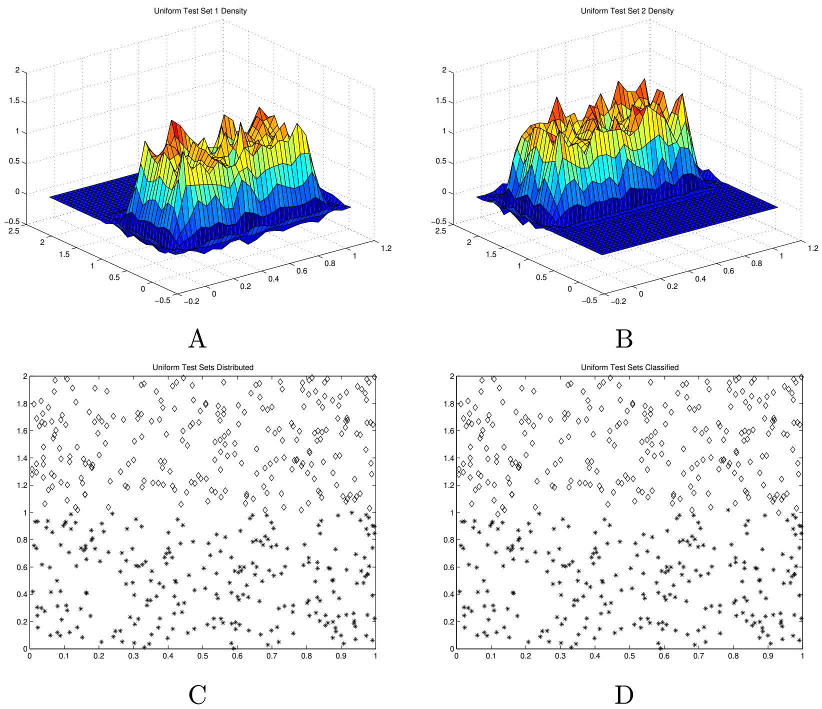


Figure 5.11: "Touching" uniformly distributed sets. (A) and (B) represent surface plots for each data set. (C) provides the true labels of the test data set and (D) the labels that the classification algorithm predicts. Here there a 0.6% error.

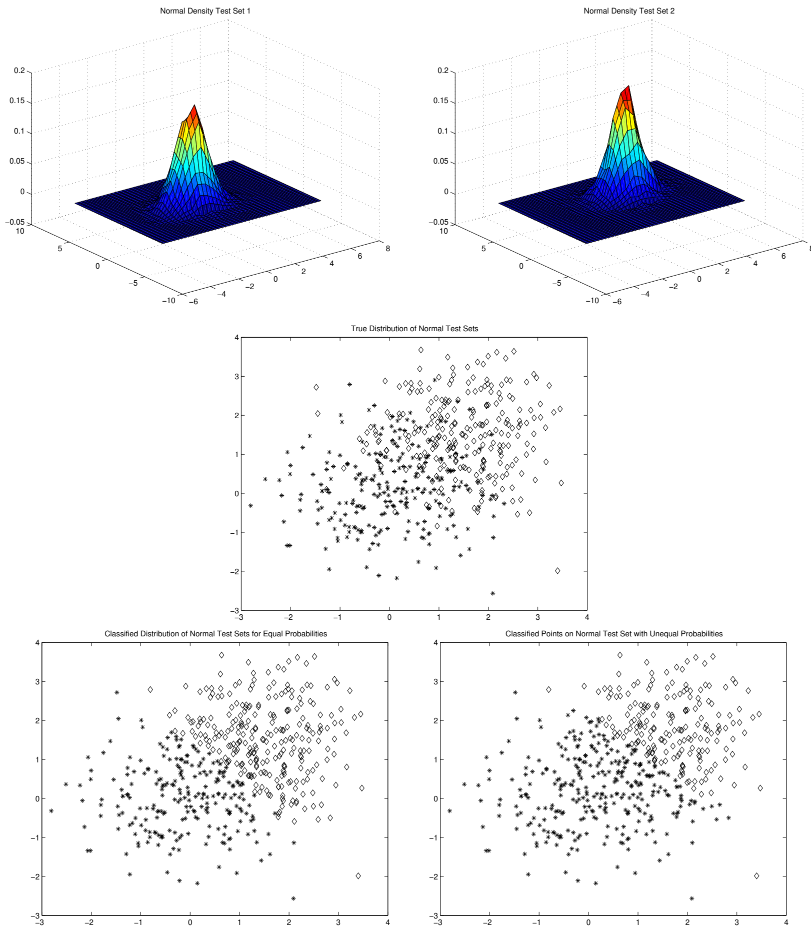


Figure 5.12: Overlapped normally distributed sets. Top row: density functions for both sets. Middle: real distribution of points in the test set. Bottom: classification of points with equal probability weightings and with $P(G_1) = 0.75$ and $P(G_2) = 0.25$. Errors are 18% and 19% respectively.

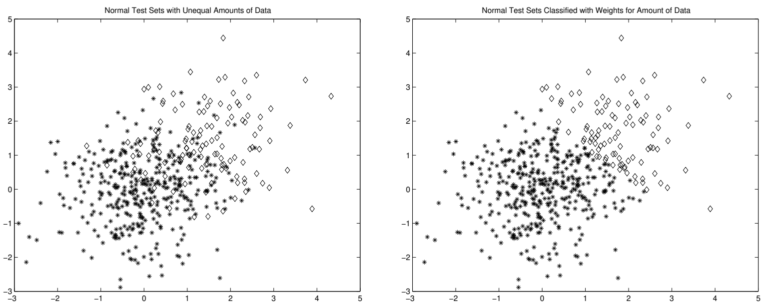


Figure 5.13: True (left) and predicted (right) labels for unequally weighted overlapping normal distributions as in Figure 5.12 with similarly weighted sets of test data. Error is 12.5%

with test data sets to judge the errors of classification algorithms. Visual results for data set number 30 are provided in Figure 5.14.

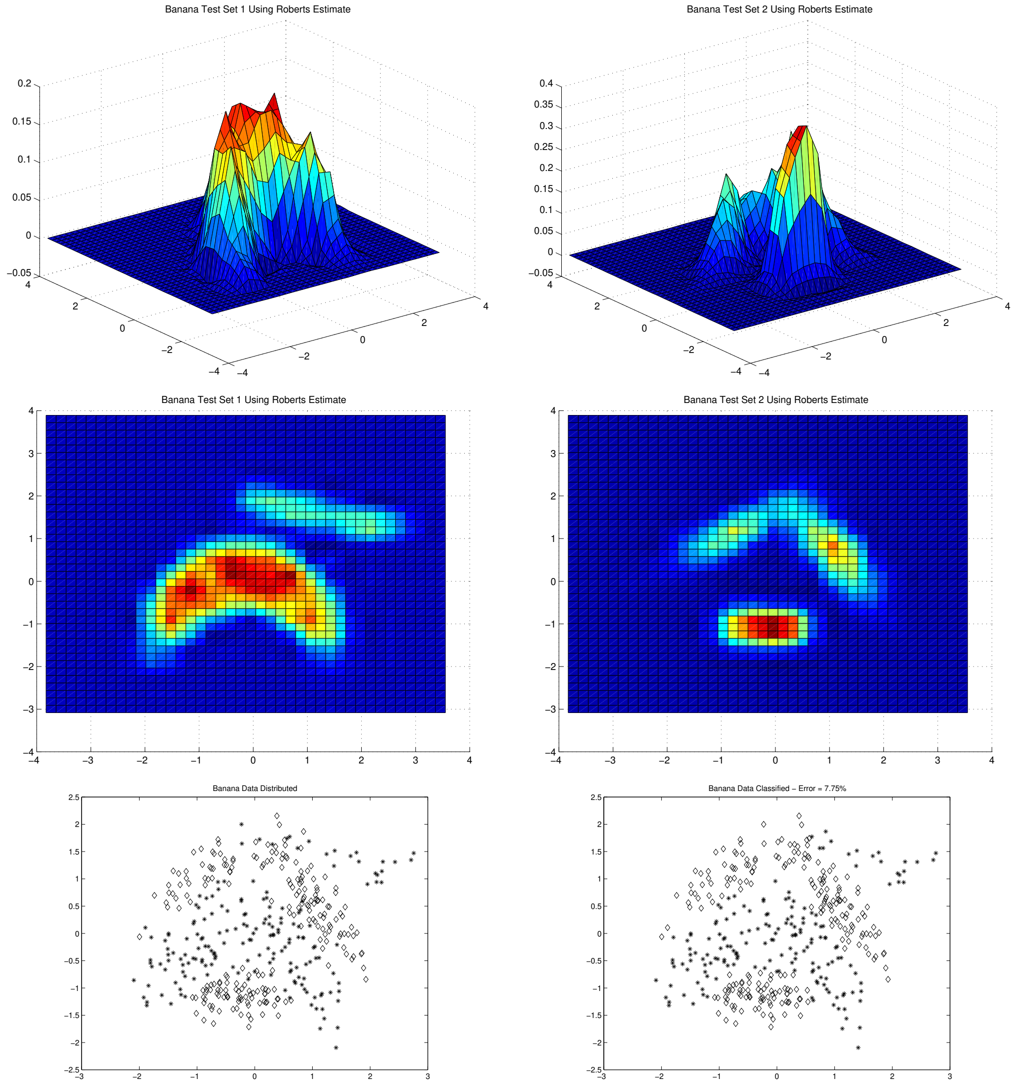


Figure 5.14: Classification study for banana data set 30. Top rows represent standard views of the surfaces. Middle rows are overhead views and bottom rows are the real (left) and classified (right) membership of the points in the test set. Error is 7.75%.

In fact, the use of the banana test set allows us to make a comparison with other estimators. [MRW+99] provides classification results for a number of standard estimators with averages and standard deviations over the 100 data sets. In our case, the Roberts estimator classifier has been run on 10% of these sets and appears to significantly outperform the other estimators. There are some differences involved in the experiments here - [MRW+99] used the first 5 sets to produce smoothing parameters before performing the classification. The results above used least-squares cross-validation on each data set separately. The following table gives the performance.

Classifier	Error
RBF	10.8 ± 0.7
Regularised AdaBoost	10.9 ± 0.4
Gaussian Kernel	11.5 ± 0.7
Fisher Kernel Discriminants	10.8 ± 0.5
Roberts Density Estimator	9.4 ± 1.4

Figure 5.15: Classification error rates on banana data comparing the Roberts estimator with standard classification schemes.

Chapter 6

Real-World Data

We have shown that the estimators we have developed perform well on simulated experiments. In real-world applications, however, datasets often have much less structure than the simulations we produce. It is therefore important to demonstrate the performance of these methods on real-world data. The experiments performed below have been carried on on a database of vehicle insurance policies maintained by the CSIRO for a leading insurance firm. There are two experiments that have been undertaken here. In the first section we will use representations of density in an exploration of the real valued variables in the data set and in an evaluation of the performance of our tools. The second section will consist of an application in the area of classification. For some fields containing boolean variables, we will create two bivariate density estimates (on selected fields), each of which will correspond to a set of points taking one value of the boolean variable. We will then use these densities to try to classify some unused set of the data as in §5.3. All the densities will be created using 10,000 randomly sampled records from the full data set.

6.1 Exploring the Data Set

The vehicle insurance data set in CSIRO archives consists of 1,601,277 records in 22 fields. These cover data about the policy holder (age and area code), the vehicle (make, year, engine type and capacity) and the policy (the sum insured for, the cost of claims made, duration and how long the policy has been in effect). There are various other fields concerning the type of finance applying to the vehicle, the no-claim bonus level on the policy and so forth. We have selected four real-valued fields with which to produce bivariate densities:

- Claim Cost - the cost of claims made on each policy (zero when no claim has been made)
- Age - the numeric value of the policy holder's age.
- Sum Insured.
- Vehicle Point Score - the insurance company's assessment of the risk associated with the vehicle; created by expert assessment.

