Practical Aspects Of Kernel Smoothing For Binary Regression And Density Estimation

Thesis

Practical Aspects of Kernel Smoothing for Binary Regression and Density Estimation

David F. Signorini, M.A.(Hons.)

Thesis submitted for the degree of Doctor of Philosophy

Statistics

July 14, 1998

Acknowledgements

This thesis has spanned a considerable period of time, and there are many people who have helped me along the way.

Daniel Lunn receives both credit and blame for beginning the whole process, first by sending me into exile in Australia to become a statistician, then helping me to return to the Open University. Don McNeil and John Simes are responsible for my initial determination to pursue a higher degree.

Charles Warlow, my current head of department, has offered both moral and practical support with the difficult task of juggling a full-time medical statistics job with a part-time Ph.D., and without his encouragement I would be many months or even years from completion.

Chris Jones, my supervisor, has been a constant source of inspiration and help both in person when I was in Milton Keynes, and by email, post and telephone during my time in Edinburgh. His criticism has always been constructive, his suggestions have invariably turned out to be improvements, and he has been very, very patient.

Finally, my family and friends have always offered words of encouragement and never nagged. Above all, my wife Kate has endured a seemingly endless succession of late evenings and working weekends with admirable forbearance. I began this before we met, I was still doing it when we were married, and if it is now all over she deserves a great deal of the credit, and much more attention.

Practical Aspects of Kernel Smoothing for Binary Regression and Density Estimation

David F. Signorini

This thesis explores the practical use of kernel smoothing in three areas: binary regression, density estimation and Poisson regression sample size calculations.

Both nonparametric and semiparametric binary regression estimators are examined in detail, and extended to two bandwidth cases. The asymptotic behaviour of these estimators is presented in a unified way, and the practical performance is assessed using a simulation experiment. It is shown that, when using the ideal bandwidth, the two bandwidth estimators often lead to dramatically improved estimation. These benefits are not reproduced, however, when two general bandwidth selection procedures described briefly in the literature are applied to the estimators in question. Only in certain circumstances does the two bandwidth estimator prove superior to the one bandwidth semiparametric estimator, and a simple rule-of-thumb based on robust scale estimation is suggested.

The second part summarises and compares many different approaches to improving upon the standard kernel method for density estimation. These estimators all have asymptotically 'better' behaviour than the standard estimator, but a small-sample simulation experiment is used to examine which, if any, can give important practical benefits. Very simple bandwidth selection rules which rely on robust estimates of scale are then constructed for the most promising estimators. It is shown that a particular multiplicative bias-correcting estimator is in many cases superior to the standard estimator, both asymptotically and in practice using a data-dependent bandwidth.

The final part shows how the sample size or power for Poisson regression can be calculated, using knowledge about the distribution of covariates. This knowledge is encapsulated in the moment generating function, and it is demonstrated that, in most circumstances, the use of the empirical moment generating function and related functions is superior to kernel smoothed estimates.

Contents

List of Figures

Introduction

This thesis explores three main areas of application of kernel smoothing; binary regression, density estimation and generalised linear model power and sample size calculations.

Part I investigates the kernel smoothing approach to binary regression, with the emphasis on calibration (estimating the probability function equally well over a range of covariate values) rather than discrimination (classifying cases into successes or failures). In Chapter 1 the problem is defined, previous simple approaches to solution are described, and extensions to estimators with two bandwidths rather than one are derived. A simulation experiment is used to assess the practical performance of the estimators, in addition to the theoretical derivation of their asymptotic behaviour.

Local polynomial approaches to the problem, which have become very popular in recent years, are discussed in Chapter 2, and it is shown that the estimators of Chapter 1 are a special case of these more complicated estimators. Chapter 3 extends these estimators to the two bandwidth case. Once again, a simulation experiment is used to compare practically the various estimators of this and the previous chapters.

To separate the problem of estimator choice from that of bandwidth selection, the simulation experiments which are used to compare the estimators were conducted under a "best-case" scenario, whereby the performance was

assessed using the optimal bandwidth in each case. Chapter 4 addresses the issue of bandwidth selection, to see if the promising results of the previous chapters can be realised in practice using a data-dependent bandwidth selection procedure. Two general approaches which have been suggested in the literature but never followed through in detail or compared to each other are taken, and the results are used to make some general recommendations for the use of these estimators in practice.

Part II considers the use of more complex kernel density estimators for use when the target density is not Gaussian. Many different improvements to the standard kernel density estimator are described in Chapter 5, and once again a simulation experiment is used to compare their practical performance. Chapter 6 applies a simple approach to bandwidth selection to the more promising of these higher order estimators and draws some more general conclusions about whether the methodological development of kernel density estimation should concentrate on either improved estimators or improved bandwidth selection procedures for existing methods.

Finally, Part III explores a method of calculating sample size or power for Poisson regression models. Chapter 7 outlines the procedure and demonstrates its reliance upon the moment generating function of the distribution of covariates, and Chapter 8 briefly discusses how this function can be estimated using kernel smoothing methods.

A reduced version containing the main ideas of Chapter 7 has been previously published in *Biometrika*, 1991, with Signorini as the sole author. The majority of the simulation results of Chapter 5 were part of a joint publication of Signorini and Jones in *The Journal of the American Statistical Association*, 1997, which considers both the theoretical and practical merits of most of the estimators discussed.

Part I

Binary Regression

Chapter 1

Nonparametric Binary Regression

1.1 Introduction

The problem of binary regression, modelling the relationship between a dichotomous response and a set of covariates, is a widely used statistical technique, especially in the areas of medical statistics, biostatistics and epidemiology (see Collett [1] or Cox and Snell [2] for numerous examples). One of the most common methods of analysis used in practice is logistic regression, a special case of the generalized linear model (McCullagh and Nelder [3]). This, and other related methods such as probit analysis, are fully parametric and assume a linear relationship between some function of the response, in this case the inverse logit function, and the covariates:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta^T X,$$

where p is the probability of success for an individual with covariate vector X, and β is the vector of parameters which must be estimated. The logit function may be replaced by any monotonic function which maps $[0, 1]$

to the real line, such as the inverse of the Gaussian distribution function (probit analysis) or $[\log(-\log(1-p)]$, the complementary log-log transformation. Whichever function is chosen, however, the method still requires an assumption that the linear relationship does indeed hold and, as so often in statistical modelling, if it does not, then the model can give misleading results.

Furthermore, when using these parametric models, it is often desirable to examine such linearity assumptions. Consider for example multiple linear regression. To check linearity for each covariate, we can produce a scatterplot of each variable against the response. What is the analogous plot for multiple logistic regression? Figure 1.1 shows a typical such example, using data taken from Fan, Heckman and Wand [4]. The response is coded as 1 for survival and 0 for death, and the covariate is a transformation of the area of third degree burns for 435 patients admitted to the University of California General Hospital Burns Centre. The actual survival values are exactly 0 or 1, but the points on the plot have been jittered vertically to show repeated values. Is a logistic regression model appropriate for this dataset? The problem is obvious. It is virtually impossible to fit by eye a definitive curve to this data. In the linear case, when faced with two continuous variables, it is reasonably easy to spot departures from the model, but here the fact that each response is either 0 or 1 compromises that ability. Copas [5] pointed this out and went on to propose a nonparametric smoothing estimate of the probability of survival as an objective way of calculating a fitted value for the conditional probability of survival given the covariate for this kind of data.

This section builds upon the work of Copas, and others, to examine various kernel-based methods of smoothing such data. In this chapter, we

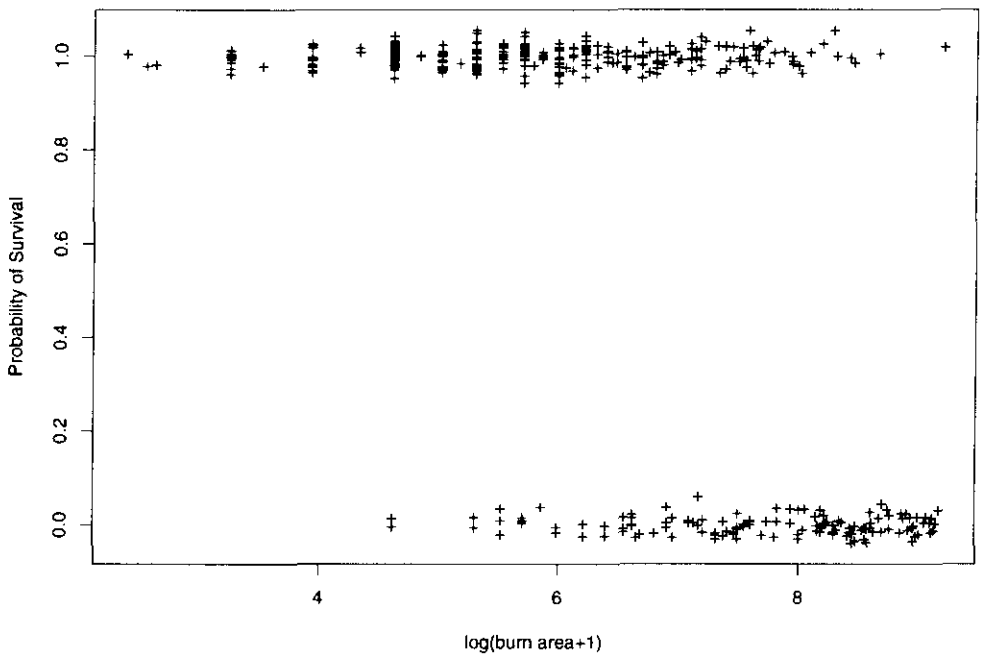

Figure 1.1: Survival versus Log(Burn Area + 1)

consider fully nonparametric methods of binary modelling, based upon kernel density estimates and their ratios. Asymptotic biases and variances and integrated mean square errors are derived and compared for the estimators, and their practical small-sample performance is assessed by means of a simulation experiment. In the next chapter, we look at semiparametric methods, where a kernel weighted quasi-likelihood is used in place of the fully parametric likelihood. These semiparametric estimators are compared to the fully nonparametric methods once again both theoretically and with a practical simulation experiment.

We begin by defining the notation to be used.

We are interested in studying the relationship between a binary variable Y and a (for the moment) single covariate X. We wish to estimate the

conditional probability

$$\lambda(x) = \text{Prob}(Y = 1 \mid X = x).$$

If we assume that the population consists of two sub-populations, the 'successes' for which $Y = 1$ and the 'failures' for which $Y = 0$, then denote the probability density function in the covariate space for the successes by $f(x)$, for the failures by $g(x)$ and for the whole population by $h(x)$. If the proportion of successes in the population is π_1, and $\pi_2 = 1 - \pi_1$ then

$$h(x) = \pi_1 f(x) + (1 - \pi_1)g(x) = \pi_1 f(x) + \pi_2 g(x).$$

Thus for any x we have

$$\lambda(x) = \frac{\pi_1 f(x)}{h(x)} \tag{1.1}$$

This equation forms the basis for the nonparametric approach to the problem. Parallels can be drawn between equation (1.1) and discriminant analysis, where it is the ratio f/g rather that f/h which is important for classification purposes, as initially described by Fix and Hodges [6, 7, 8]. There is a vast literature on the subject of nonparametric discrimination, much of which is reviewed by Ripley [9]. This work, however is focused on regression rather than classification, and so we shall measure performance by the accuracy of the estimation of $\lambda(x)$ for all values of x, in contrast to the discrimination problem, where the aim is to minimise the costs of misclassification.

To perform the estimation, assume we have a sample S of size s of response-covariate pairs $(Y_i, X_i); i = 1, \ldots, s$. Assume that the sample has been arranged so that the first m pairs are successes ($Y_i = 1$) and the last $n = s - m$ pairs are failures ($Y_i = 0$). Write $X_1, \ldots, X_m$ as $W_1, \ldots, W_m$ and $X_{m+1}, \ldots, X_s$ as $Z_1, \ldots, Z_n$. Note that $W_i = Y_i X_i$ and $Z_i = (1 - Y_{m+i})X_{m+i}$.

Thus we have $\{W_i\}$, a random sample from f, $\{Z_i\}$ a random sample from g and $\{X_i\}$ a random sample from the mixture density h.

1.2 Nonparametric Estimators

It is immediately obvious that we can write equation (1.1) as

$$\lambda(x) = \frac{\pi_1 f(x)}{\pi_1 f(x) + \pi_2 g(x)},$$

and if we replace π_1 and π_2 with the empirical estimates m/s and n/s, then we have three densities to estimate: f twice and g once. The standard kernel density estimate of f based on the sample $W_1, \ldots, W_m$ is

$$\hat{f}_a(x) = \frac{1}{ma} \sum_{i=1}^{m} K\left(\frac{W_i - x}{a}\right). \tag{1.2}$$

where a is called the bandwidth and determines how much smoothing takes place, and $K(u)$ is a symmetric probability density function, quite often taken to be a polynomial in u with domain $[-1, 1]$. Intuitively, a kernel function is centred at each data point W_i and the density estimate is taken to be the weighted sum of these functions.

Hence replacing each term in equation (1.2) with a suitable estimate we get

$$\hat{\lambda}_{a,b,c}(x) = \frac{m\hat{f}_a(x)}{m\hat{f}_b(x) + n\hat{g}_c(x)}. \tag{1.3}$$

Notice that we allow the estimates of f in the numerator and the denominator to have different bandwidths.

In practice, the use of three independent bandwidths is unnecessary and we concentrate rather on special cases of the above estimators. Two sets of constraints on a, b and c give 'sensible' estimators.

The first and simplest case is to set $a = b = c$, in which case equation

(1.3) can be written as

$$\hat{\lambda}_a(x) = \frac{\sum_{i=1}^{s} Y_i K\{a^{-1}(x - X_i)\}}{\sum_{i=1}^{s} K\{a^{-1}(x - X_i)\}} = \frac{m\hat{f}_a(x)}{s\hat{h}_a(x)}. \tag{1.4}$$

This form of the estimator, which is in fact the standard Nadaraya-Watson kernel regression estimate [10, 11], was first suggested by Copas [5].

It is well known that when estimating a density f, the bandwidth a should decrease as either the sample size s and some measure of 'roughness', such as $R(f'') = \int f''(x)^2 dx$, increase. Thus, we would expect Copas' estimator to have somewhat sub-optimal performance, especially for situations in which f and g (or m and n) differ, depending as it does on the same bandwidth for estimating both f and g.

This line of reasoning leads naturally to the second form of the estimator, with $a = b$, and

$$\hat{\lambda}_{a,c}(x) = \frac{\hat{f}_a(x)}{\hat{f}_a(x) + (n/m)\hat{g}_c(x)}. \tag{1.5}$$

As the following work will show, this estimator appears to be the most appropriate of the nonparametric models.

Finally, setting $b = c$ also gives two bandwidths, one for estimating f and one for estimating h. Unfortunately, using a different bandwidth for the estimation of f in the numerator and the denominator leads to ill-defined estimates of probability, as we now show.

The estimators defined by (1.4) and (1.5) are both constrained to lie in the interval [0,1] by the fact that $\hat{f}_a(x)$ appears in both the numerator and the denominator. This does not apply to the final case, with alarming results. Intuitively, if $a \neq b$ then one estimate of f will be smoother than the other, resulting in regions where $\hat{f}_a/\hat{f}_b$ is greater than 1.

Consider a very simple case, using finite domain kernels, bandwidth a to estimate f and bandwidth b to estimate h. Suppose we have a point x_0

and a single data value X_j, such that $x_0 - X_j = d > 0$. This situation is illustrated in Figure 1.2, where two kernels of differing widths are centred at X_j and p_1 and p_2 denote two possible locations for x_0. Firstly let $x_0 = p_1$ and $a > b$, then $a > d > b$. Now the contribution of the point X_j to $\hat{f}_a(x_0)$ will be $(ma)^{-1}K(d/a)$, and the contribution to $\hat{f}_b(x_0)$ will be zero. Thus if X_j is the only data point within a distance b of x_0, then $\hat{g}_b(x_0)$ is also zero and $\hat{\lambda}(x_0)$ will be infinite.