All of the simulations will be done using the log of the Claim Cost and Sum Insured variables in order to reduce the effect of outlying data points. We will later perform classification experiments, trying to predict the corresponding variables in the fields:

- Use - whether or not the vehicle is being used for business or private purposes.
- Under 25 - whether or not the vehicle is insured for a driver under 25 years of age.
- Claimant - whether or not a claim has been made against the policy (created from Claim Cost).

6.2 1D Density Estimates

We will firstly produce univariate densities for each of the variables. These have been done using a first-derivative penalty function and have been smoothed by sight. Figure 6.1 demonstrates the examples.

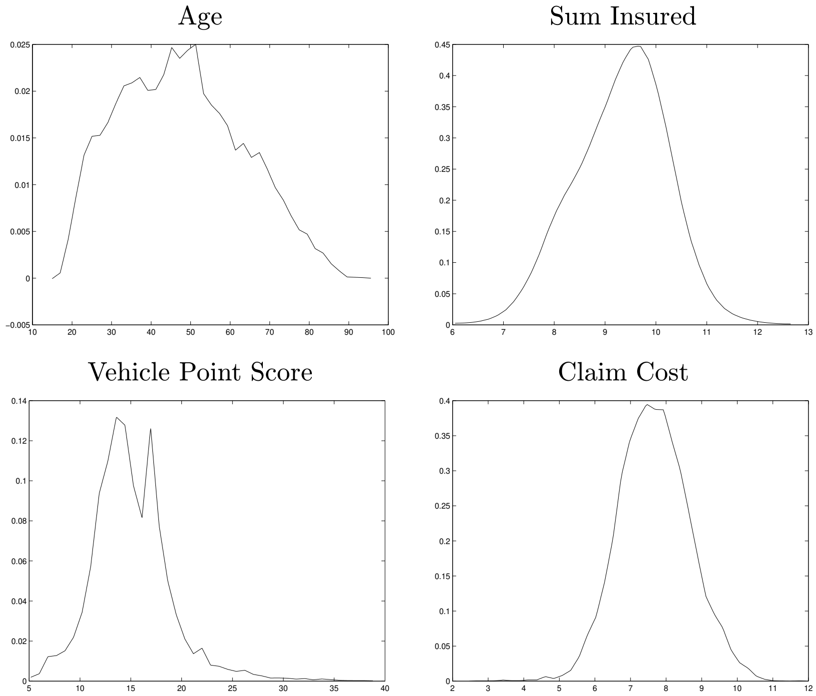


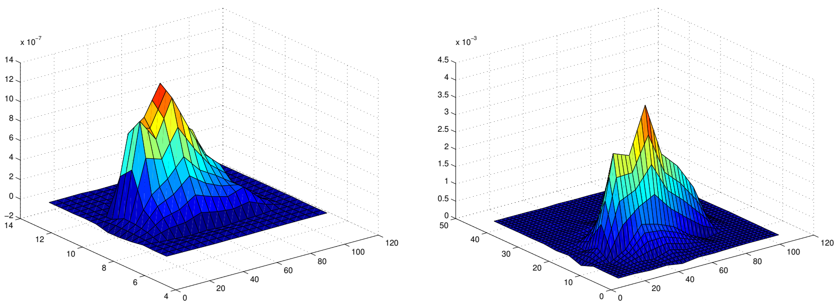
Figure 6.1: One-dimensional density estimates for the variables Age ($\lambda = 0.5$), Sum Insured ($\lambda = 0.1$), Vehicle Point Score ($\lambda = 0.5$) and Claim Cost ($\lambda = 0.01$).

Here Claim Cost has been calculated only on those records whose entries were nonzero. We note that unusually high smoothing parameters have been necessary, in comparison to those used with test data sets, to produce these results. This indicates a certain degree of roughness in the data. We see that while the Claim Cost and Sum Insured variables both have reasonably smooth functions for smaller smoothing parameters, even the larger values of λ have failed to smooth out the densities associated with Age and the Vehicle Point Score. In fact, these both appear to be possibly multi-modal and we will see the patterns represented here re-appearing in bivariate representations.

6.3 2D Density Estimates

Having established some feeling for the densities in one dimension, we will move to the bivariate case. The aim here is not only to demonstrate that sensible densities are produced, which will demonstrate interesting features, but also to examine the behaviour of the Roberts estimator and the nonconforming thin plate spline outlined in Chapter 3 in a practical example. There are six potential pairings of the variables we are using. We will restrict parings with the Claim Cost variable only to records for which Claim Cost is non-zero, with the rest of the representations coming from the full data set. The densities of subsets of the data with particular features will be left to the classification experiments.

We will begin in Figure 6.2 by presenting the densities for Age compared respectively with Sum Insured and Vehicle Point Score. In the following representations, the first variable listed is always found on the right-hand axes of the graph. Here we can see that the densities produced seem relatively smooth in the second variable, but we notice the existence of a levelling off in the 20–40 region, followed by a peak between 40 and 60. Perhaps this represents a relative increase in affluence in that age bracket before people start to give up driving. The bump ends at around 60, and this may be due to retirees affording less valuable vehicles.



A B

Figure 6.2: Density estimates on Age (right-hand axes) vs Sum Insured (A) and Vehicle Point Score (B) with smoothing parameters chosen by least-squares cross-validation.

Somewhat more interesting patterns may be found in the joint density of Sum Insured and Vehicle Point Score, shown in Figure 6.3. There appear to be two significant peaks in the distribution and we have verified here that this is not merely a product of the random sample. It is difficult to view even a two-dimensional density on a flat surface and we have introduced a view from above demonstrating that the second peak occurs at an offset to the first in both variables. It appears that while most policies are clustered around one particular set of values, there is a significant minority who have cheaper, riskier, vehicles. This is not surprising except to the extent that the population is divided into two relatively distinct clusters. We note that the Roberts estimator has produced a significantly less smooth estimate. This is partly explained by two things: (i) it is computationally cheaper to evaluate and has therefore been produced with a smaller finite element mesh, (ii) it employs a weaker smoothing term and for this reason may be predisposed to give rougher results.

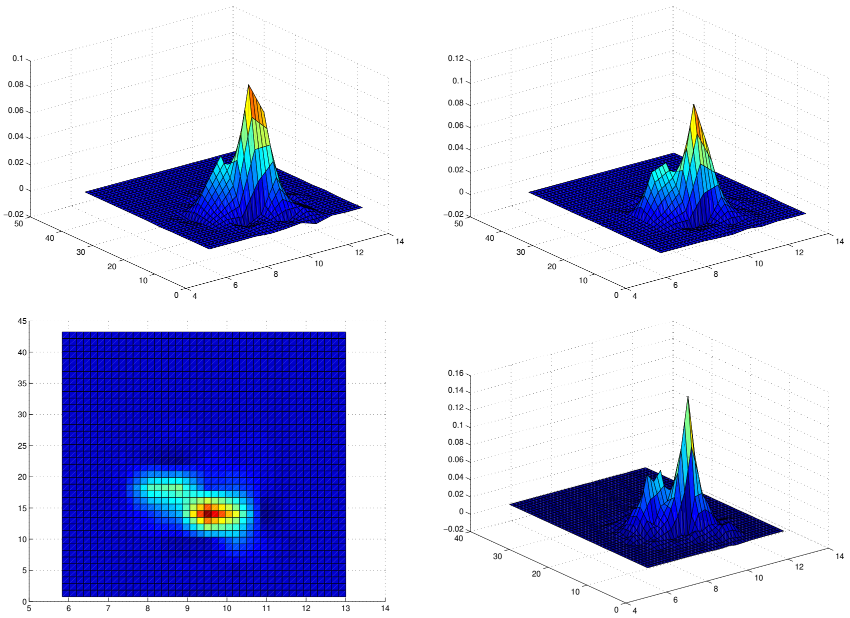


Figure 6.3: Density functions on Sum Insured (right/horizontal axes) and Vehicle Point Score (left/vertical axes) in various representations. The density function for two different random samples generated by the nonconforming method (top), a view from above demonstrating the multi-modality of the distribution (bottom left) and the Roberts estimator density (bottom right).

We will now turn to an examination of the densities involving claims that have been made and their relationship to the cost of those claims. We will begin by examining the joint densities of Claim Cost versus Age and versus Vehicle Point Score in Figure 6.4. It is again possible to see the univariate patterns with respect to the density of Age and the other variables being repeated. The Roberts estimator has once more proved to be somewhat rougher than the nonconforming method for substantially the same reasons. One feature of interest is the apparent "prong" leading towards the younger end of the Claim Cost and Age density in the Roberts estimator. Taking different random samples indicated that this was largely a product of the particular points, although the function is somewhat broader in that region than the nonconforming method would indicate. This may well be a product of the relatively coarse finite element mesh used for the nonconforming method.

A somewhat more interesting result may be found for the joint Claim Cost/Sum Insured density in Figure 6.5. A number of features are important here. To begin with, an examination of the first two rows indicates that the large mesh sizes have had a significant effect on the representation of the density. Indeed, a comparison of the nonconforming method with the Roberts estimator

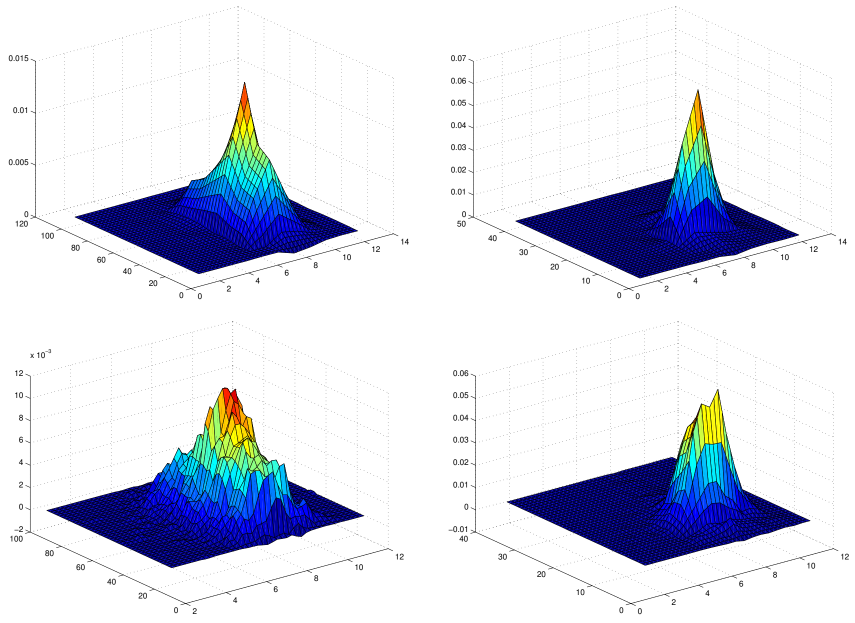


Figure 6.4: Density functions for Claim Cost (right-hand axes) with Age (left) and Vehicle Point Score (right) using the nonconforming method (top) and the Roberts estimator (bottom).

using the same grid indicates that they achieve similar results, although some greater amount of smoothing is retained by former. In terms of the data, there is a very interesting phenomenon concerning the "line" which is readily apparent from the overhead views and which the coarser nonconforming method does not make very clear. The explanation for this "cluster" is that the insurance company has a cap on the total pay-out that they will make for each policy and this is directly related to the sum insured. Points along this line, therefore, represent policies on which the maximum pay-out has been made, whereas the great majority of claims will be made on less than the full value of the policy. It is unclear in this case whether a desirable density estimate would produce a separate cluster for these points, or if they should merely be an edge to the original density. The plots here tend to suggest that there is a significant "thinning" of points between these two groups - probably an indication that past a certain point of damage it will be more profitable to replace the vehicle than to repair it.

6.4 An Experiment in Classification

We present results of a brief classification exercise using these algorithms. Three Boolean variables have been selected from the data on which to perform this experiment: Under 25, Type of Use and an indicator of whether or not a claim has been made. Clearly, successfully predicting whether or not a policy will contain cover for a driver under 25, or will list the purpose of the vehicle as business or private is not a very useful application. Predicting whether a claim will be made might have some use. Nevertheless, these represent a good test of the algorithms in a classification context. In all of the following, classification schemes have been run by producing

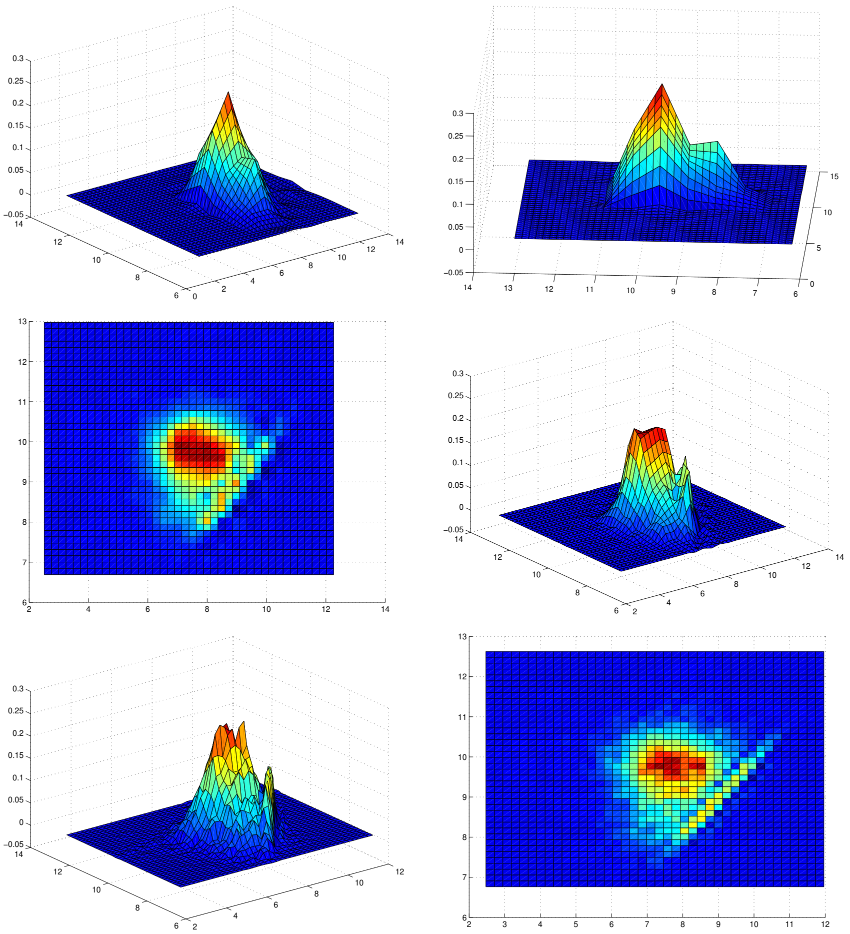


Figure 6.5: Density functions for Claim Cost (right/horizontal axes) and Sum Insured (left/vertical axes). Views using the nonconforming method on a course grid (top row), on a fine grid (middle) and the Roberts estimator (bottom).

densities for each population using a sample size of 10,000 (except in cases where this figure represents close to all of the population) and choosing the smoothing parameter by least-squares cross-validation. They have then been tested on a separate data set consisting of 250 data points from the remaining points in each population. Because we are taking equally sized sub-samples of the data to test the accuracy of these classifications, we have assigned the probability of the two groups to be equal. This would not be the case for a real-world application. Error has as usual been measured in terms of the proportion of points misclassified.

The classification algorithms that we have run on simple, simulated data sets have proven to be reasonably successful, but in those cases the two sets of data were separated fairly distinctly. We would not expect this to be the case in real-world data, particularly not for two-dimensional simulations. Indeed, we have generally not succeeded in achieving a better than 35% error rate in any of the simulations tried and no better than 46% in classifying whether or not a policy holder has made a claim. Looking at the distribution of points, however, this is not particularly surprising. Greater accuracy may well be achieved by creating higher-dimensional densities (which will no longer be easily visualisable) and incorporating some discrete variables into the decision algorithm. We will present a couple of examples of the results that we are able to achieve. Figures 6.6, 6.7, 6.8, 6.9 and 6.10 are all presented in the same format. The densities for points taking each value in the Boolean variable are presented in the top two figures, with the data points which have been classified listed in the bottom two figures, real values indicated in the right-hand figure (labelled by $*$ and $\diamond$), predicted values on the left. In the case where claim cost has been used as a predictor, the samples are derived only from policies on which claims have been made. Excepting Figure 6.6, which was produced using the nonconforming method, the simulations have made use of the Roberts estimator due to its computational efficiency.

Figure 6.10 has been included here despite an unfair advantage in using age to predict whether or not the policy includes a driver under 25 years of age. It naturally has a much greater rate of success. We have included it due to the demonstration of the demographic of those with drivers under 25 years of age. The division of this group into those policy holders who are themselves under 25 and those between 40 and 60 is very clear and gives a good indication that such drivers are registered for themselves, or by their parents. This is not very surprising. The lack of disparity between the sum insured for these two groups is a little more surprising, especially as both appear to span the main body of the range of values in Sum Insured. The youth, it might appear, are wealthier than might be thought. We have indicated the error (measured as a percentage of points incorrectly classified in each figure).

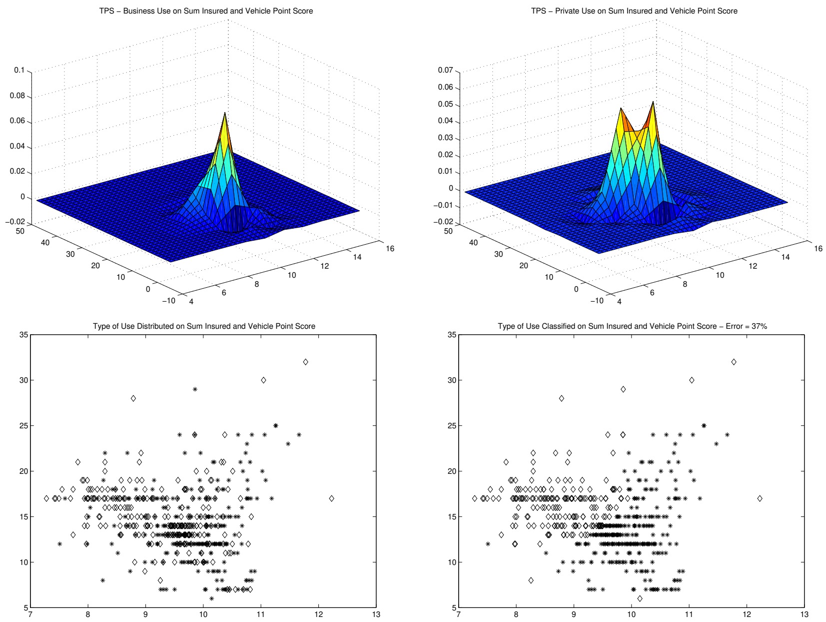


Figure 6.6: Classifying Type of Use using Sum Insured (right/horizontal axes) and Vehicle Point Score (left/vertical axes) as predictors. Plots of density functions for business use (top left - label is $*$), private use (top right - label is $\diamond$) and the true labels (bottom left) and predicted labels (bottom right) of the test set. Error is 37%.

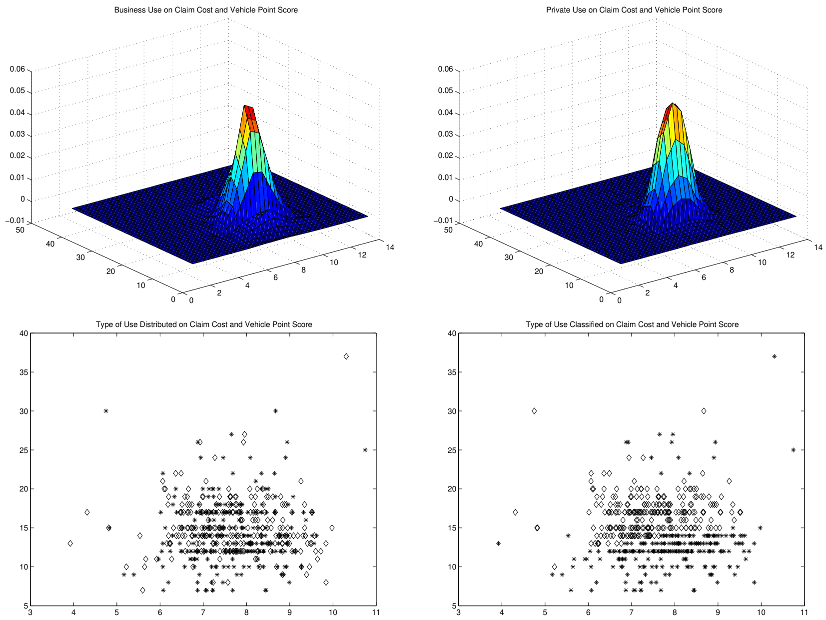


Figure 6.7: Classifying Type of Use using Claim Cost (right/horizontal axes) and Vehicle Point Score (left/vertical axes) as predictors. Plots of density functions for business use (top left - label is $*$), private use (top right - label is $\diamond$) and the true labels (bottom left) and predicted labels (bottom right) of the test set. Error is 35%.

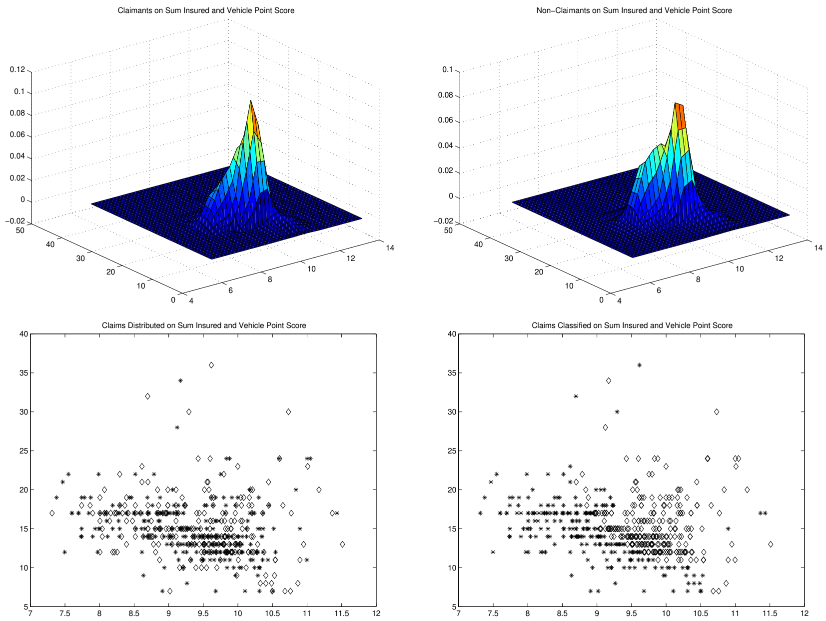


Figure 6.8: Classifying whether a claim has been made using Vehicle Point Score (left/vertical axes) and Sum Insured (right/horizontal axes) as predictors. Plots of density functions for claimants (top left - label is $*$), non-claimants (top right - label is $\diamond$) and the true labels (bottom left) and predicted labels (bottom right) of the test set. Error is 46%.