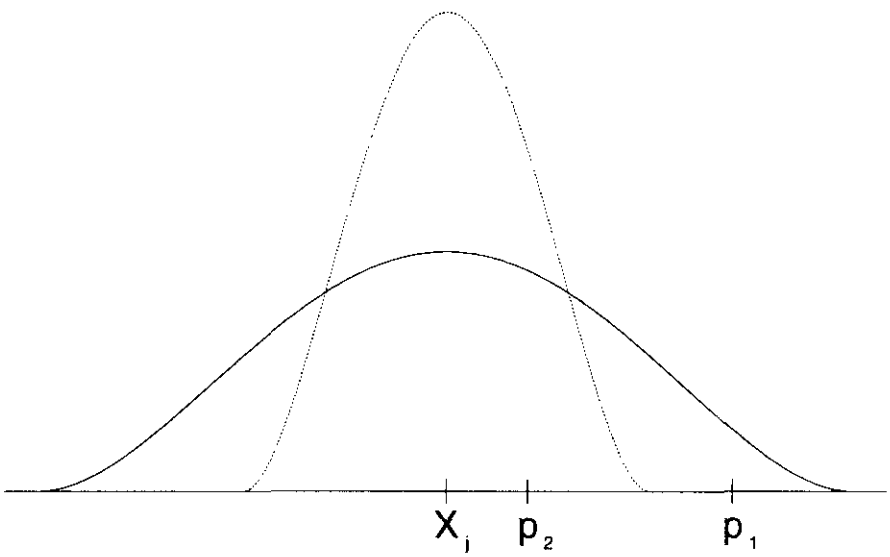

Figure 1.2: The problem of allowing differing bandwidths for the estimation of f in the numerator and denominator

Alternatively, let $x_0 = p_2$ and $b > a$, then $b > a > d$. In this case the contribution of X_j to $\hat{f}_a(x_0)$ will be $(ma)^{-1}K(d/a)$, and the contribution to $\hat{f}_b(x_0)$ will be $(mb)^{-1}K(d/b)$. Again, if $\hat{g}_b(x_0) = 0$, then $\hat{\lambda}(x_0) = \frac{bK(d/a)}{aK(d/b)}$. Thus, in the case where $x_0 = X_j$ and $d = 0$, $\hat{\lambda}(x_0) = b/a > 1$. It is trivial to extend these examples to Gaussian kernels to show that they can also give rise to estimates of λ greater than one.

This phenomenon is illustrated for a trivial data set in Figure 1.3. These

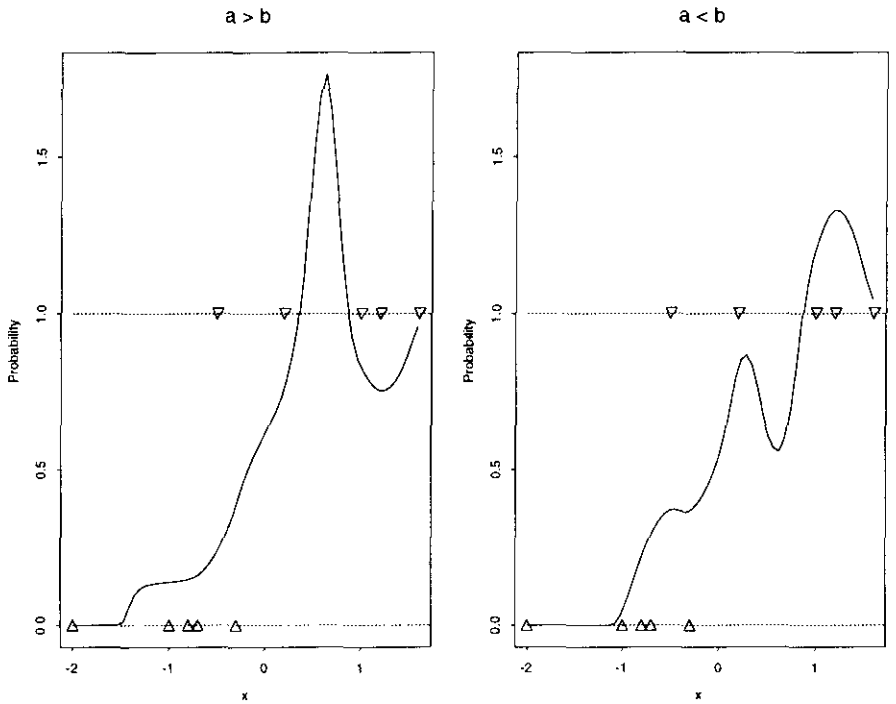

Figure 1.3: Plots of $\hat{\lambda}$ for case $b = c$

problems apply to any situation in which $a \neq b$, and so, for these reasons, we shall consider only the first two estimators in further analysis.

Without wishing to raise the practical issue of selecting the bandwidths at this point, it is worthwhile pointing out, given that we have shown that binary regression can be considered as a problem of estimating the densities from two populations, whether it is sensible to treat them separately. That is, use standard kernel density estimation methods to estimate f and g independently, and then plug these estimates into equation (1.5). This is appealing both from an intuitive standpoint, and from the pragmatic observation that simple kernel density estimation is a well-researched topic, with many sophisticated methods of bandwidth selection which could be used. This estimator is essentially that defined in equation (1.5), but using a bandwidth selection procedure which attempts to optimise the estimation of f and g separately rather than the estimation of λ.

Although the single bandwidth estimator (1.4) was first suggested for use by Copas [5] in 1983, there has been little development of the topic, at least in the fully nonparametric regression sense. Rodriguez-Campos and Cao-Abad [12] construct pointwise bootstrap confidence intervals for the estimator, as well as extending the method to the situation of more than two response categories. Kappenman [13] uses a cross-validated, likelihood-based method to select the bandwidth and also extends the method to the bivariate case by using product kernels in the obvious way. Much more work has been carried out in the area of semiparametric binary regression and this topic is the subject of the next chapter.

1.3 Asymptotic Behaviour

The point was made in the introduction that we are considering binary regression, and that what matters in terms of accuracy, is the ability to estimate $\lambda(x)$ for a large range of x. Thus, we shall measure the accuracy of estimation in terms of mean squared error (MSE) at a point, and mean integrated squared error (MISE) globally. We shall analyse the various estimators from an asymptotic viewpoint, using standard results from the kernel density estimation literature (e.g. Wand & Jones [14]).

Let $m, n \to \infty$ in such a way that $m/n \to \rho$. So, if $m \sim \text{Binomial}(s, \pi_1)$ and $n = s - m$, then $s \to \infty$, $m/s = \pi_1 + O_p(s^{-\frac{1}{2}})$, $n/s = \pi_2 + O_p(s^{-\frac{1}{2}})$ and $m/n = \rho + O_p(s^{-\frac{1}{2}})$. The bandwidths a, b, and c (denoted generically by d) are actually functions of s which tend to zero as $s \to \infty$ but slowly enough that $sd \to \infty$. Colloquially, this is to ensure that although d is decreasing towards zero, the number of data points s is increasing at such a rate that the expected number of data values in the interval $[x - d, x + d]$ tends to infinity. For simplicity, we assume common support for f and g on which

neither is zero and perform all integrations over this support (we take it that we know λ is 0, 1 or undefined but of no interest elsewhere). As usual in this kind of investigation, we assume both f and g have two continuous derivatives and that K is chosen such that quantities involving it exist and are finite.

Noting that the Copas estimator is merely a special case of equation (1.5), we begin by examining the MSE of $\hat{\lambda}_{a,c}(x)$. In the following, for the purpose of clarity, we suppress the argument of the various functions of x i.e. $f(x)$ is denoted simply by f. Also, since the main problem is to estimate the various densities, asymptotically the errors in the estimation of π_1 and π_2 are ignored, as the empirical estimates of these quantities, m/s and n/s have errors which are asymptotically of a higher order, $O(s^{-1/2})$, to those of the density estimates. Thus in all that follows, $\hat{\pi}_1$ and $\hat{\pi}_2$ may be replaced by their true values π_1 and π_2 respectively, and vice versa.

From equation (1.5) we have that

$$\begin{aligned}\hat{\lambda}_{a,c} &= \frac{\pi_1 \hat{f}_a}{\hat{h}_{a,c}} \\ &= \frac{\pi_1 f\left(1+(\hat{f}_a - f)/f\right)}{h\left(1+(\hat{h}_{a,c}-h)/h\right)}.\end{aligned}$$

Using the fact that asymptotically the discrepancies $(\hat{f}_a - f)$ and $(\hat{h}_{a,c} - h)$ are small, we can expand this to

$$\begin{aligned}\hat{\lambda}_{a,c} &\simeq \lambda\left(1+\frac{\hat{f}_a - f}{f}\right)\left(1-\frac{\hat{h}_{a,c}-h}{h}+\left[\frac{\hat{h}_{a,c}-h}{h}\right]^2+\ldots\right) \\ &= \lambda\left(1+\frac{\hat{f}_a - f}{f}-\frac{\hat{h}_{a,c}-h}{h}\right)+\text{terms of smaller order.}\end{aligned}$$

Now, expanding $\hat{h}_{a,c}$ into its components, we can express $\hat{\lambda}$ as

$$\begin{aligned}\hat{\lambda} &\simeq \lambda+\frac{\pi_1(\hat{f}_a - f)}{h}-\frac{\lambda\pi_1(\hat{f}_a - f)}{h}-\frac{\lambda\pi_2(\hat{g}_c - g)}{h} \\ &= \lambda+\frac{1}{h}\left[(1-\lambda)\pi_1(\hat{f}_a - f)-\lambda\pi_2(\hat{g}_c - g)\right] \qquad (1.6)\end{aligned}$$

The standard results about the asymptotic behaviour of density estimates, namely that

$$\mathrm{E}[\hat{f}_h - f] \simeq \frac{h^2\sigma_K^2}{2} f'' + o(h^2),$$

and

$$\mathrm{var}(\hat{f}_h) \simeq \frac{R(K)f}{nh} + o\left((nh)^{-1}\right),$$

(e.g. Wand and Jones, [14], Section 2.5) where $\sigma_K^2 = \int t^2 K(t)dt$ and $R(K) = \int K^2(t)dt$, can now be applied. From equation (1.6) and taking $a \sim c$, we have

$$\begin{aligned} \mathrm{E}\{\hat{\lambda}_{a,c} - \lambda\} &\simeq \frac{1}{h}\left[(1-\lambda)\pi_1 \frac{a^2\sigma_K^2}{2} f'' - \lambda\pi_2 \frac{c^2\sigma_K^2}{2} g''\right] + o(a^2) \\ &= \frac{\pi_1\pi_2\sigma_K^2}{2}\left(\frac{a^2 f''g - c^2 f g''}{h^2}\right) + o(a^2). \end{aligned} \tag{1.7}$$

Furthermore, the asymptotic variance is given by

$$\begin{aligned} \mathrm{var}\{\hat{\lambda}_{a,c}\} &= \frac{(1-\lambda)^2}{h^2}\pi_1^2 \mathrm{var}(\hat{f}_a) + \frac{\lambda^2}{h^2}\pi_2^2 \mathrm{var}(\hat{g}_c) \\ &= \frac{(1-\lambda)^2}{h^2}\pi_1^2 \frac{R(K)}{ma} f + \frac{\lambda^2}{h^2}\pi_2^2 \frac{R(K)}{nc} g + o((sa)^{-1}) \\ &= \frac{\lambda(1-\lambda)}{h} R(K)\left[\frac{\pi_1(1-\lambda)}{ma} + \frac{\pi_2\lambda}{nc}\right] + o((sa)^{-1}). \end{aligned}$$

However, as stated above, we can replace π_1 and π_2 with m/s and n/s respectively to get

$$\mathrm{var}\{\hat{\lambda}_{a,c}\} = \frac{R(K)\lambda(1-\lambda)}{sh}\left(\frac{(1-\lambda)}{a} + \frac{\lambda}{c}\right) + o((sa)^{-1}). \tag{1.8}$$

Note that the covariance term between $\hat{f}_a$ and $\hat{g}_c$ vanishes due to the independence of the samples from the 'successes' and the 'failures'.

We can see immediately that this estimator follows the same pattern as all smoothing estimators, with an asymptotic trade-off between bias, with order $O(a^2)$, and variance, with order $O((sa)^{-1})$. Small bandwidths give

low bias but high variance, and large bandwidths give low variance but large bias.

For Copas' estimator, with $a = c$ the above expressions simplify to

$$\mathrm{E}\{\hat{\lambda}_a\} \simeq \lambda + a^2 \frac{\pi_1 \pi_2 \sigma_K^2}{2} \left(\frac{f''g - fg''}{h^2} \right) + o(a^2) \tag{1.9}$$

and

$$\mathrm{var}\{\hat{\lambda}_a\} = (sa)^{-1} R(K) \frac{\lambda(1-\lambda)}{h} + o((sa)^{-1}). \tag{1.10}$$

Noting that (1.9) may be written as

$$\mathrm{E}\{\hat{\lambda}_a\} \simeq \lambda + a^2 \sigma_K^2 \left(\frac{1}{2}\lambda'' + \frac{\lambda' h'}{h} \right) + o(a^2), \tag{1.11}$$

we can see that these expressions correspond exactly to the well known results about the asymptotic behaviour of the Nadaraya-Watson regression estimator, as given for example, by Fan [15].

Given the trade-off between bias and variance of these estimators, it is reasonable to assess performance by using at a point the mean squared error (MSE), and globally the integrated MSE, since

$$\begin{aligned} \mathrm{MSE}\{\hat{\lambda}_{a,c}(x)\} &= \mathrm{E}\left\{ [\hat{\lambda}_{a,c}(x) - \lambda(x)]^2 \right\} \\ &= \mathrm{bias}^2 \left[\hat{\lambda}_{a,c}(x) \right] + \mathrm{var} \left[\hat{\lambda}_{a,c}(x) \right]. \end{aligned} \tag{1.12}$$

To consider some very simple cases, define

$$\phi(x) = (2\pi)^{-1/2} \exp(-x^2).$$

Let the distribution of the successes follow a standard Gaussian, $X_{Y=1} \sim N(0,1)$, and let $X_{Y=0} \sim N(\mu, 1)$ and $\pi_1 = \pi_2 = 0.5$. This model is linear on the logistic scale, since

$$\mathrm{logit}(\lambda(x)) = (\mu^2/2) - \mu x.$$

Then $f(x) = \phi(x), g(x) = \phi(x-\mu)$ and equation (1.12) becomes

$$\begin{aligned}\mathrm{MSE}\{\hat{\lambda}_{a,c}(x)\} &= \frac{\sigma_K^4}{4}\frac{\phi(x)^2\phi(x-\mu)^2}{[\phi(x)+\phi(x-\mu)]^4}\left(a^2(x^2-1)-c^2(x^2-2x\mu+\mu^2-1)\right)^2 \\ &\quad + \frac{2R(K)}{s}\frac{\phi(x)\phi(x-\mu)}{[\phi(x)+\phi(x-\mu)]^4}\left(\frac{\phi(x-\mu)}{a}+\frac{\phi(x)}{c}\right).\end{aligned}$$

For the case where $\mu = 1, a = c = 1$ and $n = 200$, Figure 1.4 plots the asymptotic bias and variance for this model.