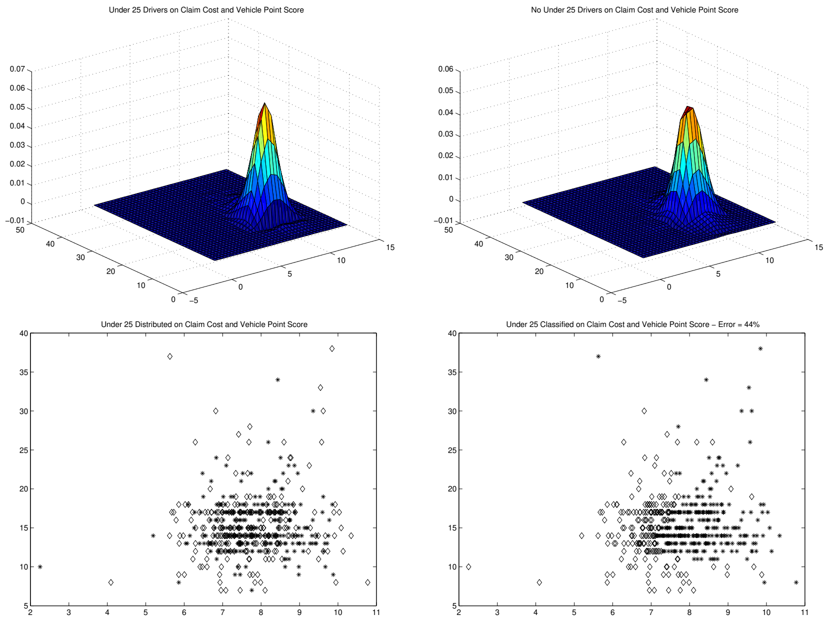


Figure 6.9: Classifying Under 25 using Claim Cost (right/horizontal axes) and Vehicle Point Score (left/vertical axes) as predictors. Plots of density functions for drivers under 25 (top left - label is $*$), no under 25 drivers (top right - label is $\diamond$) and the true labels (bottom left) and predicted labels (bottom right) of the test set. Error is 44%.

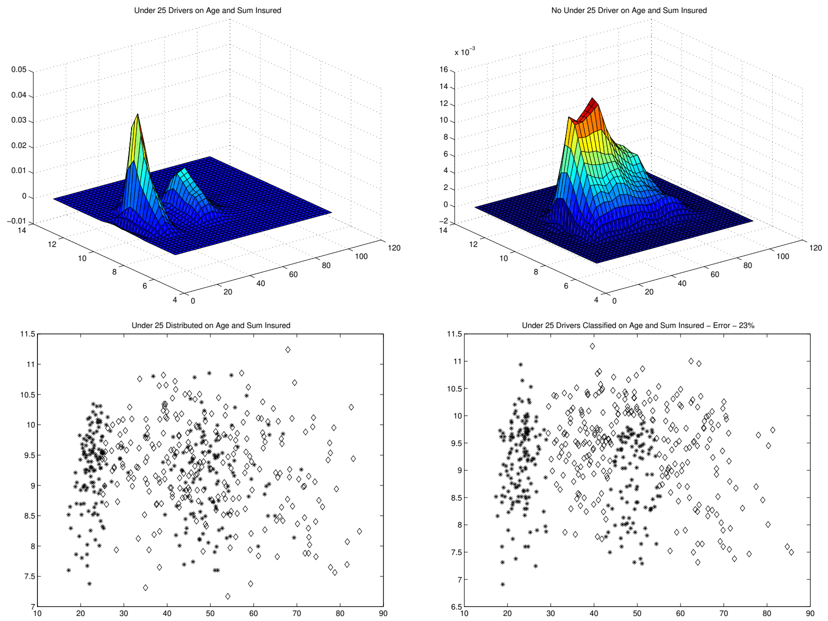


Figure 6.10: Classifying Under 25 using Age (right/horizontal axes) and Sum Insured (left/vertical axes) as predictors. Plots density functions for under 25 drivers (top left - label is $*$), no under 25 drivers (top right - label is $\diamond$) and the true labels (bottom left) and predicted labels (bottom right) of the test set. Error is 23%.

Chapter 7

Conclusion

Clearly, the proposed estimators described here are potentially powerful tools. They represent a scalable method of estimating a density function while maintaining important properties such as empirical moments. Further, they are amenable to many of the standard finite element techniques such as nonconforming methods. In this sense they are very powerful. Nevertheless, these density splines must still be considered to be in their developmental stages. There remains a set of important practical and theoretical questions which have not so far been dealt with in this thesis. While we have demonstrated some applications for the density splines, there are others which might be explored. We will make an attempt to cover these issues here.

7.1 Unanswered Questions

A number of important questions concerning the behaviour of the methods above remain unanswered in both a practical and a theoretical sense. One of the important factors here is in the area of asymptotic results; we have yet to find a proof that the methods will converge in the limit as $n \to \infty$. Beyond this, the methods for automating a choice of smoothing parameter we have developed do not perform particularly satisfactorily and the reasons for this will be discussed. There are also questions concerning the correct finite element mesh size to choose in relation to the size of the data set. Given that we are choosing a finite element mesh size, it is possible to ignore the Sobolev embedding theorem in the finite element subspace. The consequences of doing so warrant some investigation. Related to this question are the concerns raised generally about the behaviour of these methods in high-dimensional spaces and the way to reduce the problems associated with working in such spaces.

7.1.1 Convergence

One of the concerns involved in any estimator of any function is whether or not, given enough computational power or sufficient data, the method may be said to converge to the exact solution. In fact, for the purposes of statistics, this is not necessarily a concern - there is very rarely more than a finite set of data available for use. Nevertheless, convergence results are useful in the sense that they tell us that the solutions we produce should approach something like the right answer. While we have managed to demonstrate that convergence can be achieved when aggregated data is used, and while the studies in Chapter 5 indicate that the methods do converge in general, at least away from the tails of distributions, we do not yet have a good proof of this.

Clearly, if convergence can be demonstrated for the case of the theoretical splines defined by (2.6) the finite element approximations can be chosen to converge at an appropriate rate. A natural approach, then, is to try to manipulate the exact solution. We have discussed in §2.2.3 the properties of the kernel estimates derived from the use of $\mathbb{R}^d$ as a domain. In particular, it is known to meet all the conditions of Theorem 1.3.3 except for

$$\begin{gathered} \textstyle\int_{\mathbb{R}^d} |K(y)|dy < \infty \\ \textstyle\int_{\mathbb{R}^d} |y^\alpha K(y)|^{1-1/p} dy < \infty, \ \|\alpha| = m. \end{gathered}$$

The second of these will, in most cases, provide the first. These cannot be readily deduced from the variational equations governing K, but we suspect they this may be shown with Fourier transforms.

The case of non-bounded domains would still not be covered by such a result. We can, however, provide some heuristic motivation to believe that a result of this kind holds. Recall that $\hat{f}_n$ is governed by the equation:

$$\int_\Omega \hat{f}_n(x)s(x) + \lambda C(\hat{f}_n, s) = \frac{1}{n}\sum_{i=1}^n s(X_i), \ \forall s \in H^m(\Omega). \tag{7.1}$$

Now, we have that

$$E\left(\frac{1}{n}\sum_{i=1}^n s(X_i)\right) = \int s(x)\rho(x)dx$$

and this is merely a constant. Let us re-arrange this and substitute $s(x) = \hat{f}_n(x) - \rho(x)$ to obtain:

$$E\left(\int_\Omega (\hat{f}_n(x) - \rho(x))^2 x\right) + \lambda EC(\hat{f}_n, \hat{f}_n - \rho) = 0.$$

It seems likely that we can now allow $\lambda \to 0$ and demonstrate a zero MISE. The dependence of $\hat{f}_n$ on λ makes this difficult, however. The fact that this does not depend on n also remains somewhat problematic. An alternative approach is to consider the finite element case. In this case, we can use the finite dimensionality of the space to reduce λ and get an asymptotic limit:

$$\lim_{n\to\infty} \int_\Omega \hat{f}_n^h(x)s^h(x)dx = \int_\Omega \rho(x)s^h(x)dx, \forall s^h \in V^h.$$

$E\hat{f}_n^h$ could then be regarded as the projection of ρ onto V^h. Reducing the finite element mesh in some relationship to n would then allow us to demonstrate convergence.

7.1.2 Choices of λ

Several methods for choosing λ automatically were developed in Chapter 4 and it was seen that they exhibited fairly similar behaviour. In that sense, then, the numerical experiments have been relatively satisfactory in demonstrating that we only need to consider the practical aspects of an algorithm in order to apply it. However, the actual behaviour of the functionals is concerning. A more detailed examination of their values as $\lambda \to 0$ suggests that they decrease asymptotically. In other words, their optimum value of λ is zero. Examining the visual behaviour of the estimates as λ is varied in §4.5 does not suggest that this will be corrected by the use of other functionals. Essentially, varying λ, once it is small, changes the density estimate relatively little and we therefore would not expect that a different functional would behave very differently for small λ.

The explanation for this anomalous behaviour is that as $\lambda \to 0$ the density estimate essentially becomes the projection of

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}(x)$$

onto V^h. Clearly, there is a limit to which V^h may approximate a sequence of δ-functions and this can then be regarded as a limit on how bad the cross-validation score can become. Experiments by performed varying the finite element mesh size revealed that the functional did not change significantly in the behaviour that it exhibited. However, for very fine meshes the functionals do develop a minimum around the value at which they appear to level off. Figure 7.1 demonstrates a comparison of the least-squares cross-validation functionals for successively halving the mesh size. Zooming in on the critical value for the finest mesh demonstrates that a minimum does in fact occur. This is not the case for larger meshes.

This suggests that relatively fine mesh sizes would be necessary to produce a functional which provides a minimum which may be found in a computationally inexpensive manner. Nevertheless, we are faced with the task of producing an optimal λ for a given finite element mesh. The problem may be rectified in one of two ways. Firstly, some factor may be introduced to artificially create a minimum - possibly some principle may be found to justify it. Alternatively, it would be possible to try to locate the edge of the lower plateau - potentially by maximising the second derivative of the function. The disadvantage of this is that it would be done in the absence of mathematical justification.

7.1.3 Finite Element Mesh Sizes

There is one parameter which, although discussed obliquely, has yet to be dealt with. This is the size of the finite element mesh that is to be used. We obviously desire an approximation to the actual density splines and in this sense the smaller the mesh the better the approximation. Nevertheless, choosing arbitrarily small meshes will have consequences both for the amount of memory that such algorithms use and for the scalability of producing a solution. What size mesh,

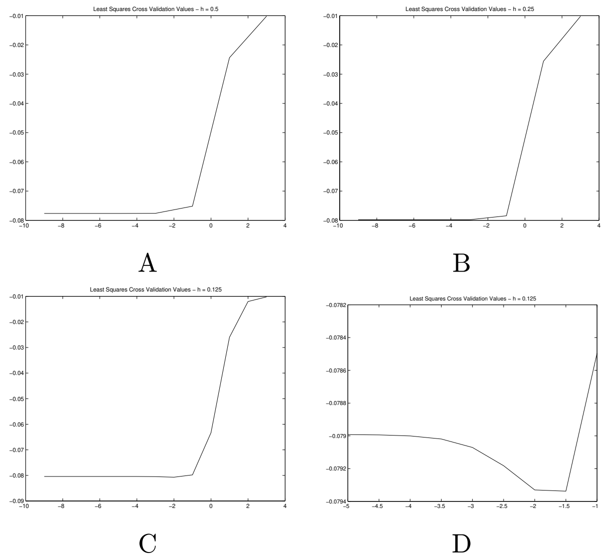


Figure 7.1: Least-squares cross-validation functionals against $\log\lambda$ for the Roberts estimator for a normal density varying the value of h. A: $h = 0.5$, B: $h = 0.25$, C and D: $h = 0.125$.

then, is necessary? In particular, it would be nice to find some relationship between the mesh size that should be employed and the data set on which it is to be employed.