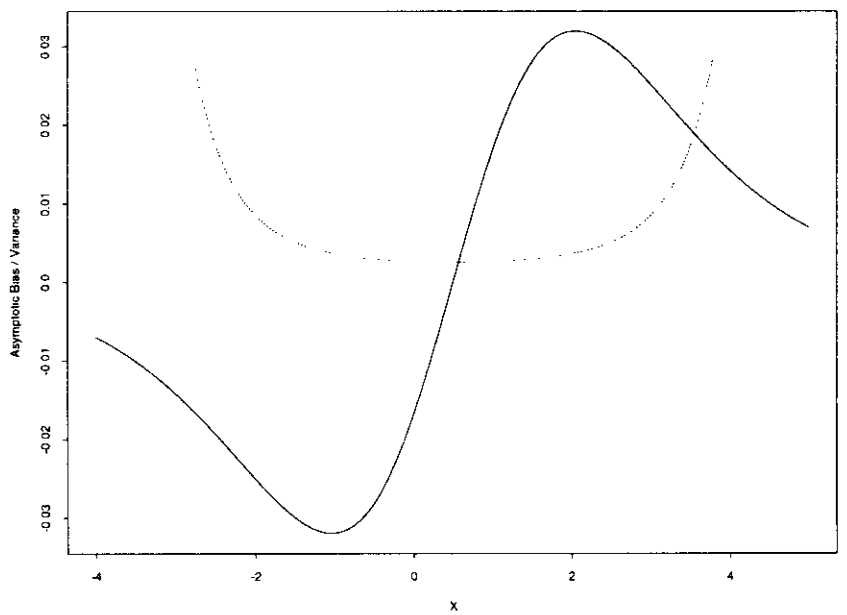

Figure 1.4: Asymptotic bias (solid line) and variance (dotted line) for Gaussian linear shift model

Clearly, for this model although the absolute bias is maximal at around -1 and +2, the variance is increasing exponentially in the tails of the density h. Globally, this means that we immediately run into problems when trying to integrate the MSE given by (1.12). To ensure that the integrals remain finite, we must calculate a weighted MISE, and to do this we use the weight function $h(x)^2$. Thus, we work in terms of

$$\mathrm{WMISE}(\hat{\lambda}) = \int h^2(z)\mathrm{E}\{\hat{\lambda}(z)-\lambda(z)\}^2 dz.$$

This has the appealing property that it weights the discrepancy between the estimate and the true λ according to the overall density of the covariate X. Using (1.7) and (1.8), this gives

$$\begin{aligned} \mathrm{WMISE}(\hat{\lambda}_{a,c}) &= \frac{\sigma_K^4}{4}\mathrm{R}\left(a^2\pi_1(1-\lambda)f'' - c^2\pi_2\lambda g''\right) \\ &+ \frac{\mathrm{R}(K)}{s}\left[a^{-1}\int \lambda(1-\lambda)^2 h + c^{-1}\int \lambda^2(1-\lambda)h\right] \end{aligned} \quad (1.13)$$

which reduces in the $a = c$ case to

$$\mathrm{WMISE}(\hat{\lambda}_a) = \frac{\sigma_K^4}{4}a^4\pi_1^2\pi_2^2\mathrm{R}\left(\frac{f''g - fg''}{h}\right) + \frac{\mathrm{R}(K)}{sa}\int \lambda(1-\lambda)h \quad (1.14)$$

Clearly, since equation (1.14) is merely a constrained version of equation (1.13), we have the inequality

$$\inf_{a,c}\mathrm{WMISE}(\hat{\lambda}_{a,c}) \leq \inf_{a}\mathrm{WMISE}(\hat{\lambda}_a)$$

with equality if the unconstrained minimum lies on the line $a = c$. Intuitively, this would imply that f and g must be quite similar, and this is demonstrated in the simulation experiment to follow. This also recalls the remarks made earlier about treating the problem as one of two independent density estimation cases. It is clear now that this is really just a bandwidth selection issue for the estimator $\hat{\lambda}_{a,c}$; this *must* give a smaller WMISE than the others, so the real question of interest is *by how much ?*

The complex nature of the expressions derived for the WMISE makes theoretical comparison of the estimators tricky, and so we rely mainly upon simulation for our conclusions. For the example used above to demonstrate the bias-variance trade off, we have that

$$\begin{aligned} \mathrm{WMISE}\{\hat{\lambda}_{a,c}(x)\} &= \frac{\sigma_K^4}{16}\int \frac{\phi(x)^2\phi(x-\mu)^2}{[\phi(x)+\phi(x-\mu)]^2}\left(a^2(x^2-1) - c^2(x^2-2x\mu+\mu^2-1)\right)^2 dx \\ &+ \frac{R(K)}{2s}\int \frac{\phi(x)\phi(x-\mu)}{[\phi(x)+\phi(x-\mu)]^2}\left(\frac{\phi(x-\mu)}{a} + \frac{\phi(x)}{c}\right)dx \end{aligned} \quad (1.15)$$

These integrals have no closed form solution and must be calculated numerically. For this particular example, however, we can simplify matters by noting that a and c are exchangeable parameters, and so the global minimum must have $a = c$. This reduces equation (1.15) to

$$\text{WMISE}\{\hat{\lambda}_{a,c}(x)\} = a^4 \frac{\sigma_K^4}{16} I_1 + a^{-1} \frac{R(K)}{2s} I_2,$$

where

$$\begin{aligned} I_1 &= \int \frac{\phi(x)^2 \phi(x-\mu)^2}{[\phi(x) + \phi(x-\mu)]^2} \left(2x\mu - \mu^2\right)^2 dx \quad \text{and} \\ I_2 &= \int \frac{\phi(x)\phi(x-\mu)}{\phi(x) + \phi(x-\mu)} dx. \end{aligned}$$

Simple calculus can be used to show that this implies that the optimal bandwidth is thus of the order $s^{-1/5}$, namely

$$a_{\text{opt}} = s^{-1/5} \left(\frac{2R(K)}{\sigma_K^4} \quad \frac{I_2}{I_1} \right)^{1/5},$$

a result which again parallels the asymptotic behaviour of the component density estimates. Note that I_1 and I_2 are essentially functions of the difference in means, μ. If we take $s = 200$ and calculate the asymptotic WMISE-optimal bandwidth for various values of μ using numerical integration to evaluate I_1 and I_2, we get the non-monotic relationship between a_{opt} and μ shown in Figure 1.5.

When there is very little separation between the two densities, a large bandwidth is optimal. This declines as the mean difference increases, then increases again. A plausible intuitive explanation for this may be that as the mean separation increases, the overlap between the distributions decreases, and estimation of the densities in the tails becomes more important, requiring a smaller bandwidth with consequently lower bias. As the distributions move further apart, however, and there is almost no overlap, a larger and

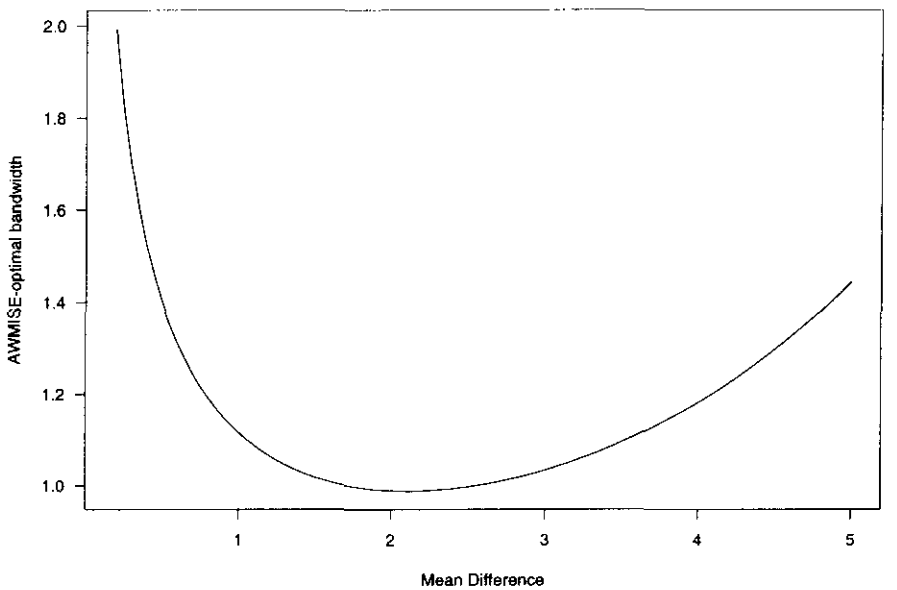

Figure 1.5: AWMISE-Optimal bandwidth as a function of mean separation for the Gaussian linear shift model

larger optimal bandwidth can be used to reduce the variance without increasing the bias.

To consider a case for which the WMISE-optimal bandwidths are not equal, let the distribution of the successes follow a standard Gaussian, as before, $X_{Y=1} \sim N(0,1)$, and $\pi_1 = \pi_2 = 0.5$, but now let $X_{Y=0} \sim N(0,\sigma^2)$. Define ϕ_σ as the density of a Gaussian random variable with mean 0 and variance σ^2. Thus $\phi_\sigma(x) = (1/\sigma)\ \phi(x/\sigma)$, and the model is now quadratic on the logistic scale, with

$$\text{logit}(\lambda(x)) = \frac{x^2}{2}\left(\frac{1}{\sigma^2} - 1\right) + \log\sigma.$$

Tedious calculation shows that the WMISE in this case is

$$\begin{aligned}\text{WMISE}\{\hat{\lambda}_{a,c}(x)\} &= \frac{\sigma_K^4}{16}\int \frac{\phi(x)^2\phi_\sigma(x)^2}{[\phi(x)+\phi_\sigma(x)]^2}\left(a^2(x^2-1) - \frac{c^2}{\sigma^2}(x^2/\sigma^2-1)\right)^2 dx \\ &\quad + \frac{R(K)}{2s}\int \frac{\phi(x)\phi_\sigma(x)}{[\phi(x)+\phi_\sigma(x)]^2}\left(\frac{\phi_\sigma(x)}{a} + \frac{\phi(x)}{c}\right) dx. \qquad (1.16)\end{aligned}$$

In this situation the bandwidths are not exchangeable and the WMISE-optimal values will not satisfy $a = c$, although they will still have the same asymptotic order, namely $O(s^{-1/5})$. Figure 1.6 shows a contour plot of the WMISE for various values of a and c for the case where $\sigma^2 = 0.25$ and $s = 200$. In this case the asymptotic WMISE-optimal bandwidths are approximately $a = 1.25$ and $c = 0.6$, confirming our intuition that a smaller bandwidth is required to estimate the density with the smaller variance.

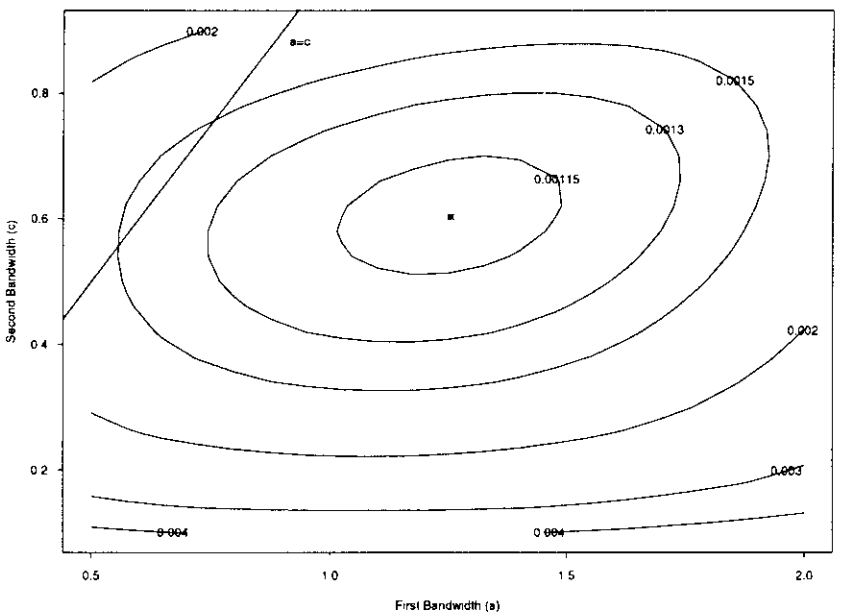

Figure 1.6: Asymptotic WMISE as a function of two bandwidths for the Gaussian variance change model

1.4 Practical Performance

The intractability of the expressions for the asymptotic WMISE of the estimators does not allow easy comparison. This is further complicated by the knowledge that the three estimators ($\lambda_a, \lambda_{a,c}$ and λ_{sep} below) are identical

but for the constraints upon the bandwidth. Thus, we can be certain that the two bandwidth solution will minimise the WMISE, but from a practical point of view we would wish to know whether the gain over a single bandwidth or separable bandwidths problem is worth the added difficulty of estimating two bandwidths.

1.4.1 Methods

To pursue this problem, a simulation experiment was performed to assess the small sample performance of the three estimators in practice. For a variety of differing models, the relative performances of the estimators discussed in the previous section were evaluated and compared. The three nonparametric estimators considered were therefore :

- $\lambda_a(x)$ - only one bandwidth.
- $\lambda_{a,c}(x)$ - two bandwidths.
- Estimating f and g independently as two separate densities - λ_{sep}.

Twenty four models were simulated; in each case f was taken to be a standard Gaussian density, mean zero, variance one, and G, the random variable with density g, and π_1 were as shown in Table 1.1, where MW(k) refer to the Gaussian mixture distributions used by Marron and Wand [16], which provide a wide range of non-symmetric and multimodal distributions. These densities are studied more extensively in Chapter 6.

These distributions give rise to a wide variety of probability functions, as shown in Figure 1.7. Note that the first nine models are linear in x on the logistic scale, whilst the next three are quadratic.

Model		G	π_1
Linear Shift	1	$N(0.5, 1)$	0.5
	2	$N(0.75, 1)$	0.5
	3	$N(1, 1)$	0.5
	4	$N(1.25, 1)$	0.5
	5	$N(1.5, 1)$	0.5
Different Proportions	6	$N(1, 1)$	0.2
	7	$N(1, 1)$	0.4
	8	$N(1, 1)$	0.6
	9	$N(1, 1)$	0.8
Different Variance	10	$N(0.5, (0.2)^2)$	0.5
	11	$N(0.5, (0.5)^2)$	0.5
	12	$N(0.5, (0.8)^2)$	0.5
Cauchy	13	Cauchy(0.6)	0.5
	14	Cauchy(0.8)	0.5
	15	Cauchy(1.0)	0.5
	16	Cauchy(1.2)	0.5
Marron-Wand	17	MW(2)	0.5
	18	MW(3)	0.5
	19	MW(4)	0.5
	20	MW(5)	0.5
	21	MW(6)	0.5
	22	MW(7)	0.5
	23	MW(8)	0.5
	24	MW(9)	0.5

Table 1.1: Distributions used for g in simulation experiment

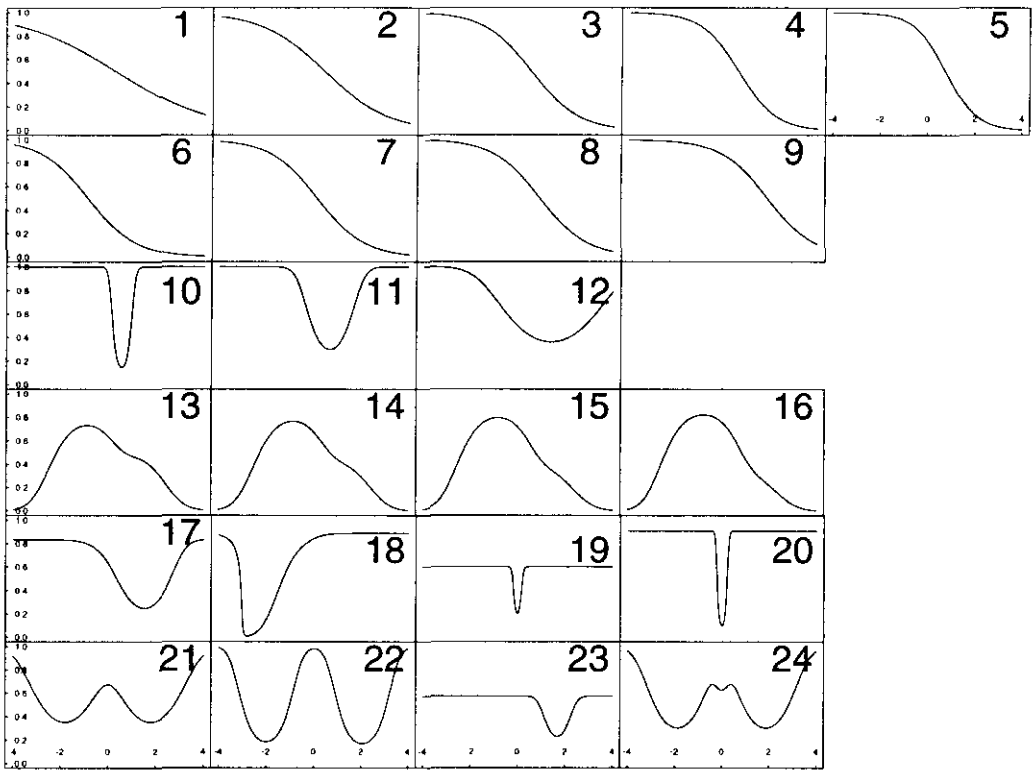

Figure 1.7: Plots of λ for the simulated models

For each model, samples of size 200 were drawn from the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$. As the true $\lambda(x)$ and $h(x)$ were known in each case for any estimator method and choice of bandwidths, the weighted ISE (WISE) could be calculated. This was approximated by a weighted sum of squared errors over a grid of 401 points on the range [-4,4]:

$$\text{WISE} \approx \sum_{j=1}^{401} h^2(x_j) \left[\hat{\lambda}(x_j) - \lambda(x_j)\right]^2. \tag{1.17}$$

To separate the question of method comparison from that of bandwidth selection, equation (1.17) was used to find the WISE-optimal bandwidth(s) for each of the first two estimators, by using a grid search procedure. This ensured a 'best-case' scenario where each estimator was allowed to produce its minimum WISE for comparison. Practically, bandwidth selection procedures are unlikely to produce such an optimal bandwidth, but this design

provides an objective assessment of the methods themselves. In a similar fashion, for the third model, involving two independent density estimation problems, the true $f(x)$ and $g(x)$ were used to determine ISE-optimal bandwidths which were then used to calculate $\hat{\lambda}$. Each model was used to produce 100 data sets and a quartic kernel $\left(\frac{15}{16}(1-x^2)^2 I(|x|<1)\right)$ was used throughout.

At the extremes of the chosen interval, near -4 and 4, and also for small bandwidths, it was sometimes the case that the estimates would be undefined. That is, because of the use of a finite domain kernel, the estimate of h is zero in some region. For the purposes of the simulation, the estimate of λ in these regions was set to the true value and so they made no contribution to the WISE.

1.4.2 Results

For each of 100 simulated datasets from each of 10 densities, the WISE was calculated for the three different estimators. What can be done to summarise this data? As discussed above, the two bandwidth version of the estimator, namely $\lambda_{a,c}$, will always achieve the minimum WISE and it is the increase in WISE for the other two estimators in which we are interested. Moreover, the actual values of WISE($\hat{\lambda}_{a,c}$) were skewed to the right, with some models for some datasets proving very difficult to estimate accurately, resulting in large errors. Thus, the *relative* increase in WISE, as a percentage of WISE($\hat{\lambda}_{a,c}$) was used. This measure was found in all cases to be skewed to the right and so the median was chosen rather than the mean as a fair overall summary.

Tables 1.2 and 1.3 show the median increase in WISE, as a percentage of the two-bandwidth fully optimal value, caused by using either λ_a or λ_{sep}.

Model	Percentage Increase		Wilcoxon Test	
	λ_a	λ_{sep}	W-Statistic	p-value
Linear Shift				
1 : $\mu = 0.5$	38.36	75.32	-5.72	0.00000
2 : $\mu = 0.75$	14.23	19.57	-3.58	0.00035
3 : $\mu = 1$	8.09	13.75	-3.20	0.00139
4 : $\mu = 1.25$	2.66	16.95	-5.76	0.00000
5 : $\mu = 1.5$	2.52	10.51	-6.58	0.00000
Different Proportions				
6 : $\pi_1 = 0.2$	19.13	20.40	-1.24	0.21388
7 : $\pi_1 = 0.4$	8.20	13.11	-3.51	0.00044
8 : $\pi_1 = 0.6$	5.10	11.48	-2.69	0.00713
9 : $\pi_1 = 0.8$	9.03	14.42	-0.57	0.57166
Different Variance				
10 : $\sigma = 0.2$	140.65	24.44	7.69	0.00000
11 : $\sigma = 0.5$	41.81	21.11	5.48	0.00000
12 : $\sigma = 0.8$	34.79	61.11	-1.87	0.06166

Table 1.2: Median percentage increase in optimal WISE for single and separate bandwidth methods over two bandwidth method, and Wilcoxon signed rank test of single versus separate methods for Models 1 to 12

Model	Percentage Increase		Wilcoxon Test	
	λ_a	λ_{sep}	W-Statistic	p-value
Cauchy				
13 : $\mu = 0.6$	12.17	21.07	-2.42	0.01572
14 : $\mu = 0.8$	13.47	28.08	-3.22	0.00127
15 : $\mu = 1$	7.73	14.42	-2.40	0.01647
16 : $\mu = 1.2$	3.54	15.84	-4.20	0.00003
Marron-Wand				
17 : MW(2)	10.90	13.93	-0.51	0.61205
18 : MW(3)	14.37	445.14	-8.68	0.00000
19 : MW(4)	41.37	17.06	5.75	0.00000
20 : MW(5)	83.06	9.05	8.52	0.00000
21 : MW(6)	5.53	8.29	-1.76	0.07921
22 : MW(7)	13.03	6.38	3.56	0.00038
23 : MW(8)	12.29	19.81	-4.28	0.00002
24 : MW(9)	5.20	7.32	-1.40	0.16118

Table 1.3: Median percentage increase in optimal WISE for single and separate bandwidth methods over two bandwidth method, and Wilcoxon signed rank test of single versus separate methods for Models 13 to 24

1.4.3 Discussion

The most obvious and important conclusion to be drawn from these results is that, with the exception of a few models, the single bandwidth estimator λ_a out-performs the naive approach of treating the problem as two separate density estimations. More rigorously, a Wilcoxon Rank Sum test was performed to compare the WISE values for λ_a and λ_{sep}. Note that it is unnecessary to test the optimal WISE values from $\lambda_{a,c}$ against the other two estimators, as we can be certain that they are smaller; what is important is whether they are so reduced that we would consider that the added complication of two bandwidths to estimate makes their use worthwhile in a practical sense.

The cases in which separate estimation is beneficial are when the distribution of g is very different from that of f, namely when g is Gaussian with small variance, or has a high kurtosis, as in densities 4 and 5 of the Marron-Wand models. This implies that they require substantially different bandwidths and so the single bandwidth method will fail. As intuitively expected, λ_a is best when the two densities are similar. For the models with differing variances, very large increases in WISE over that for $\lambda_{a,c}$ were observed.

The size of the differences between the optimal WISEs in each case however, suggest that, provided a suitable method can be found for automatically selecting two bandwidths simultaneously, neither of the two other alternatives should be applied.

To explore this further, consider Model 2. In this case λ is linear on a logistic scale, and both f and g are Gaussian with variance 1, yet the median increase in WISE of the single bandwidth method over the two bandwidth method is approximately 14% for λ_a and 20% for λ_{sep}. To show

how the bandwidths vary between the single and the double bandwidth methods, we order the simulation results by their relative increase in WISE and select the centre portion, from the lower quartile to the upper quartile. For these 50 data sets, Figure 1.8 plots the WISE-optimal bandwidths for the two bandwidth method and connects them to the corresponding values for the single bandwidth method, which obviously all lie along the line $a = c$. The increases in WISE caused by using λ_a in place of $\lambda_{a,c}$ ranged in these datasets from 3.4% to 45%. Clearly, even in this very easy to estimate case, the minima of the WISE-surface in the two dimensional bandwidth space do not always lie in the region of the line $a = c$.

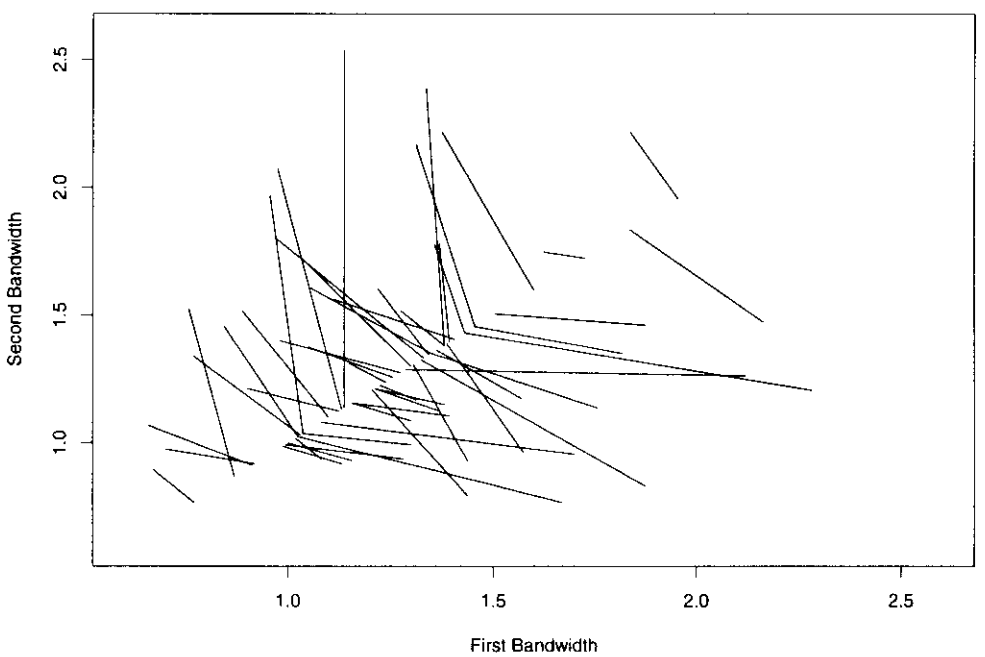

Figure 1.8: Relationship between the WISE-optimal bandwidths of the $\lambda_{a,c}$ and λ_a for datasets from Model 1 which show a relative increase in WISE of between 3.4 and 45%

This phenomenon of large differences between the bandwidths is replicated when the ISE-optimal bandwidths used by λ_{sep} are compared to those used by $\lambda_{a,c}$. In this case, the separate estimation seems to produce bandwidths which are often too small when compared to the WISE-optimal ones. For the example above, the WISE-optimal bandwidth for estimating f is on average 0.2 larger that the ISE-optimal value (range of differences -0.6 to 2.5), and for estimating g the WISE-optimal value is on average 0.35 larger (range of differences -0.7 to 3.4.)

1.5 Conclusions

This chapter has described the general nonparametric binary kernel regression estimator for a single covariate, extending the standard Nadaraya-Watson estimate to the two bandwidth case. Both asymptotic and simulation results have shown that in cases where the density of the failures and the density of the successes differ substantially in terms of variability, then the use of two bandwidths is essential. Moreover, even when there are less obvious differences, there are substantial gains to be made in the WISE by using the two bandwidth estimator, providing that a reasonable bandwidth selection procedure can be devised for this scenario. This topic will be pursued in a later chapter.

Another interesting result is the fact that treating the problem as two separate density estimations is not useful in terms of optimising the WISE of the conditional expectation of the response given the covariate. This can be explained intuitively by noting that the estimates which minimise the ISEs for f and g independently may not optimise λ. Intuitively we can argue that as $f/(f+g)$ is the focus of interest, we have a situation where biases in the estimation of f can cancel out biases in the estimation of g, allowing us to

use estimators with lower variance and higher bias, which of course implies larger bandwidths.

This is shown in Figure 1.9 which shows the data relating burn area to survival from section 1.1. The solid line was calculated using $\hat{\lambda}_{sep}$ where the bandwidths were selected by using the Sheather-Jones plug-in bandwidth selection procedure [17] on survivors and non-survivors independently. This resulted in values of $a = 0.417$ and $b = 0.309$, showing that the non-survivors have a slightly smaller variability. However, the estimate of λ obtained shows significant under-smoothing. The dotted line is the result of using $\hat{\lambda}_{a,c}$, with a and c taken to be double the values calculated above. This estimate is much smoother and more biologically plausible.

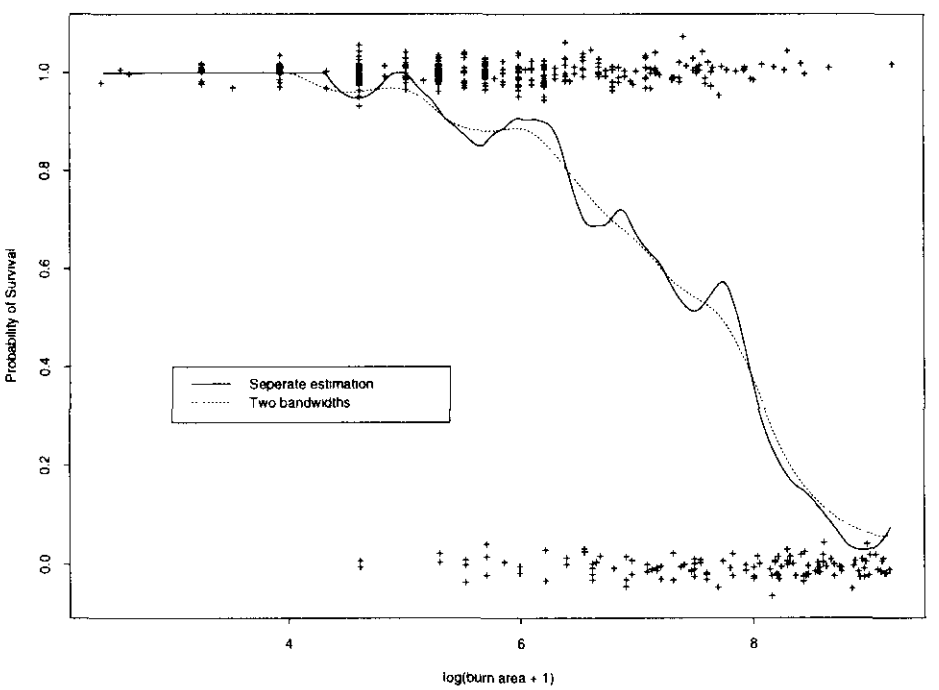

Figure 1.9: Fitted probability functions for the burn data showing under-smoothing of separate bandwidth estimation method (solid line) versus full two bandwidth method (dotted line).

One interesting point to note about the $\hat{\lambda}_{a,c}$ estimate in this case, however, is that although the curve is relatively smooth, there are significant fluctuations for values of log(Burn area + 1) between 4 and 6, where there is significant clustering of the covariate values. This is probably a result of the $\lambda' h'/h$ term in the asymptotic bias (1.11), as when the design density h shows this clustering, then this term will be large. This relationship between the design density and the asymptotic bias provides one of the motivations for the methods explored in the next chapter.