To begin with, we can place some simple bounds, necessary for convergence, which mimic the bounds that would be provided for the smoothing parameters of other estimators. It is clear that if convergence is to be obtained, we will require $h \to 0$. At the same time, we need to have

$$\frac{1}{n}\sum_{i=1}^{n} s(X_i) \to \int_\Omega s(x)\rho(x)dx$$

for each $s \in V^h$. This would imply that $nh \to \infty$. These are the same sorts of conditions required for the convergence of kernel estimators and of histograms.

So far, this is not a particularly specific requirement. In fact, a good bound on the size of the finite element mesh that is necessary to obtain convergence would require some more specific knowledge of the underlying density function (and we would like as coarse a mesh as possible). Nevertheless, it is possible to at least place a bound on the order of h in terms of a power of n.

Recall that we have already shown that a density estimate cannot converge faster than a rate of $n^{-\phi_d(m,p)}$. Now we will assume that we are using an $H^m(\Omega)$ basis for the finite element method and we can therefore achieve an $\mathcal{L}^2$-squared convergence of order h^{2m}. Setting $\hat{f}_n^h$ to indicate the finite element density spline, we have:

$$\begin{aligned} E\int_\Omega (\hat{f}_n^h(x) - \rho(x))^2 &\leq ||\hat{f}_n^h - \hat{f}_n||^2_{\mathcal{L}^2(\Omega)} + E||\hat{f}_n - \rho||^2_{\mathcal{L}^2(\Omega)} \\ &\leq Ch^{2m} + Kn^{-\frac{2m-2d/p}{2m+d-2d/p}}. \end{aligned}$$

Now we merely need to be interested in matching

$$h^{2m} \approx n^{-\frac{2m-2d/p}{2m+d-2d/p}}.$$

This gives us that

$$h \approx n^{-\frac{2m-2d/p}{2m(2m+d-2d/p)}}.$$

However, for most cases we will be interested in, $p = \infty$ and this will give us that

$$h \approx n^{\frac{-1}{2m+d}}.$$

In practice, however, we will want to bound the number of finite elements we have to use, and therefore the size of the system we are required to solve, rather than the actual mesh size itself. Let k be the number of finite elements; we have that $k \approx h^{-d}$. This gives us that

$$k \approx n^{\frac{d}{2m+d}}.$$

This will be a conservative lower bound in general. We will typically be working in a reproducing kernel Hilbert space which will require that $2m \geq d$. This will give us $k \leq n^{1/2}$. This bound does not look particularly efficient, but it is sufficient to guarantee that we do not interfere with the rate of convergence of the density spline itself. Of course, these calculations assume that we are using some conforming finite element method, but also that density splines do attain the optimal convergence rate. Greater analysis of a data set would be necessary, however, to produce a good mesh size specifically for it.

So far we have also not addressed the issue of the dependence of the mesh size on the choice of parameter λ. Clearly, we would expect that optimal choices of λ will converge to some value as we refine the finite element mesh. Indeed, as λ becomes smaller, we would expect to need a finer mesh size in order to approximate a true density spline, as it will have greater variance. One possible solution to the problems posed in §7.1.2 would be to simultaneously reduce the size of the finite element mesh as we reduce λ when we try to choose λ optimally. This would be considerably more computationally expensive. It would also require some further analysis of the relationship between λ and the finite element error.

7.1.4 Escaping the Sobolev Embedding Theorem

One of the major restrictions that we face in producing density splines is the need to conform to the Sobolev embedding theorem - that $H^m(\Omega) \subseteq C^0(\overline{\Omega})$ for $2m \geq d$. This is required to guarantee the boundedness of pointwise evaluation. Given a finite element subspace, however, pointwise evaluation is automatically bounded and we therefore know that there is an $H^m(\Omega)$ finite element solution even for $m < d/2$. We are then faced with the question of whether it is possible to make use of this fact and avoid an aggregation into a histogram. The resulting estimator on a given finite element mesh might be regarded as a limited δ-sequence.

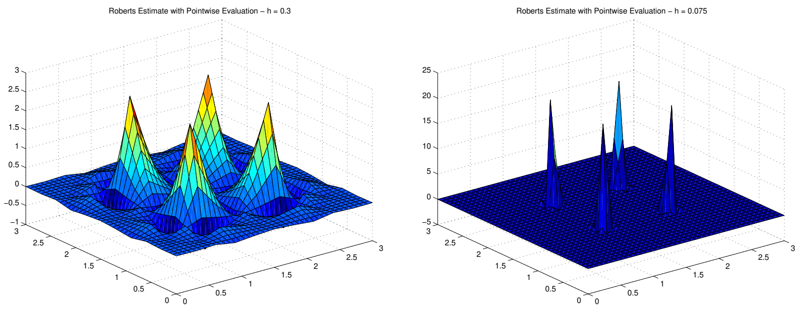


Figure 7.2: Finite element Roberts estimators using pointwise evaluation in the load vector for grid sizes $h = 0.3$ and $h = 0.075$, $\lambda = 0.1$.

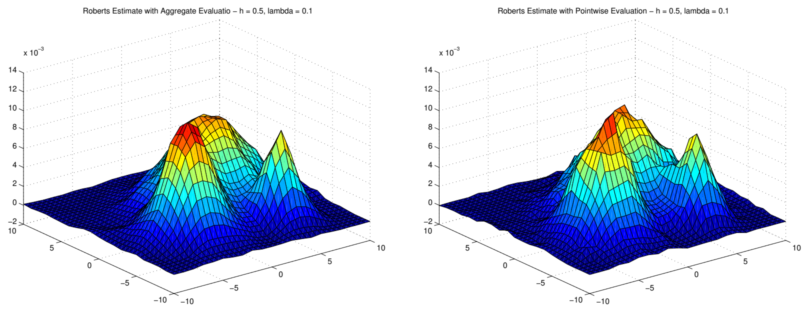


Figure 7.3: The Roberts estimator on a 10,000-point tri-modal density using aggregate and pointwise evaluation in the load vector for $\lambda = 0.1$.

It is clear, to begin with, that we cannot expect to achieve some convergence to a solution which will be unbounded in its supremum. Figure 7.2 presents an experiment designed to look at the behaviour of the finite element method for the bivariate Roberts estimator where no aggregation has been made. We have used just four points of data, maintained the smoothing parameter, and reduced the size of the finite element mesh. Clearly, the functions become unbounded at the data points.

However, we notice that for coarser grids, the estimator is bounded. Compare this with the Roberts estimator using aggregation into a histogram on the finite element grid in Figure 7.3. We see that the estimates are comparable, although pointwise evaluation makes the estimate understandably rougher.

Clearly, we will no longer have asymptotic convergence; the kernel associated with the bivariate Roberts estimator will be unbounded at zero. What is different between this case, and that of first aggregating the data is that while allowing the mesh, including the histogram, to tend to zero will give us unbounded solutions for a given data set, simultaneously letting $n \to \infty$ will still provide convergence. Nevertheless, in practical situations it may well be possible to reduce the complexity of finite element subspaces in higher dimensions without having to resort to aggregating data. This will not produce an enormous computational saving, but it will allow us to match empirical moments exactly once again. We will also avoid the need to first develop a good histogram fit in order to reduce the size of the error. As bin width rules for histograms do not behave nicely, avoiding using them may turn out to be a significant advantage.

7.1.5 The Curse of Dimensionality

We have so far encountered problems with dimensionality in the form of the Sobolev embedding theorem. In fact, there are a number of factors which make working in higher dimensions difficult and we shall see that the Sobolev embedding theorem is one of our lighter concerns. Many of the major applications of density estimates occur in the area of data mining, in which datasets may possibly possess thousands of attributes, each of which may be considered as having its own dimension. It is therefore important to study the ways in which datasets in higher dimensions behave.

One of the first problems encountered in generalising to higher dimensions is the increase in computational cost in solving finite element systems. For a given finite element mesh size, using k bins in one dimension, the size of the corresponding linear system in d dimensions will be on the order of k^d. This will quickly lead to problems both in storing and solving systems on a large scale.

The real problems occur, however, in the geometry high dimensions. Many of our standard intuitions break down in high-dimensional cases and in particular there are significant implications for producing models such as densities. The essential problem here is that as the dimensionality of data sets grow, the data tends to become sparse. Many results concerning this phenomena exist, those we will discuss here may be found in [Sil86], which contains a demonstration of these problems as they involve density estimation.

A very simple demonstration of the phenomenon of sparseness can be given very neatly by considering the unit hypercube in d dimensions. We observe that the hypercube of side $1/2$ will represent $100/2^d\%$ of the volume of the unit hypercube in d dimensions. To make this more explicit, if we are working with a dataset which is uniformly distributed over the unit hypercube in ten dimensions, then a data point in the set will have probability 0.01 of falling into a box of side 0.63. There are many further demonstrations of the pathological nature of high-dimensional geometry which will not be considered here.

The consequence for density estimation in this case is that the tails of densities become important. Let us consider the normal distribution. In the univariate case, 90% of the mass of the distribution lies in the region $|x| \leq 1.6$. The same region contains less than 1% in ten dimensions. This not only means that regions of relatively high density will have few points in them, it will also mean that it will require a much larger amount of data to ensure the convergence of density estimates. We have already seen this in the dependence of optimal convergence rates on $n^{-\phi_d(m,p)}$. [Sil86] produces a study of the number of points required to produce a relative mean squared error at the origin $(E(\hat{f}(0) - f(0))^2/f(0)^2)$ of less than 0.1 using a Gaussian kernel and choosing the window width optimally for the estimation. While only four points were sufficient in the univariate case, for ten dimensions 842,000 were required.

There are no really effective ways of dealing with this phenomenon. An initial approach is to observe that the Sobolev embedding theorem requires us to strengthen the smoothing terms we use as we increase the dimension of the dataset. This, in some sense, counterbalances the sparseness of the data. At the same time, we can further smooth the data by increasing the size of the finite element mesh so as to capture more data in each element and to compensate for the increased computational cost. This solution is not particularly effective, however - the larger mesh sizes will not represent the data well and will tend to shift modes and smooth out potentially interesting features. The alternative is to reduce the number of dimensions we consider. This will be a reasonable procedure for the purposes of data exploration in which visualisation in high dimensions presents a difficulty, anyway. When we require a density that is created on higher dimensions for a classification scheme, for example, we can employ the densities created on lower dimensions along with some interaction terms between dimensions to try to approximate it. [HT90] provides a good overview of additive models. This will get us some way, but even then we will require increasingly higher-dimensional interaction terms as the dimension of the data increases.

7.2 Further Applications

We have made some demonstration of the applications of this method to data exploration and visualisation, and to the problem of classifying new data into separate subsets. There are, of course, many further applications in which a good estimate of density may be used. [Sil86][Chapter 6] provides an overview of many of the uses to which such functions may be put. Two interesting applications are worth examining in greater detail. The first of these is the problem of clustering, which might be thought of as the inverse of the classification problem. The second, related idea is in the definition of spatial regions in which points occur. This is somewhat more obscure as an application but the scalability of the methods described here make them well-suited to the problem and warrant some discussion.