Chapter 2

Semiparametric Binary Regression

2.1 Introduction

In the previous chapter we were concerned only with nonparametric modelling of the binary regression function, motivated explicitly by situations where the standard logistic regression model, with its assumption of linearity, is not appropriate. Fully nonparametric methods allow the data to dictate the shape of the resulting estimator, independent of any assumptions about the link between the probability function and the covariate.

Consider the simple case of linear regression with a single covariate. In the presence of non-linearity, we could attempt to model by fitting polynomial terms in the covariate X. As the regression function $Y = m(x)$ became less linear, however, we would require polynomial terms of high degree to ensure an adequate fit.

An alternative would be to *locally* fit a low-degree polynomial. To calculate this estimate at a point x_0, we weight the points in the neighbourhood

of x_0 using a suitable kernel function, determine the weighted least squares fit using a low-degree polynomial, and use the fitted value of the local polynomial at the point x_0 as the fitted value $\hat{m}(x_0)$. Formally, the estimate of the regression function $\hat{m}(x_0)$ is the value of β_0 in the solution of the order p local polynomial fit given by the minimisation of

$$\sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right]^2 K_h(X_i - x_0),$$

where Y_i and X_i, $i = 1, \ldots, n$, are the regression data, h is the bandwidth and p is the order of the local polynomial.

This method is explored by Fan [15] where a kernel weighted local linear ($p = 1$) regression is used, and expanded upon at length by Fan and Gijbels [18]. It can be shown that the method may be expressed explicitly as a weighted average smooth of the data, and various asymptotic properties of the estimate can be derived.

This idea of locally linear smoothing can also be extended to generalised linear models. Fan, Heckman and Wand [4] show that, by considering a kernel weighted quasi-likelihood (which for the Gaussian case with identity link is equivalent to the least squares formulation given above), local polynomials can be incorporated into the model specification. In a similar vein to the definition above, the inverse link function is expressed as a low-degree polynomial in X, substituted into the quasi-likelihood function, weighted by the kernel function and maximised to give the point estimate.

To apply this to binary data, let $\mu(x)$ be the true probability function and take $V(\mu) = \mu(1 - \mu)$ and link $\eta(x) = \text{logit}[\, \mu(x) \,]$, giving the quasi-likelihood

$$Q = \sum_{i=1}^{n} \left(Y_i \ln \left[\mu(X_i) \right] + (1 - Y_i) \ln[1 - \mu(X_i)] \right). \tag{2.1}$$

To calculate the probability of success at x we must maximise with respect to $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ the weighted quasi-likelihood

$$\begin{aligned} Q(X, \beta, h) &= \sum_{i=1}^{n} \Big\{ K_h(X_i - x) \left[Y_i \left(\beta_0 + \beta_1 (X_i - x) + \ldots + \beta_p (X_i - x)^p \right) \right] \\ &\quad - K_h(X_i - x) \left[\ln \left(1 + e^{\beta_0 + \beta_1 (X_i - x) + \ldots + \beta_p (X_i - x)^p} \right) \right] \Big\}. \quad (2.2) \end{aligned}$$

Unfortunately there is no explicit closed-form solution to these equations, except when $p = 0$, and so a numerical optimisation procedure must be applied at each value of x for which an estimate of λ is required. This implies that the computational burden of this method is considerably greater than the nonparametric methods, requiring as it does an iterative solution at each point. An efficient algorithm for this procedure is developed below in Section 2.3.

The model defined by equation (2.2) has two appealing qualities. Firstly, if we let $h \to \infty$, then the kernel function gives equal weight to all data points, independent of x, and the problem reduces to a standard logistic regression model with a polynomial of order p in the covariate X.

Secondly, if we set $p = 0$, equivalent to fitting a locally constant model, then equation (2.2) becomes

$$Q(X, \beta_0, h) = \sum_{i=1}^{n} K_h(X_i - x) \left[Y_i \beta_0 - \ln \left(1 + e^{\beta_0} \right) \right],$$

but this is maximised when

$$\sum_{i=1}^{n} K_h(X_i - x) \left[Y_i - \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right] = 0.$$

Thus, the fitted value in the case $p = 0$ is

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

which is exactly the nonparametric estimate of Copas from the previous chapter. The locally linear logistic model which we concentrate on in this

chapter, thus represents a 'half-way house' between fully parametric logistic regression and unconstrained fully nonparametric kernel smoothing.

The problem of coping with non-linearities in the relationship between a covariate and the conditional expectation of the response has been considered before, and many of the suggested solutions are reviewed by Schimek [19]. The literature deals almost exclusively with the general multivariate case and attempts are made to unify the models through a generalised linear model approach. The intention of the current work is to be much less ambitious, and to concentrate on the univariate case in an attempt to gain insights into the problems of the semiparametric approach, such as bandwidth selection, which are often glossed over in the more comprehensive treatments.

Bonneu et al. [20] compare a form of semiparametric estimator which they call the pseudo-maximum likelihood approach (PMLE) whereby the regression function $\hat{m}(x) = \mathrm{E}(Y|x)$ is given (for multivariate x) by an unknown function r on a linear combination of the x variables,

$$m(x) = \mathrm{E}(Y|x) = r_\alpha(\theta^T x).$$

They estimate r by a Nadaraya-Watson smooth with bandwidth α of the Y_i on the 'single index' $\theta^T x$ (from which these models take their name) with a simple plug-in estimate for α. The parameter vector θ is then estimated by maximising a pseudo-likelihood function. For the binary case this reduces to

$$Q_n(Y_i, x_i, \theta) = \sum_{i=1}^{n} \left[Y_i \mathrm{logit}\left(\hat{r}(\theta^T x)\right) + \log\left(1 - \hat{r}(\theta^T x)\right)\right]$$

However, we can see that any constant factor multiplying the vector θ can be absorbed into the bandwidth without changing the estimate. Therefore not all components of θ are identifiable and in practice the contraints $|\theta| = 1$ or

$\theta_1 = 1$ are used. Unfortunately, this implies that for the case of univariate x, θ and hence Q_n are both fixed, and the estimate again becomes the simple Nadaraya-Watson estimate using an unjustified choice of bandwidth.

Klein and Spady [21] develop an almost exactly similar estimator, but choose to complicate matters further by involving both higher-order kernel smoothing and consequently a trimming function to ensure asymptotic correctness. As we saw in the previous chapter when examining the nonparametric solution to this problem which estimates each density independently, it is by no means certain that the approach of using improved methods for part of the problem necessarily improve the whole.

Other work defines 'semiparametric' as meaning that the model is of the form

$$m(x) = \mathrm{E}(Y|x) = f\left(\beta^T x + g(t)\right),$$

where f is the link function, x is a vector of covariates entering the model in a linear fashion, and t a vector of non-linear covariates entering the model through the nonparametric smoother g. Estimation of g tends to proceed though a penalised version of the (quasi-)likelihood, in an analogue of classical spline smoothing. Typically, the problem is separated into two parts; estimating the smoothing function g and then maximising the (quasi-)likelihood function for fixed g. These methods have been explored by Green [22], Chen [23] and Severini & Staniswallis [24] among others, each with slightly different methods of estimating the smoothing function g, but all with both a parametric and a nonparametric component, which reduces to a fully nonparametric solution when considering the univariate case.

2.2 Asymptotic Behaviour

We have already explored the asymptotic properties of the case $p = 0$ as this corresponds exactly to the estimator λ_a from the previous chapter. The asymptotic bias and variance for this estimator are given in equations (1.10) and (1.11), and will not be repeated here.

Fan, Heckman and Wand [4] extend the calculations to the local linear case ($p = 1$) and beyond. They use the error in the estimation of η rather than λ, where

$$\eta = \text{logit}(\lambda) = \log\left[\lambda/(1-\lambda)\right],$$

but also explain how the bias and variance in terms of λ can be derived. They give different expressions for even and odd values of p, but as we will soon see, the gains made by considering locally quadratic ($p = 2$) or cubic ($p = 3$) fitting are small.

Denoting the local polynomial logistic estimator by $\lambda_{LP,p}$ for polynomials of order p, the bandwidth by a, and the total number of observations by s, for $p = 1$ we have

$$\mathrm{E}\{\hat{\lambda}_{LP,1}\} = \lambda + \frac{a^2\sigma_K^2}{2}\eta''\lambda(1-\lambda) + o(a^2), \tag{2.3}$$

and

$$\mathrm{var}\{\hat{\lambda}_{LP,1}\} = (sa)^{-1}R(K)\frac{\lambda(1-\lambda)}{h} + o((sa)^{-1}), \tag{2.4}$$

where $h(x)$ is the density of the covariate x. Thus the asymptotic variance of the estimator is the same as for $p = 0$. However, the bias expression is simpler than equation (1.11), missing out the term involving λ' and h'. The second derivative of η can be expanded into terms involving only terms in λ and its derivatives, to give

$$\mathrm{E}\{\hat{\lambda}_{LP,1}\} = \lambda + \frac{a^2\sigma_K^2}{2}\left(\lambda'' + \frac{\lambda'^2(2\lambda-1)}{\lambda(1-\lambda)}\right) + o(a^2). \tag{2.5}$$

Similarly, for $p = 2$, the relevant expressions are

$$\mathrm{E}\{\hat{\lambda}_{LP,2}\} = \lambda + a^4\sigma_L^4 \left[\frac{\eta^{(IV)}}{24} + \frac{\eta'''[\lambda(1-\lambda)h]'}{6\lambda(1-\lambda)h} \right] \lambda(1-\lambda) + o(a^4), \quad (2.6)$$

and

$$\mathrm{var}\{\hat{\lambda}_{LP,2}\} = (sa)^{-1} R(L) \frac{\lambda(1-\lambda)}{h} + o((sa)^{-1}), \quad (2.7)$$

where L is the fourth-order kernel derived from K according to the formula

$$L(x) = \frac{\mu_4 - \mu_2 x^2}{\mu_4 - \mu_2^2} K(x),$$

where $\mu_k = \int v^k K(v) dv$, the kth moment of K, so that $\mu_2 = \sigma_K^2$. It is simple to check that L is indeed a fourth-order kernel by showing that $\int u^2 L(u) du = 0$; this and other similar kernel functions are discussed in detail in Chapter 5.

For general p, similar expressions can be derived. Odd values lead to bias in terms of $\eta^{(p+2)}$ alone, whereas even values of p have an additional term involving

$$\frac{\eta^{(p+1)}[\lambda(1-\lambda)h]'}{h}.$$

As noted in the previous chapter, this term can influence the estimate of λ when there is non-uniformity of the design density h, since then $h' \neq 0$. For this reason, and due to consideration of boundary effects, only odd values of p are used in practice by many people, typically $p = 1$.

Boundary effects exist for all kernel smoothers when the support of the x variable is finite or semi-finite. In this case there will be at least one boundary point beyond which there will be no possible covariate values. The simplest case is when x is constrained to be non-negative. Then for any point within a distance h (the bandwidth) of 0, some of the kernel function centred at this point will be in the region of $x < 0$. This portion of the kernel will not contribute to the final estimate, and the closer the

point is to the boundary, the more of the kernel that is 'lost' in this fashion. This phenomenon has both theoretical and practical implications: most smoothers have poorer convergence rates in the boundary region than in the interior, and this can be readily seen in practice, such as when trying to estimate a highly right-skewed density bounded below by zero.

Local linear smoothers, however, and their extension to GLM's have the appealing property that if p is odd, then the asymptotic rate of convergence of the estimator in the boundary region is unchanged. The constants σ_K^2 and $R(K)$ are replaced by definite integrals bounded by the support of x, but otherwise equations (2.3) and (2.4) apply. This is not the case when p is even; then the asymptotic rate of convergence is slower in the boundary region than the interior, which implies that Copas' estimator, corresponding to $p = 0$ is, at least asymptotically, inferior for problems where the range of x is bounded.

Returning to the examples from Section 1.3, the simple Gaussian linear shift model will have asymptotic bias of $o(a^2)$, as in this case $\eta''(x) = 0$. For the model where the density of failures is Gaussian but with variance $\sigma^2 \neq 1$, the model is quadratic on the logistic scale, and

$$\eta''(x) = \left(\frac{1}{\sigma^2} - 1\right).$$

It is then easy to demonstrate that in this case

$$\mathrm{E}\{\hat{\lambda}_{LP,1}\} = \frac{a^2\sigma_K^2}{2} \frac{\phi(x)\phi_\sigma(x)}{[\phi(x) + \phi_\sigma(x)]^2} \left(\frac{1}{\sigma^2} - 1\right),$$

and

$$\mathrm{E}\{\hat{\lambda}_{LP,0}\} = \frac{a^2\sigma_K^2}{2} \frac{\phi(x)\phi_\sigma(x)}{[\phi(x) + \phi_\sigma(x)]^2} \left(x^2[1 - \sigma^{-4}] - 1 + \sigma^{-2}]\right).$$

Thus for the case $p = 0$ the asymptotic bias involves x^2 explicitly, whereas the $p = 1$ case involves only terms in $\lambda(x)$.

The relative bias, that is the bias for the $p = 0$ case divided by that of the $p = 1$ case, can easily be shown to be

$$1 - x^2(1 + \sigma^{-2}).$$

Thus, for the centre of the distribution of data points, the bias of the $p = 0$ estimator will actually be less than that of the $p = 1$ case, but as x moves further away from zero, the absolute size of bias will eventually become larger than that of $p = 1$, with the exact point at which this happens depending upon the value of σ^2.

2.3 Computational Issues

For the estimate of $\lambda(x)$, we require the value of β_0 from the maximising parameter vector $(\hat{\beta}_0, \hat{\beta}_1)$ for equation (2.2) when $p = 1$. Taking the first partial derivatives with respect to β_0 and β_1 gives

$$\frac{\partial Q}{\partial \beta_0} = \sum_{i=1}^{n} (Y_i - \hat{\mu}_i(x))\, K_h(X_i - x), \tag{2.8}$$

$$\frac{\partial Q}{\partial \beta_1} = \sum_{i=1}^{n} (Y_i - \hat{\mu}_i(x))\, (X_i - x) K_h(X_i - x), \tag{2.9}$$

where

$$\hat{\mu}_i(x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1(X_i - x))}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1(X_i - x))}.$$

Simultaneously equating equations (2.8) and (2.9) to zero will obviously require iterative solution, as the parameters of interest β_0 and β_1 enter in a non-linear fashion, and it is this fact that is crucial to the computational burden.

Practically, by far the most important feature is to use a kernel function with bounded support. Use of a Gaussian kernel, although ensuring that the final estimate is infinitely differentiable, implies that every data point

contributes to the estimate at a particular point. Obviously some points many bandwidths away will have exceedingly small weights, but the use of a bounded kernel means that at each estimation point only a subset of data points must be considered.

If an algorithm based upon a quartic kernel is used, this allows evaluation at each point using only a subset of the total data. Using the example from the previous section of a Gaussian linear shift model with a difference of 1 between the mean values of the two densities, for a bandwidth of $h = 1$, and $n = 200$, on average the maximum number of data values contributing to the estimate at a single point was approximately 120 (60%), and quite often considerably less. When implemented as a C++ program called from Splus running on a SPARCstation 20, the average time to estimate the regression function on a grid of 400 points was 430ms. This is more than 10 times slower than the single bandwidth fully nonparametric estimator λ_a, which averaged 38ms for the same dataset. Although we are talking in terms of milliseconds rather than seconds, this has important implications for bandwidth selection procedures such as cross-validation and simulation experiments which are both situations in which estimates are calculated for many different bandwidths.

When estimating the regression function on a grid of points, one of the main areas of user control is in the selection of starting values for β_0 and β_1 in the iterative algorithm. The example above used a default of $\beta_0 = \beta_1 = 0$, equivalent to a fitted probability of 0.5, for each estimation point, but the algorithm can be substantially accelerated by using the estimate from the previous gridpoint as the starting value. As the regression function λ is smooth and continuous, the speed of convergence should be greater. Indeed, in the example used above, applying this technique reduced the

average evaluation time to 125ms, a saving of 70%, and only 3-4 times slower than estimating λ_a. Thus, although at first glance these semiparametric estimators would seem to be far more computationally expensive, simply by choosing a fixed-width kernel and reusing the parameter estimates as starting values, we can greatly reduce the differences in evaluation time.