7.2.1 Clustering

At the beginning of Chapter 1 the idea of a density was introduced with the motivation that some regions of space appear to contain more points than others. One natural thing to do following this observation is to ask whether or not we can identify which points are close together and which not. This yields two complementary problems: classifying points into separate groups and clustering, or forming groups. In the case we deal with clustering will be interpreted purely as spatial proximity but many other metrics may be used. There are myriads of clustering algorithms with many motivations: dividing sets of data, aggregating sets of data, moving centroids around and so forth. Density estimates are used, for the most part, in what can be described as hill-climbing problems.

Suppose we have a density estimate for some dataset. Clearly our intuition about there being clumps in the data states that we should look for regions where there is a large amount of data and which are separated from the rest of the data set. These will correspond to peaks in the density function. This glib statement tells us nothing about how to then go about producing clusters, but it can be usefully employed as motivation. Typically, we want to divide the dataset totally - in other words, we want to assign to each data point some cluster to which it belongs.

Let us define clusters by "peaks" in the density function. We wish to use this to divide the data set. There are two distinctive approaches to doing this are documented in [Sil86]. The first is to group data points hierarchically. We assign to each data point X_i a "parent", which is that point amongst its k (predetermined) nearest neighbours $\{X_{ik}\}_{j=1}^{k}$ for which

$$\frac{\hat{f}_n(X_{ik}) - \hat{f}_n(X_i)}{|X_{ik} - X_i|}$$

is greatest. Effectively we are taking the steepest path between data points. Clearly, we can consider the centroids of each cluster to be those whose parents are themselves. We can now determine the cluster to which a data point belongs by following its "ancestors" until we get to a "root".

The second approach is much more closely related to hill-climbing. The basic idea is to "move" each data point uphill until the data aggregates itself. We define a sequence X_i^m by constructing an estimate $\hat{f}_n^m$ on the set of data $\{X_i^m\}_{i=1}^{n}$ and then move the data points according to

$$X_i^{m+1} = X_i^m + \frac{a \nabla \hat{f}_n^m(X_i^m)}{\hat{f}_n^m(X_i^m)}$$

where a is some predetermined parameter. Here, the clumps re-enforce themselves by increasing the steepness of $\hat{f}_n^m$ at each step. We therefore expect to find the data becoming ever more tightly grouped into clumps. The mathematically pleasing properties behind this are that if $\hat{f}_n^m$ is a normal distribution and $a = \sigma^2$ its variance, then points are moved automatically to the mean of $\hat{f}_n^m$.

Clearly, the use of a finite element method will significantly speed up the calculations for either of these methods. The proposal here is to introduce what might be considered a variant on these by using the properties of the finite element approximation that we have made to a density spline. We observe that there will be a considerable computational saving if instead of working with the data points themselves we create a hierarchy of the nodes in the finite element mesh. We define the "parent" of each node to be that member of the set of its neighbours at which the density function is highest. It is now a relatively simple matter to associate each data point with the nearest, or the highest neighbouring, node and to then climb the hierarchy defined on the nodes. This could either be interpreted as moving the data points in, the motivation of the second method, or in creating a tree. In fact, if we use undersmoothed densities we could move points and then aggregate spurious clusters by creating a second density estimate on the shifted points using a smoother density. As with the broader set of clustering algorithms, the list of potential variations is endless.

7.2.2 Area Definition and Edge Detection

The problem of defining a geometrical region, or its edge, is not often regarded as being an important application of density estimates. There is some similarity in the problem, however, to the ideas we put forward in §7.2.1 above and we will outline a situation in which the techniques here may provide a reasonable solution to the problem. The original problem was suggested by [Cla98] but in fact the application can be discussed a little more generally.

Let us suppose that, having created some clusters in a data set, we wish to define a region into which the points in each cluster fall. A proposed method of defining such a region would be to take a density estimate for the region. We could now define the edge of the region in terms of that set of points on which the density takes a particular value. The choice of this value will depend, clearly, on the application required and the smoothness of the density estimate. Should we wish to capture as many points as possible, some relatively small value could be used. One suggestion might be to create a highly undersmoothed fit and declare the edge of the region to be those places where the density estimate becomes negative. Alternatively, if we are penalised more for placing points into a cluster in which they do not belong, we might want to choose a higher value as the cut-off point. One potential automatic choice of cut-off would be the smallest value that the density takes on the points in the data set. Having this density estimate also allows us to quickly determine whether or not a point is in the desired region.

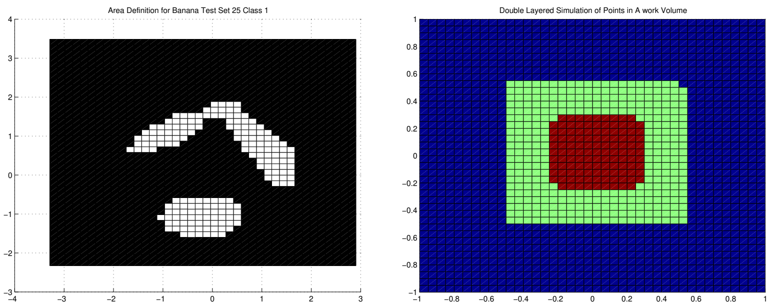


Figure 7.4: Area Definitions. Left: banana data set 25, class 1 defined as a region of $\mathbb{R}^2$. Right, a simulated 2-level listing of points in a robot work volume. Both volumes were produced using the Roberts esimator.

Of course, this theme may be varied indefinitely. We can oversmooth and use the values of the density to define fuzzy boundaries for a set. We can "layer" the density into a number of different regions and differentiate between areas of varying density and so forth. An example is given in Figure 7.4 of both the straightforward definition problem and a double-layered problem. As with the problem of clustering, we are able to make use of the finite element density to do this in an efficient numerical manner.

The concrete example which inspires this particular application derives from [Cla98]. The problem posed there is to determine the work-space of a robot arm. We are able to run the arm about its work-space and provide a large dataset of points which it will be able to reach. We now wish to use this dataset as a criteria to decide whether or not the arm is able to reach some other point. In this case, an undersmoothed density may be used with a small but positive cut-off value to give what will hopefully be a reasonable criteria for deciding whether a point is in the work-volume. The criteria may also be evaluated much more quickly than a search for nearby data points. A layered density would allow us to differentiate between regions which the robot arm can reach in only one way, versus those which may be reached using two, three, or more configurations.

7.3 Conclusions

The functions we have presented, in developing these density splines, represent an original and promising approach to the problem of density estimation. They maintain many of the properties of density estimates, which are important in applications. We have seen that we can regard density splines in the full domain as methods for choosing kernels, but that they generalise to bounded domains without sacrificing the properties above. In producing a finite element approximation to density splines we have been able to significantly increase the computational efficiency both in solving the systems required to gain an estimate and in evaluating the resulting function. In particular we have shown that we can implicitly enforce the non-negativity of the estimate and we can employ standard variations on finite element techniques to make the approximation exercise more manageable. We have demonstrated methods for choosing smoothing parameters and shown that the use of a finite element approximation makes this choice considerably faster than a naïve approach using other methods would produce.

Nevertheless, there remain a large number of unanswered questions concerning these methods. We have been unable to produce a proof of the convergence of these methods, except in special cases. While the evaluation of functions used to choose smoothing parameters has been made efficient by using finite element approximations, these approximations also introduce problems due to a lack of well-defined minima. A good theoretical treatment of the consequence of these approximation methods for such functionals is required to either produce alternative functionals or to justify the use of some other procedure than minimisation in the current ones. The relationship between the accuracy of the finite element approximation and the convergence of the method overall requires further study, as does an attempt to use the finite element method to avoid having to aggregate data before employing it in the density splines. Finally, the behaviour of these methods in high dimensions must be addressed. There appears to be a relationship between the problems encountered with the curse of dimensionality and the Sobolev embedding theorem. The extent to which they may counter each other is not clear. The effectiveness of the use of additive or multiplicative models in producing high-dimensional densities is also unexplored and could potentially produce promising results.

Despite the limitations of the analysis conducted so far, these density splines have performed well empirically, both in simulated and real-world exercises. They produce good visual pictures and therefore make a useful tool in data visualisation and exploration. They have also performed remarkably well in the problem of classifying data in synthetic trials. They have been able to find significant features in the vehicle insurance dataset, but have performed less well in classification exercises due to the limited nature of the particular algorithms employed. There are a great many more applications available to these estimates and we have seen their potential to be applied to the problems of clustering and in edge-detection and defining geometrical areas. In all of these, we are able to take advantage of the finite element approximation to increase the computational efficiency of the algorithms in a straightforward manner. Clearly, considerably more development needs to be put into the estimates. Nevertheless, the studies contained here indicate that they warrant that attention.

Appendix A

Glossary of Notations

Throughout this work, $\int_\Omega f(x)dx$ is to be interpreted as a volume integral over $\Omega \subseteq \mathbb{R}^d$ in which $x = (x_1, \ldots, x_d)$ is a vector. Where $\prod_{i=1}^n dx_i$ is used, we assume an integral over $\mathbb{R}^{nd}$ in which each x_i is d-dimensional. When we have employed function spaces such as $W_p^{(m)}(\Omega)$, the omission of a specific set Ω indicates that we are considering the space of functions on the whole of $\mathbb{R}^d$.

$\preceq$	For a_k and b_k dependent on some parameter k, $a_k \preceq b_k$ is equivalent to $a_k \leq Cb_k$ for some constant C.
$\lvert\alpha\rvert$	For $\alpha = (\alpha_1, \ldots, \alpha_k)$, $\lvert\alpha\rvert = \alpha_1 + \ldots + \alpha_k$. This notation suppresses k.
$D^\alpha f(x)$	$\frac{\partial^{\lvert\alpha\rvert} f(x)}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$.
ξ^α	$\xi_1^{\alpha_1} \ldots \xi_d^{\alpha_d}$.
$\mathcal{L}^p(\Omega)$	The space of all p-integrable functions defined on Ω.
$W_p^{(m)}(\Omega)$	The space of all functions defined on Ω whose derivatives up to order $m-1$ are all uniformly continuous and whose m^{th} derivatives are in $\mathcal{L}^p(\Omega)$.
$\lvert \cdot \rvert_{W_p^{(m)}(\Omega)}$	The semi-norm on $W_p^{(m)}(\Omega)$ defined by $\lvert f\rvert^2_{W_p(m)(\Omega)} = \sum_{\lvert\alpha\rvert=m} \lvert\lvert D^\alpha f\rvert\rvert^2_{\mathcal{L}^p(\Omega)}$.
$B_{W_p^m}(M)$	The subset of $W_p^{(m)}(\Omega)$ consisting of all functions f such that $\lvert f\rvert_{W_p^{(m)}(\Omega)} \leq M$. Ω has been suppressed here.
$H^m(\Omega)$	The space of functions defined on Ω whose derivatives up to order m are square-integrable. This can also be thought of as $W_2^{(m)}(\Omega)$.
$\lvert \cdot \rvert_{H^m(\Omega)}$	$\lvert \cdot \rvert_{W_2^{(m)}(\Omega)}$.

$\|\| \cdot \|\|_{H^m(\Omega)}$	$\left(\sum_{\alpha=0}^{m} \| \cdot \|^2_{H^\alpha(\Omega)}\right)^{1/2}$.
Y^*	For a Banach space Y, the space of all bounded linear functionals on Y.
V_h	A finite element subspace parametrised by $h \in H$; a set of numbers with accumulation point 0.
T_h	The finite element discretisation of Ω associated with V_h. Typically h is the longest side of an element in T_h.
$\frac{(\xi h \cdot \nabla)^j}{j!} f(x)$	The j^{th} term in the d-variate Taylor expansion of $f(x + \xi h)$. This is less compactly expressed as $\frac{h^j}{j!} \sum_{\|\alpha\|=j} \xi^\alpha D^\alpha f(x)$.
$R_n(f; x, x + \xi h)$	The n^{th} degree Taylor remainder in the expansion of $f(x + \xi h)$ $$R_n(x, x + \xi h) = \sum_{\|\alpha\|=n} \frac{(\xi h)^\alpha}{(n-1)!} \int_0^1 (1-t)^{n-1} D^\alpha f(x + t\xi h) dt.$$

Appendix B

Gâteaux Derivatives

The results used in §1.3.1 rely on the Gâteaux differentiability of the functional T_x with respect to variation of the X_i. Formally, the Gâteaux derivative of a functional $T : X \to Y$ at the point z in the direction w is defined as:

$$DT(z)[w] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[T(z + \epsilon w) - T(z) \right]. \tag{B.1}$$

Here $DT(z)[w] \in Y$, $DT(z) : X \to Y$, and $DT : X \times X \to Y$ are all possible interpretations of this derivative. This can be thought of a generalisation of the directional derivative in $\mathbb{R}_n$:

$$\nabla f(x) \cdot y = \lim_{\epsilon \to 0} \frac{f(x + \epsilon y) - f(x)}{\epsilon}$$

and in fact [Tap71] argues that this is a more natural generalisation of differentiation to spaces other than the real line - partial derivatives only making sense in finite-dimensional spaces.

In fact, this is somewhat different from the definition of the kernel in Theorem 1.3.1. We use there what might be thought of as a convex Gâteaux derivative and assume that that particular limit is everywhere defined. It has been employed for the purposes of the result, but it also allows us to restrict T_x to be defined on distribution functions only. The properties we require in the proof of Theorem 1.3.1 all depend on the standard definition of Gâteaux derivatives, however, and can be shown here. In particular, we require the result of the following theorem:

Theorem B.0.1. *Let X,Y be normed linear spaces and suppose that $T : X \to Y$ is Gâteaux differentiable and*

1. *$DT(z)[\cdot]$ is continuous at zero for each $z \in X$.*
2. *$DT(\cdot)[w]$ is continuous for each $w \in X$.*

Then $Df(z)$ is a bounded linear functional.

In particular, the continuity of T_x guarantees that (1) and (2) are satisfied. The first of these is easy to see by noting that for $||g_i|| \downarrow 0$, $T(z + kg_i) \to T(z)$. The second follows from observing that for $z_i \to z$, $T(z + \epsilon g) - T(z) - T(z_i + \epsilon g) + T(z_i) \to 0$.

The proof of the theorem follows [Tap71]. We begin with two lemmata:

Lemma B.0.1. *The Gâteaux derivative of T at z is a homogeneous operator:*

$$DT(z)[\alpha w] = \alpha DT(z)[w].$$

Proof. Replace ϵ by $\alpha\epsilon$ in (B.1) and observe:

$$DT(z)[w] = \lim_{\epsilon \to 0} \left[\frac{T(x + \epsilon\alpha w) - T(x)}{\alpha\epsilon} \right] = \frac{1}{\alpha} DT(z)[\alpha w].$$

□

We also require the somewhat more complicated result:

Lemma B.0.2. *Consider $T : X \to Y$ to be Gâteaux differentiable at each point $z + t(w - z)$ in the direction $w - z$ for $0 \leq t \leq 1$ and given z and w. Then for each $\delta \in Y^*$ we have that*

$$\delta(T(z) - T(w)) = \delta\left(DT(x + \theta(w - z))[w - z]\right)$$

for some $0 < \theta < 1$ and

$$||T(z) - T(w)|| \leq \sup_{0<\theta<1} ||DT(z + \theta(w - z)[w - z]||.$$

Proof. Let $g(t) = \delta(T(z + t(w - z))$. Then we clearly have that

$$g'(t) = \delta\left(D(z + t(w - z))[w - z]\right)$$

so by the mean value theorem, $g(1) - g(0) = g'(\theta)$ for some $0 < \theta < 1$, giving us the first result. Now we make use of the Hahn-Banach theorem to choose $\delta \in Y^*$ with $||\delta|| = 1$ and $\delta(T(z) - T(w)) = ||T(z) - T(w)||$ which gives us the result. □

Now we can go on to prove the theorem:

Proof of Theorem B.0.1. By (B.0.1) we have that $D(z)[0] = 0$. Using Lemma 1 we know that there exists an $r > 0$ such that $||DT(z)[w]|| \leq 1$ whenever $||w|| \leq r$ and therefore

$$||DT(z)[w]|| = \left\| \frac{||w||}{r} DT(z) \left[\frac{rw}{||w||} \right] \right\| \leq \frac{1}{r} ||w||$$

and we therefore only need to show additivity. Let us suppose $w_1, w_2 \in X$. Take $\kappa > 0$ then there exists a $\tau > 0$ such that

$$\begin{aligned}&||DT(z)[w_1+w_2]-DT(z)[w_1]-DT(z)[w_2]\\&\quad-\left[\frac{T(z+\epsilon w_1+\epsilon w_2)-T(z)}{\epsilon}\right]+\left[\frac{T(z+\epsilon w_1)-T(z)}{\epsilon}\right]+\left[\frac{T(z+\epsilon w_2)-T(z)}{\epsilon}\right]||\\&\quad\leq\kappa.\end{aligned}$$

for $\epsilon<\tau$. So that we have:

$$\begin{aligned}&||DT(z)[w_1+w_2]-DT(z)[w_1]-DT(z)[w_2]||\\&\quad\leq\frac{1}{|\epsilon|}||T(z+\epsilon w_1+\epsilon w_2)-T(z+\epsilon w_1)-T(z+\epsilon w_2)+T(z)||+\kappa\end{aligned}$$

Now using Lemma B.0.2 we can find a $\delta\in Y^*$ such that $||\delta||=1$ and

$$\begin{aligned}&||T(z+\epsilon w_1+\epsilon w_2)-T(z+\epsilon w_1)-T(z+\epsilon w_2)+T(z)||\\&=\quad\delta(T(z+\epsilon w_1+\epsilon w_2)-T(z+\epsilon w_1))-\delta(T(z+\epsilon w_2)-T(z))\\&=\quad\epsilon\delta\left[DT(z+\epsilon w_1+\theta_1\epsilon w_2)[w_2]\right]-\epsilon\delta\left[DT(z+\theta_2\epsilon w_2)[w_2]\right]\\&\leq\quad|\epsilon|||DT(z+\epsilon w_1+\theta_1\epsilon w_2)[w_2]-DT(z)[w_2]||\\&\qquad+|\epsilon|||DT(z)(w_2)-DT(z+\theta_2\epsilon w_2)[w_2]||\\&\leq\quad 2|\epsilon|\kappa\end{aligned}$$

by (2) for ϵ sufficiently small. It follows that

$$||DT(z)[w_1+w_2]-DT(z)[w_1]-DT(z)[w_2]||\leq 3\kappa$$

and we therefore have that $DT(z)$ is additive, giving us the required result. □

Appendix C

Constructions for Convergence Rates

Here we produce the construction of the function $e_{k\delta}$ used in the proof of Theorem 1.3.2. The function used here departs from that given in [Wah75b] and [Far72]. They define an explicit univariate function $g_{k\delta}$ using a recursive definition, taking successive translations of integrals of an original $C^0(\mathbb{R})$ function. [Wah75b] does not consider the multivariate case. [Far72] extends to several variables with the definition

$$e_{k\delta}(x) = \frac{1}{d}\sum_{i=1}^{d} g_{k\delta}(x_i).$$

However, his proofs rely on $e_{k\delta}$ having bounded support, which is clearly false.

We present an alternative function which is much more simply defined even when we give it explicitly and will provide the result in higher dimensions. The requirements we have are that $e_{k\delta}$ must be small in the sense of having a small $\mathcal{L}^2$ norm, have a large "peak" at 0 but also have $|e_{k\delta}|_{W_p^{(k)}} \leq M/2$. In particular, we need $e_{k\delta}$ to be parametrised by δ so that it may vary with M without increasing its support or $\mathcal{L}^2$-norm.

We begin with a function $g_k \in W_p^{(k)}$ defined to satisfy

$$\int_{-\infty}^{\infty} g_k(x)dx = 0 \tag{C.1}$$

$$\max_x |g_k(x)| = g_k(0) = 1 \tag{C.2}$$

$$\text{supp} g_k \subseteq [-1, 1]. \tag{C.3}$$

Such a function can very easily be produced explicitly. Let b_k denote the $(k+1)^{\text{st}}$ order B-spline defined on a uniform grid with support $[-1/3, 1/3]$. It will take its maximum at 0. Then we let

$$g_k(x) = \frac{b_k(x) - b_k(x - 2/3)}{b_k(0)}$$

and g will satisfy all the conditions above. We can now proceed to define our modifying function:

$$e_{k\delta}(x) = \delta^{k-d/s} \prod_{i=1}^{d} g_k\left(\frac{x_i}{\delta}\right) = \delta^{k-d/s} G\left(\frac{x}{\delta}\right).$$

Clearly, we have

$$\max_x |e_{k\delta}(x)| = e_{k\delta}(0) = \delta^{k-d/s}$$

and using (C.3) we see that

$$\mathrm{supp} e_{k\delta} \subseteq \{x : ||x||_\infty \leq \delta\},$$

which combine to give us:

$$\int_{\mathbb{R}^d} e_{k\delta}(x)^2 dx \leq 2^d \delta^{2(k-d/s)+d}.$$

It is also clear that

$$\int_{\mathbb{R}^d} e_{k\delta}(x) dx = \delta^{k-d/s} \prod_{i=1}^{d} \int_{-\infty}^{\infty} g_k(x_i) dx_i = 0$$

from (C.1). We now only need to check $|e_{k\delta}|_{W_p^{(k)}}$. In the following, $G_\delta : x \to G(x/\delta)$ has been used to clarify the functions to which differential operators are applied.

$$\begin{aligned}
\int_{\mathbb{R}^d} |D^k e_{k\delta}(x)|^p dx &= \delta^{p(k-d/s)} \int_{\mathbb{R}^d} |D^k G_\delta(x)|^p dx \\
&= \frac{\delta^{p(k-d/s)}}{\delta^{pk}} \int_{\mathbb{R}^d} |D^k G\left(\frac{x}{\delta}\right)|^p dx \\
&= \delta^{-pd/s} \delta^d \int_{\mathbb{R}^d} |D^k G(y)|^p dy \\
&= K\delta^{d(1-p/s)}
\end{aligned}$$

with the substitution $y = x/\delta$. Here we know that each $g_k(x_i) \in W_p^{(k)}$ and therefore so is G. D^k is regarded as a general k^{th} order differential operator. We will subsume the summation of all such operators acting on $e_{k\delta}$ into K. In order to check the case for $p = s = \infty$ it is enough to observe that b_k is $C^k(\mathbb{R})$ and therefore so is g_k. Doing the calculations we will get $|e_{k\delta}|_{W_\infty^{(k)}} \leq K$ for some K. Multiplying $e_{k\delta}$ by a constant depending on M in Theorem 1.3.2 will provide the result.

Appendix D

Reproducing Kernel Hilbert Spaces

One of the main pieces of machinery used in spline fitting and other nonparametric analysis is that of the reproducing kernel Hilbert space. We aim to give a brief overview of the theory of these spaces in order to provide some useful results and to add to the brief descriptions given at the beginning of Chapter 2. The results found here are largely covered in [Wah90][Chapter 1].