2.4 Simulation Experiment

As in the previous chapter, the asymptotic comparison of these estimators with the fully nonparametric ones is both intractable and not entirely relevant to the practical small-sample case, and we again proceed by simulation. Furthermore, although the locally linear logistic model is computationally feasible, it is still of interest to compare it with the simpler and faster local linear estimator, i.e. apply the usual local linear smoother *directly* to the binary data, avoiding the need to use the logit transformation. This approach can be considered as an extension of the local constant nonparametric estimator of Copas.

Unlike both the locally linear logistic and the nonparametric methods, however, the local linear method is not constrained to lie between 0 and 1. This problem often occurs close to the minimum and maximum x values, and Figure 2.1 shows a typical case.

Here, at the lower end of the x-scale, there are a group of failures ($Y = 0$) between -1.1 and -1. This clump of points lies in the tail of the distribution of failures and it is clear that it produces the 'dip' in the estimate of λ. The trend in this estimate is still increasing as x decreases, however, and the resulting rapid rise in $\hat{\lambda}$ between -1.3 and -1.1 is continued until the estimate of probability is greater than 1. This can result in large contributions to the overall WISE from a relatively small interval of estimation.

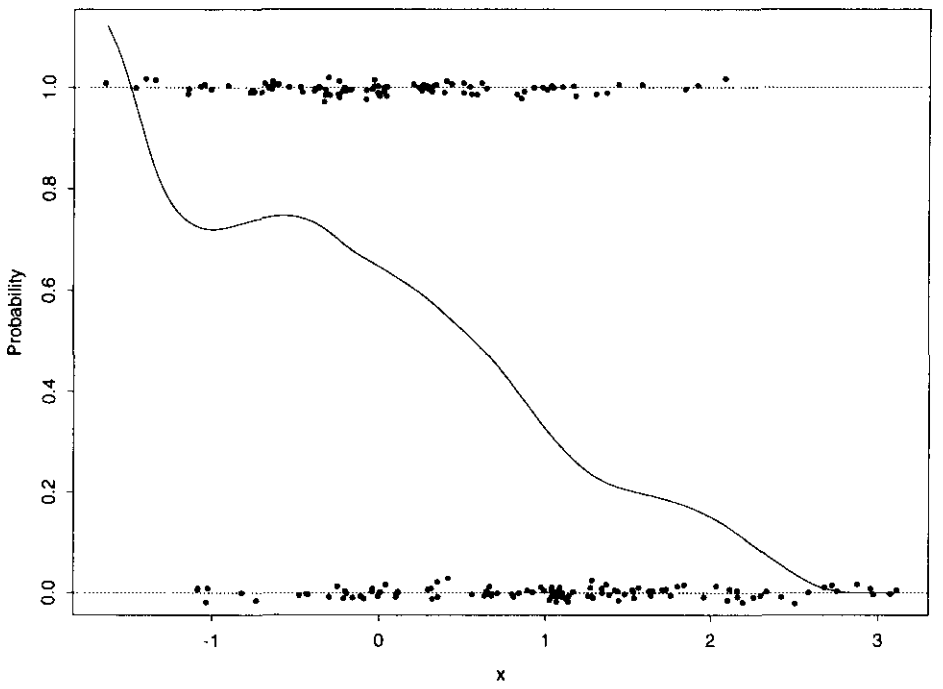

Figure 2.1: Failings of using the local linear smoother directly with binary data

To make matters worse, below $X_{(1)}$, the estimate continues to increase. Traditionally, and with very good reason, regression estimates are only calculated within the range of the data, i.e. in $[X_{(1)}, X_{(n)}]$, where $X_{(i)}$ are the order statistics of the x values. All of the methods previously discussed, with the exception of the local linear approach, are constrained to lie between 0 and 1 for the whole of the interval $[X_{(1)} - h, X_{(n)} + h]$. To prevent these boundary effects from swamping the overall WISE and distorting our results, all the squared errors for the semiparametric models were calculated on the interval $\left[\max\left(X_{(1)}, -4\right), \min\left(X_{(n)}, 4\right)\right]$. Note that this differs from that used to compare the nonparametric estimators, which are always either between 0 and 1 or undefined.

In an effort to avoid this problem of illegal probability estimates, and in

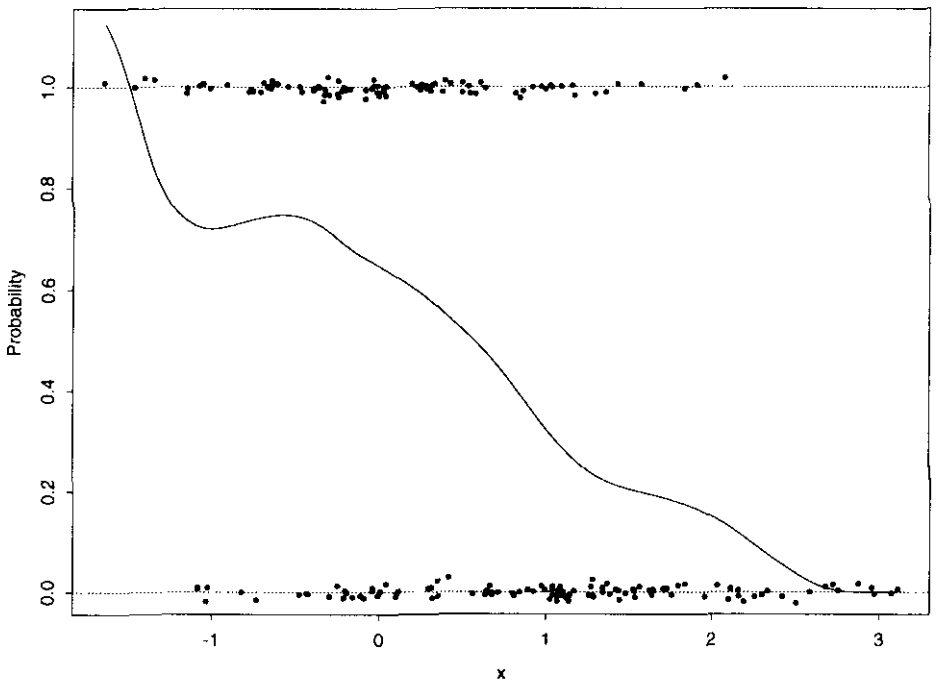

Figure 2.1: Failings of using the local linear smoother directly with binary data

To make matters worse, below $X_{(1)}$, the estimate continues to increase. Traditionally, and with very good reason, regression estimates are only calculated within the range of the data, i.e. in $[X_{(1)}, X_{(n)}]$, where $X_{(i)}$ are the order statistics of the x values. All of the methods previously discussed, with the exception of the local linear approach, are constrained to lie between 0 and 1 for the whole of the interval $[X_{(1)} - h, X_{(n)} + h]$. To prevent these boundary effects from swamping the overall WISE and distorting our results, all the squared errors for the semiparametric models were calculated on the interval $\left[\max\left(X_{(1)}, -4\right), \min\left(X_{(n)}, 4\right)\right]$. Note that this differs from that used to compare the nonparametric estimators, which are always either between 0 and 1 or undefined.

In an effort to avoid this problem of illegal probability estimates, and in

addition to the standard local linear estimator, we also considered a version which was simply truncated to zero or one when it was outside these boundaries. This is a similar approach to that taken when estimating densities using fourth or higher order kernels, where in regions of low density the estimate can sometimes be negative, as discussed, for example, by Hall and Murison [25]. Applying the local linear method to the linear shift example given in Section 2.3 gave an approximate time of 50ms per evaluation, indicating a significant computational advantage over the more correct locally linear logistic model.

So, the three semiparametric estimators which were compared to each other, and to the nonparametric estimators of the previous chapter were

- Locally linear logistic regression - λ_{LLL},
- Locally linear regression - λ_L,
- Locally linear regression truncated to $[0, 1]$ - λ_T.

To assess the practical small-sample performance of the semiparametric estimators in comparison to the nonparametric estimators, a simulation experiment was performed in which these estimators were applied to exactly the same 24 models shown in Figure 1.7, and using the same datasets from the previous experiments. A quartic kernel was used and the selection of the WISE-optimal bandwidth h proceeded as before by a grid search over a wide range of possible values. For some of the models, notably those involving either a linear shift or a difference in proportions (the first nine models), it was found that, for all reasonable values of h, the WISE function was decreasing with no apparent minimum. These models are linear on the logit scale, and hence can be excellently fitted by a logistic regression model. As stated previously, as $h \to \infty$, the locally linear logistic smoother converges

to the logistic model. It was assumed that this was the situation in these cases, and to prevent endless searching for the minimum WISE, an arbitrary upper limit of $h = 10$ was taken.

Each model was simulated 100 times, with a sample size of $n = 200$. Even using the computational techniques discussed in Section 2.3, the simulations still required periods of several days rather than hours to perform, implying that more extensive experiments would be an onerous undertaking.

2.5 Practical Performance

The results of the simulation experiments are presented separately for the comparisons between the semiparametric estimators, and between the semiparametric and the nonparametric estimators. Median WISE values are compared and a Wilcoxon Signed Rank test used to test for statistically significant differences between the various estimators.

2.5.1 Comparisons between Semiparametric Estimators

Taking first the locally linear logistic smoother λ_{LLL}, and comparing this with both the unmodified version of the local linear estimate λ_L and the truncated version λ_T , we get the results shown in Tables 2.1 and 2.2. Median WISE values are presented, as is the median percentage increase over the locally linear logistic estimator for each of the local linear estimators. The p-values reported are for a Wilcoxon test of the difference in WISE values between λ_{LLL} and the local linear estimators.

The first observation to be made from these results are the very small differences between the unconstrained and the truncated forms of the local linear estimator, suggesting that although this may be theoretically a problem, it is less troublesome in practice. As expected, the truncated ver-

Model	λ_{LLL}	λ_L			λ_T		
	Median WISE	Median WISE	Increase (%)	p-value	Median WISE	Increase (%)	p-value
Linear Shift							
1 : $\mu = 0.5$	341	303	-8.61	0.00000	303	-8.61	0.00000
2 : $\mu = 0.75$	301	283	-1.73	0.00893	282	-1.73	0.00849
3 : $\mu = 1$	277	298	-0.04	0.60243	286	-1.28	0.49928
4 : $\mu = 1.25$	245	234	3.66	0.61930	229	3.21	0.50585
5 : $\mu = 1.5$	231	300	44.30	0.00016	299	44.14	0.00028
Different Proportions							
6 : $\pi_1 = 0.2$	201	261	45.35	0.00002	261	44.72	0.00004
7 : $\pi_1 = 0.4$	199	229	15.20	0.53939	219	14.70	0.60243
8 : $\pi_1 = 0.6$	253	276	18.38	0.07351	274	18.38	0.10796
9 : $\pi_1 = 0.8$	234	251	33.71	0.03260	248	33.38	0.04338
Different Variance							
10 : $\sigma = 0.2$	1291	1325	9.70	0.00003	1325	9.70	0.00003
11 : $\sigma = 0.5$	826	886	0.60	0.77140	878	0.29	0.83253
12 : $\sigma = 0.8$	620	568	-3.28	0.00000	568	-3.71	0.00000

Table 2.1: Comparison of local linear logistic method with local linear methods, Models 1 to 12. WISE values are $\times 10^6$, and p-value are from a Wilcoxon signed rank test compared to λ_{LLL}.

Model	λ_{LLL}	λ_L			λ_T		
	Median WISE	Median WISE	Increase (%)	p-value	Median WISE	Increase (%)	p-value
Cauchy							
13 : $\mu = 0.6$	416	386	-1.75	0.00407	384	-1.75	0.00122
14 : $\mu = 0.8$	421	396	-2.95	0.01017	395	-2.99	0.00463
15 : $\mu = 1$	452	485	3.11	0.15410	482	2.21	0.23486
16 : $\mu = 1.2$	373	422	5.15	0.00567	422	4.33	0.01298
Marron-Wand							
17 : MW(2)	595	617	4.73	0.00003	617	4.68	0.00005
18 : MW(3)	292	383	19.99	0.00050	371	14.40	0.00650
19 : MW(4)	3242	3227	-3.28	0.08973	3126	-4.28	0.00000
20 : MW(5)	2192	2652	18.15	0.00000	2585	14.07	0.00000
21 : MW(6)	974	942	-1.76	0.00000	926	-1.79	0.00000
22 : MW(7)	957	891	-3.62	0.05311	891	-3.63	0.04707
23 : MW(8)	864	822	-0.17	0.02076	822	-0.17	0.01775
24 : MW(9)	964	959	-0.81	0.03704	946	-0.82	0.02765

Table 2.2: Comparison of local linear logistic method with local linear methods, Models 13 to 24. WISE values are $\times 10^6$, and p-value are from a Wilcoxon signed rank test compared to λ_{LLL}.

sions give slightly smaller WISEs, but the median percentage decrease is never more than 5%. For applications of binary regression which form only a component of a larger procedure, the lack of differentiability caused by truncation may be a problem, but for exploratory data analysis and model checking the truncated estimate λ_T may be reasonably used.

Comparing the local linear logistic estimator λ_{LLL} with the local linear estimates, we can see that, although the logistic estimator λ_{LLL} is not always optimal, when it is worse, the median decrease in WISE is always less than 10%, and when it is an improvement over the linear methods, the improvement can be quite large.

There appears to be no clear pattern where the locally linear smoothing estimators are highly inferior. For models 1 to 4 λ_L performs quite adequately, but for models 5 to 9 λ_{LLL} is clearly superior. All these models, however, are linear on the logistic scale, and it may be that for the models involving different proportions the fact that the underlying overall density h is skewed rather than symmetric can explain the difference in median WISE values.

Improvements over locally linear logistic smoothing can also be seen for models 12, 13 and those involving the Marron-Wand densities 6 through 9 (models 21 to 24). These last four models are the only ones to give g a clear multimodal structure, and this feature may dominate the need for consistent probability estimates between 0 and 1, with the result that the local linear models alone can achieve good performance, although the logistic estimators are not very far behind.

Interestingly, models 18, 19 and 20, all of which have the general form of a relatively sharp trough of probability in an otherwise constant function (see Figure 1.7), show somewhat different comparative performances. For

models 18 and 20, the local linear logistic estimator is clearly better, with a median improvement in the WISE of approximately 20%. For model 19, however, there is no clear difference between the estimators. The fact that, unlike models 18 and 20, the trough of probability for model 19 does not span the entire range from 0 to 1 may account for this peak being unresolvable with only 200 data points, as then the estimators will be fitting a constant probability.

2.5.2 Comparisons with Nonparametric Estimators

To compare non- and semiparametric models, we must first standardise the error measure. The optimal WISEs were recalculated for both λ_a and $\lambda_{a,c}$, but on the interval $\left[\max\left(X_{(1)}, -4\right), \min\left(X_{(n)}, 4\right)\right]$ only. Note that this could imply that the WISE-optimal bandwidth could change as well as the WISE value itself. These values were then compared to the results previously obtained for λ_{LLL}, with the results shown in Tables 2.3 and 2.4.

The results here are somewhat clearer than the previous section. When compared to the single bandwidth version of the nonparametric estimator λ_a, the locally linear estimator is nearly always superior, and when it is not, the losses are relatively small. Obviously λ_{LLL} is performing better for models 1 to 9, but this is hardly surprising. This is a situation where the parametric part of λ_{LLL} is correct, so we would expect the semiparametric model to outperform the nonparametric one. For the rest of the models, the benefits of the locally linear logistic method are smaller but still mostly favourable.