Let H be a Hilbert space of real-valued functions on a space V such that pointwise evaluation, $L_t f \to f(t)$, is a bounded linear functional. Then we say that H is a reproducing kernel Hilbert space. We note that many standard Hilbert spaces will not be reproducing kernel spaces - L^2 is not, for example. However, if $2m \geq d$ the Sobolev embedding theorem gives us that $H^m(\Omega) \subset C^0(\overline{\Omega})$ by a compact injection. We therefore have that functional evaluation is bounded in $H^m(\Omega)$ and it is a reproducing kernel Hilbert space.

Having produced such a space, we let R_t be the representor of evaluation at t: $L_t f =< R_t, f >$ by a theorem of Riesz. Now we can define a function R on $V \otimes V$ by $R(s,t) =< R_s, R_t >$. This will be known as the reproducing kernel of H. The reasons behind the name are now obvious.

Let us define a function $Q : V \otimes V \to \mathbb{R}$ to be positive definite if for all $a_1, \ldots, a_n \in R$ and all $t_1, \ldots, t_n \in V$,

$$\sum_{i,j=1}^{n} a_i a_j Q(t_i, t_j) \geq 0.$$

Then it is clear that

$$\begin{aligned} \sum_{i,j=1}^{n} a_i a_j R(t_i, t_j) &= \sum_{i,j=1^n} a_i a_j < R_{t_i}, R_{t_j} > \\ &= \| \sum_{i=1}^{n} a_i R_{t_i} \|^2 \\ &\geq 0 \end{aligned}$$

and R is positive definite. In fact, any positive definite, symmetric function on a space $V \otimes V$ will be called a reproducing kernel and we can associate a reproducing kernel space with it. This can be done by constructing the Hilbert space H_R out of Span$\{R_t(\cdot) = R(t, \cdot)\}_{t \in V}$ with the inner product:

$$\left\langle \sum_{i=1}^{m} a_i R_{t_i}, \sum_{j=1}^{m} b_j R_{s_j} \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j R(t_i, s_j),$$

which is well-defined due to the positive definiteness of R and we can easily see that $< R_t, f >= \sum_{i=1}^{n} a_i R(t, t_j) = f(t)$.

A particularly nice property of reproducing kernel Hilbert spaces is that norm convergence implies pointwise convergence:

$$|f_n(t) - f(t)| = | < f_n - f, R_t > | \leq ||f_n - f||||R_t||.$$

There are further results in the spectral theory of reproducing kernels, which have been used to create orthogonal series estimators and so forth. In our particular case, we are interested in spline approximation in which reproducing kernel Hilbert spaces are a useful tool. We now will demonstrate the construction of reproducing kernels from particular penalty terms as were used in §2.1.

We will begin by motivating the development from a univariate example. Consider $H^m([0, 1])$. This is a reproducing kernel Hilbert space for any $m \geq 1$. We can express any function $f \in H^m([0, 1])$ by the Taylor expansion:

$$f(x) = f(0) + \sum_{i=1}^{n} f^{(i)}(0) + \int_0^1 \frac{(x-t)_+^{(m-1)}}{(m-1)!} f^{(m)}(t) dt.$$

Then for functions in the space H_1 defined as satisfying $f^{(i)}(0) = 0$, $i = 0, \ldots, m-1$ we have:

$$f(x) = \int_0^1 \frac{(x-t)_+^{(m-1)}}{(m-1)!} f^{(m)}(t) dt$$

and H_1 is a Hilbert space with norm $\int_0^1 f^{(m)}(x)^2 dx$. Now let

$$G(x, t) = \frac{(x-t)_+^{(m-1)}}{(m-1)!}$$

be the Green's function satisfying $\int_0^1 G(x, t) f^{(m)} dt = f(x)$. Then the reproducing kernel can be found as

$$R(x, y) = \int_0^1 G(x, t) G(y, t) dt.$$

This can be shown by observing that

$$\langle R_x, f\rangle = f(x) = \int_0^1 G(x,t) f^{(m)}(t)dt$$

and therefore $R_x^{(m)}(y) = G(x,y)$. We can construct a reproducing kernel for all of $H^m([0,1])$ by taking H_0 to be the Hilbert space of polynomials spanned by the orthonormal basis $\phi_1, \dots, \phi_m$ and endowing it with the inner product

$$\langle f, g\rangle = \sum_{i=0}^{m-1} f^{(i)}(0) g^{(i)}(0).$$

Then it is clear that $H^m([0,1]) = H_0 \oplus H_1$ when it is endowed with the norm:

$$||f||^2 = \sum_{i=0}^{m-1} f^{(i)}(0)^2 + \int_0^1 f^{(m)}(x)^2 dx$$

and H_0 and H_1 are orthogonal. In fact this norm is equivalent to the usual norm on $H^m([0,1])$ and we can find a reproducing kernel for the whole space as:

$$R(x,y) = \sum_{i=1}^{m} \phi_i(x)\phi_i(y) + \int_0^1 G(x,t)G(y,t)dt.$$

In the general case, we are interested in the penalty term

$$C(u,v) = \sum_{|\alpha|,|\beta|=m} \int_\Omega a_{\alpha\beta}(x) D^\alpha u(x) D^\beta v(x) dx,\ a_{\alpha\beta} \in \mathcal{L}^\infty(\Omega).$$

En Lemma 2.3.1 we demonstrated that $C(\cdot,\cdot)^{1/2}$ is equivalent to $||\cdot||_{H^m(\Omega)}$ on the subspace:

$$H_1 = \{u \in H^m(\Omega) : \int_\Omega D^\alpha u(x)dx = 0, |\alpha| \le m\}.$$

Define Λ to be the differential operator

$$\Lambda = (-1)^m \sum_{|\alpha|,|\beta|=1} D^\beta \left(a_{\alpha\beta} D^\alpha\right)$$

and let $G(x,t)$ be the Greens function satisfying

$$\int_\Omega f(t) \Lambda_x G(x,t) dt = f(x)$$

with appropriate boundary conditions so that $C(f, G(x,\cdot)) = f(x)$. In this case $R_x(\cdot) = G(x,\cdot)$ and the reproducing kernel is $R(x,y) = C(G(x,\cdot), G(y,\cdot))$.

Turning to the null space of $C(u,u)$, H_0, we can write any $f \in H^m(\Omega)$ as $f = u_0 + q$ for some $q \in H^1$. Let us endow H_0 with the inner product

$$\langle p, q \rangle = \sum_{|\alpha|<m} \int_\Omega p(x)dx \int_\Omega q(x)dx.$$

If we iterate the inequality

$$\int_\Omega u(x)^2 dx \preceq \sum_{|\alpha|=1} \int_\Omega D^\alpha u(x)^2 dx + \left[\int u(x)dx \right]^2$$

it is easy to see that the associated norm is equivalent to $||\cdot||_{H^m(\Omega)}$. We now have an equivalent norm on $H^m(\Omega)$ and once again H_0 and H_1 are perpendicular. The reproducing kernel follows as above with $p_1, \ldots, p_n$ an orthonormal basis with respect to the new inner product. In doing this we have assumed that $H^m(\Omega)$ is a reproducing kernel Hilbert space. Were this not the case, $R(x,y)$ may be unbounded. We have also suppressed an assumption here that $C(\cdot,\cdot)$ is symmetric. If this is violated, we will have two one-sided reproducing kernels and the theory becomes more complicated.

Bibliography

[BB79] Jacques Bleuez and Denis Bosq. Conditions nécessaires et suffisantes de convergence de l'estimateur de la densité par la méthode des fonctions orthogonales. *Revue Roumaine de Mathémathiques Pures et Appliquées*, 24(6):869 – 886, 1979.

[BF91] Franco Brezzi and Michel Fortin. *Mixed and Hybrid Finite Element Methods*. Springer Series in Computational Mathematics. Springer-Verlag, New York-Berlin-Heidelberg-London-Paris-Tokyo-Hong kong-Barcelona, 1991.

[BKS71] Liliana I. Boneva, David Kendall, and Ivan Stefanov. Spline transformations: Three new diagnostic aids for the statistical data-analyst. *Journal of the Royal Statistical Society*, 33:1 – 17, 1971.

[Cia75] Ph. Ciarlet. *Lectures on the Finite Element Method*. Tata Institute of Fundamental Research, Bombay, 1975.

[Cla98] Andrew Clark, 1998. Personal communication.

[DG85] Luc Devroye and László Györfi. *Nonparametric Density Estimation: The $\mathcal{L}^1$ View*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York-Chichester-Brisbane-Toronto-Singapore, 1985.

[Duc76] Jean Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In W. Schempp and K. Zeller, editors, *Constructive Theory of Several Variables, Proceedings of the 1996 Oberwalch Conference*, pages 85 – 100, Berlin, 1976. Springer-Verlag.

[DW82] Nira Dyn and Grace Wahba. On the estimation of functions of several variables from aggregated data. *SIAM Journal of Mathematical Analysis*, 13(1):134 – 152, 1982.

[Dyn99] Nira Dyn, 1999. Personal communcation.

[Far72] R. H. Farrell. On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *The Annals of Mathematical Statistics*, 43(1):170 – 180, 1972.

[Fri94] Jerome H. Friedman. Flexible metric nearest neighbor classification. Technical report, Deptartment of Statistics, Stanford University, 1994.

[Gri85] P. Grisvard. *Elliptic Problems in Nonsmooth Domains.* Monographs and Studies in Mathematics. Pitman, Boston-London-Melbourne, 1985.

[GT77] D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order.* Grundlerhren der mathematischen Wissenschaften. Springer-Verlag, Berlin-Heidelberg-New York-Tokyo, 1977.

[HRA97] M. Hegland, S. Roberts, and I. Altas. Finite element thin plate splines for surface fitting. In John Noye, Michael Teuber, and Andrew Gill, editors, *Computational Techniques and Applications: CTAC97*, pages 289 – 296, Singapore-New Jersey-London-Hong Kong, 1997. Wold Scientific.

[HT90] T. J. Hastie and R. J. Tibshirani. *Generalised Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, London-New York-Tokyo-Melbourne-Madras, 1990.

[MRW+99] Sebastian Mika, Gunnar Rätsch, Jason Weston, Bernhard Schölkopf, and Kaus-Robert Müler. Fisher discriminant analysis with kernels. In *IEEE Workshop on Neural Networks for Signal Processing*, 1999.

[Par62] Emanuel Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065 – 1076, 1962.

[RHA99] S. Roberts, M. Hegland, and I. Altas. An H^1 finite element approximation of a thin plate spline. Preprint, 1999.

[Sco92] David W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York-Chichester-Brisbane-Toronto-Singapore, 1992.

[Sil86] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Monographs on Statistics and Applied Probability. Chapman and Hall, London-New York, 1986.

[Tap71] R. A. Tapia. *The Differentiation and Integration of Nonlinear Operators*, pages 45 – 102. Academic Press, New York-London, 1971.

[Tri78] Hans Triebel. *Interpolation Theory, Function Spaces, Differential Operators.* North Holland Mathematical Library. North Holland Publishing Company, Amsterdam-New York-Oxford, 1978.

[TS92] George R. Terrell and David W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236 – 1265, 1992.

[TT90a] Richard A. Tapia and James R. Thompson. *Nonparametric Probability Density Estimation.* Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore and London, 1990.

[TT90b] James R. Thompson and Richard A. Tapia. *Nonparametric Function Estimation, Modeling, and Simulation.* Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[Wah75a] Grace Wahba. Interpolating spline methods for density estimation I. equi-spaced knots. *The Annals of Statistics*, 3(1):30 – 48, 1975.

[Wah75b] Grace Wahba. Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. *The Annals of Statistics*, 3(1):15 – 29, 1975.

[Wah81] Grace Wahba. Numerical experiments with the thin plate histospline. *Communications in Statistics - Theory and Methods*, 24:2475 – 2514, 1981.

[Wah90] Grace Wahba. *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathemtics. Society for Industrial and Applied Mathematics, Philadelphia, 1990.