When compared to the two bandwidth version of the nonparametric estimator, however, there are clear cases where the locally linear logistic estimator is much worse than the two bandwidth version. For models 10,

Model	λ_{LLL}	λ_a			$\lambda_{a,c}$		
	Median WISE	Median WISE	Increase (%)	p-value	Median WISE	Increase (%)	p-value
Linear Shift							
1 : $\mu = 0.5$	341	458	35.35	0.00001	311	-5.70	0.15213
2 : $\mu = 0.75$	301	531	110.16	0.00000	403	64.93	0.02576
3 : $\mu = 1$	277	596	103.14	0.00000	548	72.07	0.00013
4 : $\mu = 1.25$	245	388	69.62	0.00001	372	49.95	0.00015
5 : $\mu = 1.5$	231	462	130.19	0.00000	453	121.15	0.00000
Different Proportions							
6 : $\pi_1 = 0.2$	201	335	72.78	0.00001	287	28.16	0.11334
7 : $\pi_1 = 0.4$	199	500	157.67	0.00000	411	124.95	0.00001
8 : $\pi_1 = 0.6$	253	557	99.66	0.00000	496	73.64	0.00000
9 : $\pi_1 = 0.8$	234	336	59.49	0.00000	279	34.34	0.01877
Different Variance							
10 : $\sigma = 0.2$	1291	1318	7.81	0.29830	472	-48.68	0.00000
11 : $\sigma = 0.5$	826	962	4.85	0.23216	453	-41.67	0.00095
12 : $\sigma = 0.8$	620	761	2.76	0.35056	393	-41.53	0.00790

Table 2.3: Comparison of local linear logistic method with nonparametric methods, Models 1 to 12. WISE values are $\times 10^6$, and p-value are from a Wilcoxon signed rank test compared to λ_{LLL}.

Model	λ_{LLL}	λ_a			$\lambda_{a,c}$		
	Median WISE	Median WISE	Increase (%)	p-value	Median WISE	Increase (%)	p-value
Cauchy							
13 : $\mu = 0.6$	416	476	22.23	0.00003	384	2.26	0.94381
14 : $\mu = 0.8$	421	550	33.37	0.00014	415	5.82	0.75306
15 : $\mu = 1$	452	529	26.69	0.00003	479	10.49	0.12728
16 : $\mu = 1.2$	373	539	28.28	0.00003	484	17.07	0.00458
Marron-Wand							
17 : MW(2)	595	672	15.04	0.00006	477	-4.73	0.95750
18 : MW(3)	292	335	4.24	0.27347	269	-6.89	0.02486
19 : MW(4)	3242	3136	-2.05	0.08342	2101	-32.81	0.00000
20 : MW(5)	2192	3074	16.36	0.01894	1373	-27.66	0.00001
21 : MW(6)	974	1036	-0.75	0.55772	937	-10.29	0.00911
22 : MW(7)	957	783	-2.70	0.33482	693	-20.97	0.00004
23 : MW(8)	864	818	-2.70	0.44631	601	-19.46	0.00001
24 : MW(9)	964	917	-5.92	0.12814	804	-12.09	0.00010

Table 2.4: Comparison of local linear logistic method with nonparametric methods, Models 13 to 24. WISE values are $\times 10^6$, and p-value are from a Wilcoxon signed rank test compared to λ_{LLL}.

11 and 12, which are *quadratic* on the logistic scale, the nonparametric estimator is approximately 50% worse for each model. Similarly for the less monotonic models in which g is taken from the Marron-Wand suite of densities, the two bandwidth method can improve significantly upon the single bandwidth semiparametric one.

2.6 Conclusions

We have shown in this chapter that semiparametric methods are serious contenders for general binary regression problems. The gains in terms of coherent probabilities and weighted integrated squared errors from using the logistic formulation rather than the simple linear version of the smoother more than outweigh the losses in terms of slower computation.

Moreover, the single bandwidth locally linear logistic estimator is at least as good as the fully nonparametric single bandwidth estimator, and with improved asymptotic bias and boundary effects. Indeed, the estimator is only bettered by the two bandwidth version in certain cases, once again where the densities of successes and failures differ markedly. For these cases it is possible to modify the locally linear logistic estimator to also incorporate two bandwidths and this is done in the next chapter.

The fact, however, that a single bandwidth procedure can do as well as a more complicated two bandwidth procedure in most circumstances, has important practical consequences. It is obviously easier to select a single bandwidth than a pair of bandwidths. For the admittedly sub-optimal example of cross-validation, we would be evaluating the optimising function on a grid rather than a sequence of points and obviously squaring the number of evaluations required.

When considering the WISE-optimal bandwidths, the locally linear lo-

gistic method achieves uniformly larger bandwidths over all models. This can be attributed to the fact that the smoothing is effectively taking place in the logit-transformed space rather than the x-space directly. Bandwidth selection for these models is discussed in Chapter 4, but Fan, Heckman and Wand [4] derive a crude plug-in estimate of bandwidth, which for the burn data we have previously used, gives an estimated bandwidth for λ_{LLL} of 1.242. This can be contrasted with the estimates of individual bandwidths from the last chapter of 0.417 for successes and 0.309 for failures.

We have demonstrated both that in the nonparametric case that two bandwidths are better than one, but that if we are constrained to only one bandwidth, then the semiparametric locally linear logistic estimator is best. For completeness, we now extend the locally linear logistic model to two bandwidths, before finally attempting to find practical bandwidth estimators which come close to achieving these optimal performances.

Chapter 3

Two Bandwidth Semiparametric Binary Regression

3.1 Introduction

Consider the basic quasi-likelihood equation for the single bandwidth locally linear logistic model, a kernel-weighted form of equation (2.1),

$$Q(x) = \sum_{i=1}^{n} (Y_i \ln [\mu([X_i - x])] + (1 - Y_i) \ln[1 - \mu([X_i - x])]) K_h(X_i - x),$$

where $\text{logit}[\mu(z)]$ is a polynomial of order p in z.

This can be thought of as having two components; one for the successes ($Y_i = 1$) and one for the failures ($Y_i = 0$). As was demonstrated for the nonparametric case, for situations where the distributions of successes and failures have substantially different shapes, a two bandwidth procedure can often improve estimation. To apply this philosophy to the above estimator, we replace the general weighted quasi-likelihood with a single kernel and

bandwidth h by a pair of kernels with bandwidths a and c.

$$Q(x) = \sum_{i=1}^{n} \{Y_i \ln [\mu([X_i - x])]\} K_a(X_i - x) + \sum_{i=1}^{n} \{(1 - Y_i) \ln[1 - \mu([X_i - x])]\} K_c(X_i - x). \quad (3.1)$$

To see the connection with the nonparametric estimator, if we set $p = 0$, so that $\mu([X_i - x]) = \mu = (1 + e^{-\beta_0})^{-1}$, then equation (3.1) is maximised when

$$\frac{\partial Q}{\partial \beta_0} = 0,$$

and hence the solution satisfies

$$(1 - \hat{\mu}) \sum_{i=1}^{n} K_a(X_i - x) Y_i = \hat{\mu} \sum_{i=1}^{n} K_c(X_i - x)(1 - Y_i).$$

But the term on the left-hand side is simply $(1 - \hat{\mu}) m \hat{f}_a$, and the term on the right-hand side is $\hat{\mu} n \hat{g}_c$, and so we have

$$\hat{\mu}(x) = \frac{m \hat{f}_a}{m \hat{f}_a + n \hat{g}_c} = \lambda_{a,c}(x).$$

Thus, once again the $p = 0$ case corresponds to the nonparametric case, this time with two bandwidths.

Alternatively, if we were to use the other form of the quasi-likelihood for binomial models, namely

$$Q(y, \mu) = y \operatorname{logit}(\mu) + \log(1 - \mu),$$

and weight each term on the right hand side with a kernel of differing bandwidths, then setting $p = 0$ would result in the solution being $\lambda_{a,b}$, a two bandwidth estimator we have previously rejected as it is not bounded between 0 and 1.

In terms of the WISE performance of this, our most complicated estimator, it is clear that for the $p = 1$ case that we are in exactly the same

situation as for the nonparametric case ($p = 0$). The two bandwidth locally linear logistic estimator, which we shall denote by $\lambda_{LLL,2}$, will always give a smaller optimal WISE than the identical but restricted single bandwidth version, so the question of interest is concerned with the absolute size of this improvement.

3.2 Practical Performance

The two-bandwidth locally linear logistic estimator was implemented and the practical performance assessed in the context of the previously described simulated datasets.

3.2.1 Computational Issues

For the two bandwidth case with $p = 1$, the estimation of $\hat{\mu}(x)$ is very similar to the single bandwidth case, λ_{LLL}. The evaluation of the first and second derivatives which are required for the iterative solution at each estimation point is merely decomposed into separate 'success' and 'failure' parts and combined to produce the iterative adjustments. Thus the computational complexity is not substantially increased and evaluation times are similar.

Essentially, equations (2.8) and (2.9) become respectively,

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= \sum_{i=1}^{m} (1 - \hat{\mu}_i(x))\, K_h(X_i - x) - \sum_{i=m+1}^{m+n} \hat{\mu}_i(x) K_h(X_i - x), \\
\frac{\partial Q}{\partial \beta_1} &= \sum_{i=1}^{m} (1 - \hat{\mu}_i(x))\, (X_i - x) K_h(X_i - x) - \sum_{i=m+1}^{m+n} \hat{\mu}_i(x)(X_i - x) K_h(X_i - x),
\end{aligned}$$

where the data is ordered so that the first m points are the successes and the next n points the failures. Similar expressions can be derived for the second derivative terms.

3.2.2 Simulation Experiment

Using exactly the same models and simulated datasets as before, the minimum WISE for the two bandwidth estimator $\lambda_{LLL,2}$ was determined using a quasi-Newton minimisation algorithm starting from the single bandwidth. The previous approach of using a grid search was computationally infeasible for this estimator.

As for the single bandwidth case, an arbitrary limit of $h = 10$ was used to bound the bandwidths for the models which were linear on the logistic scale, thus all WISE functions were optimised over the square $[0, 10] \times [0, 10]$.

Furthermore, to return to the performance criterion used in Chapter 1, the WISE was calculated on the range $[-4, 4]$, and not restricted to the range of the data, as this latter restriction was only introduced to deal with the estimators which were not constrained to lie between 0 and 1.

3.2.3 Results

A summary of the WISE-optimal estimates for both λ_{LLL} and $\lambda_{LLL,2}$ is shown in Tables 3.1 and 3.2.

The reduction in WISE for the two-bandwidth version of the locally linear logistic estimator is dramatic. The largest gains seem to be made in the first nine simulated models, all of which are truly linear on the logistic scale. The smallest gains are achieved in the highly skewed and multimodal failure densities of models 17 to 24, but even in these cases the median reduction in WISE is between 7.5% and 27.5%.

Closer examination of these remarkable gains in the accuracy of estimation, however, shows an interesting practical phenomenon. To see this, consider Figure 3.1 which shows the WISE-optimal single and double bandwidth estimators λ_{LLL} and $\lambda_{LLL,2}$ for a dataset from Model 1. This dataset

Model	**Median WISE**		*Median Percentage*
	λ_{LLL}	$\lambda_{LLL,2}$	Improvement
Linear Shift			
1 : $\mu = 0.5$	341	66	69.00
2 : $\mu = 0.75$	301	68	55.11
3 : $\mu = 1$	277	65	64.66
4 : $\mu = 1.25$	245	62	68.73
5 : $\mu = 1.5$	231	66	60.85
Different Proportions			
6 : $\pi_1 = 0.2$	201	32	76.83
7 : $\pi_1 = 0.4$	199	23	77.51
8 : $\pi_1 = 0.6$	253	55	69.92
9 : $\pi_1 = 0.8$	234	37	80.27
Different Variance			
10 : $\sigma = 0.2$	1291	384	47.76
11 : $\sigma = 0.5$	834	445	41.35
12 : $\sigma = 0.8$	620	269	51.97

Table 3.1: Median optimal WISE values and median percentage improvement for comparison of single and two bandwidth local linear logistic methods, Models 1 to 12. WISE values are $\times 10^6$.

Model	**Median WISE**		Median Percentage
	λ_{LLL}	$\lambda_{LLL,2}$	Improvement
Cauchy			
13 : $\mu = 0.6$	416	221	41.22
14 : $\mu = 0.8$	421	235	27.61
15 : $\mu = 1$	452	223	29.09
16 : $\mu = 1.2$	373	234	32.02
Marron-Wand			
17 : MW(2)	609	355	25.82
18 : MW(3)	293	157	27.45
19 : MW(4)	3249	2662	16.52
20 : MW(5)	2214	1598	22.75
21 : MW(6)	986	765	8.10
22 : MW(7)	974	697	17.40
23 : MW(8)	868	644	12.58
24 : MW(9)	976	797	7.66

Table 3.2: Median optimal WISE values and median percentage improvement for comparison of single and two bandwidth local linear logistic methods, Models 13 to 24. WISE values are $\times 10^6$.

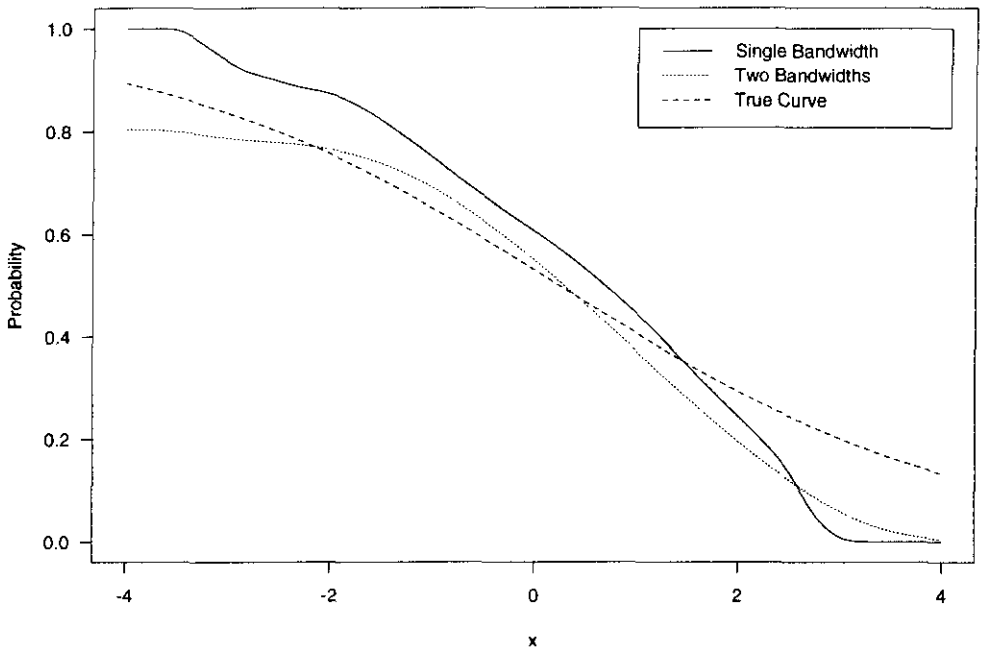

Figure 3.1: Comparison of single and two bandwidth WISE-optimal estimates with the true probability curve for a simulated dataset from Model 1

was selected from those where the relative reduction in WISE was around the upper quartile of the observed range. The single bandwidth estimator has a minimum achievable WISE of 1302×10^{-6} at $a = 1.8$, whereas the two bandwidth version can reduce this by over three-quarters to 296×10^{-6} at the bandwidths $a = 5.4, c = 4.1$. The plot suggests, however, that this is achieved by shifting the mean level of the estimate rather than any fine-tuning of the shape of the estimate itself. Indeed, the average value of λ_{LLL} over the whole range of x is 0.548, whereas $\lambda_{LLL,2}$ has an average of 0.478.

The WISE-surface, plotted in Figure 3.2 as a function of the two bandwidths, shows that the minimum WISE values lie along a valley which is parallel but not coincident with the line of equality $a = c$, where $a > c$.

Returning to the dataset, we note that in this case there were 109 successes and only 91 failures. Thus the empirical estimate of π_1, the propor-

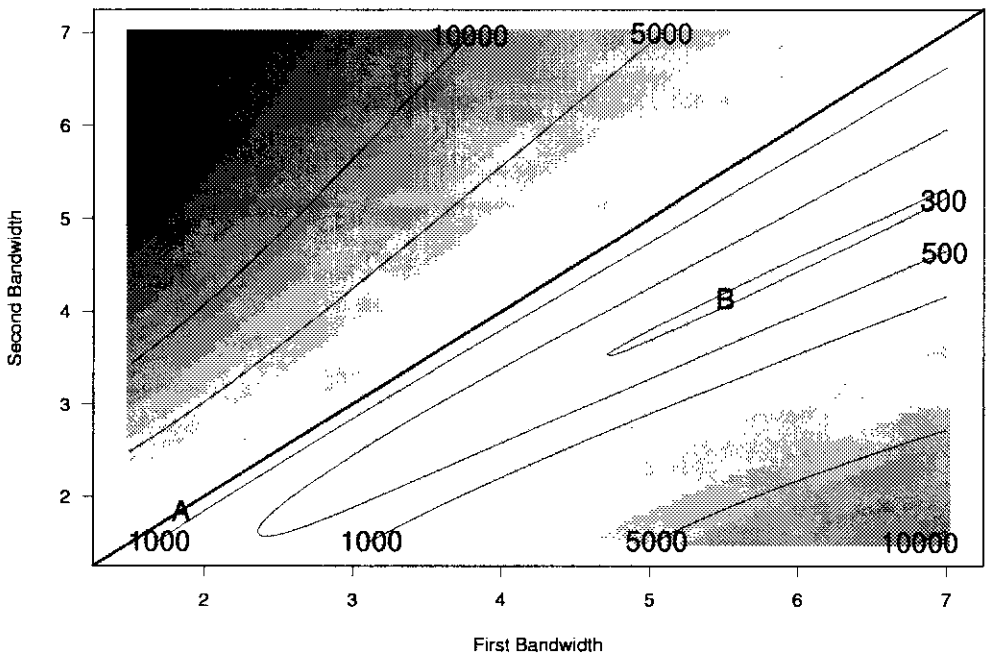

Figure 3.2: WISE contour plot for a simulated dataset from Model 1. WISE values are $\times 10^6$.

tion of successes, is $109/200 = 0.545$, very close to the mean of the single bandwidth estimate. As asymptotically-optimal bandwidths are functions of sample size, this partly explains the fact that the WISE-optimal solution occurs when $a > c$. So it would seem that the two bandwidth estimator is achieving an improvement in the WISE not by estimating the densities of failures and successes more accurately, but by correcting the errors in the estimation of π_1 itself.

To explore this phenomenon further, for each simulated dataset, the absolute value of the difference between the observed number of successes and the expected number (100 for all except models 6 to 9) was calculated. The correlation between this measure of the contribution of the sampling variability in π_1 to the accuracy of the probability estimates and the absolute

improvement in WISE allowed by using two bandwidths instead of one was then calculated. These correlations are presented in Table 3.3.

Clearly, for most models the improvements achieved for the two bandwidth version of the locally linear logistic estimator over the single bandwidth version have a significant component which is due to the estimation of π_1. It is only for Models 10, 11, 18, 19 and 20 that the gain seems to be due entirely to the capability to adapt to the densities f and g separately, as in this case the variance associated with g is substantially less than 1.

Asymptotically, it would appear that the $O(n^{-1/2})$ errors in the estimation of π_1, which up to now we have assumed to be dominated by the errors in the estimation of f and g, are actually having a significant influence on the error of the estimate.

Even for the models where the correlation is small, however, the apparent gains in WISE are at times counter-intuitive. Figure 3.3 shows an example from Model 11, where the density g is from a Gaussian distribution with mean 0.5 and standard deviation 0.5.

In this case the single bandwidth WISE-optimal solution is given by $h = 0.8375$, with a WISE of 4029×10^{-6}. This can be reduced by nearly 50% by the two bandwidth estimator $\lambda_{LLL,2}$ with $a = 1.68$ and $c = 0.524$ which gives a WISE of 2079×10^{-6}. Notice that the two bandwidth solution achieves this reduction by adjusting to the different variabilities of f and g, rather than a simple calibration, as in this dataset there are only 102 successes and 98 failures. This improves the estimation of $\lambda(x)$ in the main peak, but at the expense of the anomalous minor peak at $x = 2$. As the ISE is weighted by $h(x)^2$, however, the increase in error from the single bandwidth case is more than cancelled out by the improvement in the range $x \in [0, 1]$, where the overall density h is maximal. Thus the 'improvement'

Model	Correlation	Model	Correlation
Linear Shift		**Cauchy**	
1 : $\mu = 0.5$	0.911	**13** : $\mu = 0.6$	0.561
2 : $\mu = 0.75$	0.821	**14** : $\mu = 0.8$	0.574
3 : $\mu = 1$	0.747	**15** : $\mu = 1$	0.384
4 : $\mu = 1.25$	0.631	**16** : $\mu = 1.2$	0.423
5 : $\mu = 1.5$	0.549	**Marron-Wand**	
Different Proportions		**17** : MW(2)	0.562
6 : $\pi_1 = 0.2$	0.733	**18** : MW(3)	-0.050
7 : $\pi_1 = 0.4$	0.650	**19** : MW(4)	-0.149
8 : $\pi_1 = 0.6$	0.663	**20** : MW(5)	0.002
9 : $\pi_1 = 0.8$	0.833	**21** : MW(6)	0.463
Different Variance		**22** : MW(7)	0.372
10 : $\sigma = 0.2$	0.045	**23** : MW(8)	0.689
11 : $\sigma = 0.5$	0.073	**24** : MW(9)	0.411
12 : $\sigma = 0.8$	0.621		

Table 3.3: Pearson correlation coefficients between absolute imbalance in observed successes and absolute decrease in optimal WISE for two bandwidths over one

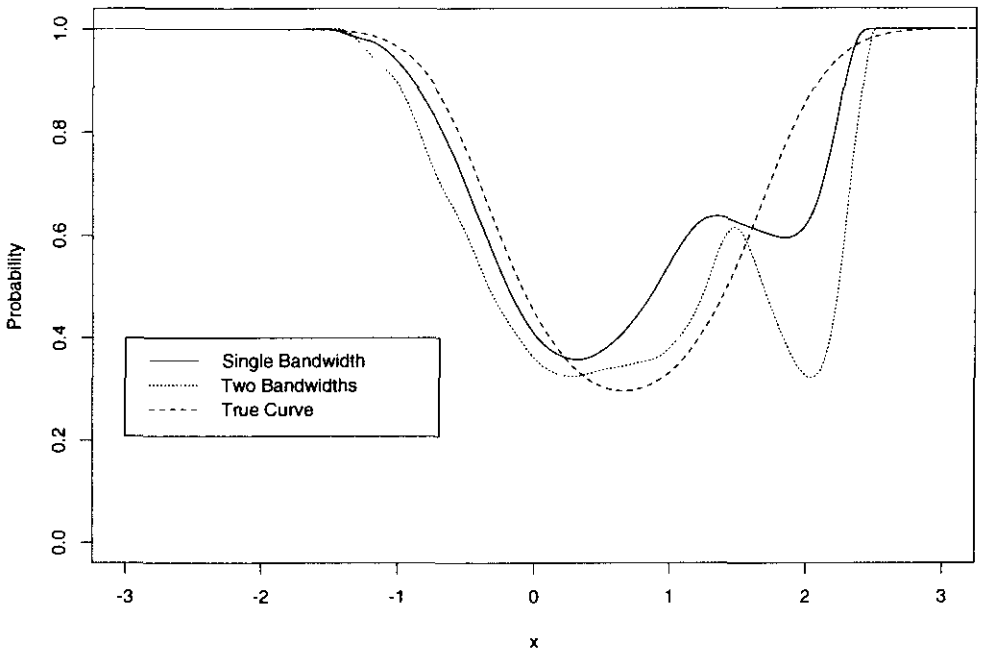

Figure 3.3: Comparison of single and two bandwidth WISE-optimal estimates with the true probability curve for a simulated dataset from Model 11

in WISE does seem to be at the expense of a subjectively large error in predicted probability of success at $x = 2$, where the true probability is 0.9, but the fitted value from $\lambda_{LLL,2}$ is less than 0.4.

3.3 Conclusions

We have extended the locally linear logistic estimator of the previous chapter to a two bandwidth version. The particular form of the two bandwidth weighted quasi-likelihood chosen was derived by analogy with the $p = 0$ case where $\lambda_{a,c}$ is bounded but $\lambda_{a,b}$ is not.

At first glance, the simulation results suggested that very dramatic gains could be realised in almost all circumstances by using two bandwidths. Closer inspection, however, revealed that the improvements are driven not

by better estimation of the component densities f and g, but by improved estimation of the proportion of successes π_1. This has important practical consequences as, in almost all real practical situations, π_1 will be unknown. The maximum likelihood estimate $\hat{\pi}_1 = m/s$ was, however, only improved upon in the simulation experiment because the true probability function $\lambda(x)$ (and hence π_1) was known and we sought a WISE-optimal solution.

Thus, although the extension of the locally linear logistic estimator to two bandwidths is useful for completeness, the true practical improvements achieved are likely to be considerably less than that observed in the simulation experiment, and will probably not justify the increased complexity of this estimator.

Chapter 4

Bandwidth Selection for Binary Regression

4.1 Introduction

Previous chapters have explored the properties of a variety of single and double bandwidth estimators for the binary regression problem. In the simulation experiments, by using the WISE-optimal bandwidth in every case, the problem of choosing an estimator was separated from that of choosing a bandwidth. It is clear that in certain circumstances more complex methods such as the two bandwidth nonparametric estimator or the locally linear logistic estimator may lead to improved estimation, but can these benefits be realised in practice with a data-dependent bandwidth selection procedure ?

This chapter extends and evaluates two contrasting approaches to the problem of bandwidth selection in binary regression: cross-validation and plug-in methods. The methods are applied to both the single and double bandwidth versions of the nonparametric and semiparametric kernel binary regression estimators, and the results compared in terms of WISE to that

best-possible case achieved from the simulation experiments of the previous two chapters.

4.2 Methods of Bandwidth Selection: Background

The majority of authors who have considered binary regression problems do not seem to have examined the problem of data-dependent bandwidth selection in any great detail, and indeed there are only two major suggestions of how to proceed, neither of which has undergone extensive practical evaluation.

4.2.1 Cross-validation

Kappenman [13], who extends Copas' original estimator $\hat{\lambda}_a(x)$ to two or more dimensions, takes as his starting point the log-likelihood of the data

$$L(a;X) = \sum_{j=1}^{n} \left(Y_j \log[\hat{\lambda}_a(X_j)] + (1 - Y_j) \log[1 - \hat{\lambda}_a(X_j)] \right). \tag{4.1}$$

However, it is clear that this will be maximised when the estimator $\hat{\lambda}_a(X_j)$ is equal to 1 for those X_j where $Y_j = 1$, and is equal to 0 for those where $Y_j = 0$. This can only be achieved if the bandwidth a tends to zero, and both $\hat{f}$ and $\hat{h}$ are each a sum of Dirac delta functions located at the successes for $\hat{f}$ and at all data points for $\hat{h}$. To avoid this, Kappenman suggests using the leave-one-out estimator

$$\hat{\lambda}_a^{(-j)}(X_j) = \frac{\sum_{i=1, i \neq j}^{n} Y_i K_a(X_j - X_i)}{\sum_{i=1, i \neq j}^{n} K_a(X_j - X_i)}. \tag{4.2}$$

Modifying equation (4.1), we then have the *likelihood cross-validation* (LCV) function to be maximised over a,

$$L(a;X) = \sum_{j=1}^{n} \left(Y_j \log[\hat{\lambda}_a^{(-j)}(X_j)] + (1 - Y_j) \log[1 - \hat{\lambda}_a^{(-j)}(X_j)] \right). \tag{4.3}$$

It is clear that this LCV method can be extended to the two bandwidth nonparametric estimator $\hat{\lambda}_{a,c}(x)$, and to both the single and double bandwidth semiparametric estimators $\lambda_{LLL}(x)$ and $\lambda_{LLL,2}(x)$. In each case the bandwidths are chosen so as to maximise the leave-one out likelihood

$$L(q;X) = \sum_{j=1}^{n} \left(Y_j \log[\hat{\lambda}^{(-j)}(X_j)] + (1 - Y_j) \log[1 - \hat{\lambda}^{(-j)}(X_j)] \right), \qquad (4.4)$$

where $\hat{\lambda}$ is the appropriate estimator, and q is a vector of either one or two bandwidths.

The idea of likelihood cross-validation is discussed in the context of density estimation by Silverman [26], where it is demonstrated heuristically that likelihood cross-validation is equivalent to minimising the Kullback-Leibler information loss function $\int f(x) \log[\hat{f}(x)/f(x)]dx$.

In density estimation, the more usual form of the technique is *least-squares cross-validation* [27, 28] which begins with the aim of estimating the mean integrated squared error (MISE)

$$\mathrm{E}\left(\int [f(x) - \hat{f}_a(x)]^2 dx \right). \qquad (4.5)$$

This equation can be expanded into three integral terms, involving only the unknown density f, only the estimate $\hat{f}_a(x)$, and a cross term involving both, respectively. This final term is estimated by cross-validation, and the whole expression minimised over the bandwidth a.

Although we have used a similar criterion throughout this work as a measure of estimator performance, to ensure the existence of the integral we have used a weighting function which is the square of the (unknown) combined density h. This implies that the expansion of the MISE function into terms involving only known or only estimated quantities does not apply, in that h must also be estimated. This precludes the use of a cross-validation

method based upon the weighted MISE, and as we shall see presently, also complicates matters for the plug-in estimator.

4.2.2 Plug-in Methods

In the initial development of the semiparametric approach, Fan, Heckman and Wand [4] describe a simple plug-in bandwidth selector based upon the asymptotic expression for the mean integrated squared error (MISE). Rather than use the error in the estimation of $\lambda_{LLL}(x)$ directly, however, they use the WISE of the estimate of the linear functional η, where

$$\eta(p) = \log\left(\frac{p}{1-p}\right).$$

In this case, the asymptotic expressions for bias and variance are given by

$$\mathrm{E}\left[\hat{\eta}_{LLL} - \eta\right](x) = \frac{a^2\sigma_K^2}{2}\eta''(x) + o(a^2) \tag{4.6}$$

$$\mathrm{var}\left[\hat{\eta}_{LLL}\right](x) = (sa)^{-1}R(K)\frac{1}{\lambda(x)[1-\lambda(x)]h(x)} + o((sa)^{-1}). \tag{4.7}$$

These values can then be substituted into the weighted MISE equation, expressed as a function of the bandwidth a, as

$$\mathrm{MWISE}(a) = \int \left(\mathrm{E}\left[\hat{\eta}_{LLL} - \eta\right](x)^2 + \mathrm{var}\left[\hat{\eta}_{LLL}\right](x)\right) w(x)h(x)\,dx, \tag{4.8}$$

where $w(x)$ is a "weighting function". The authors never actually specify what this weight function should be, other than to say that both $w(x)$ and $h(x)$ are included for "stability purposes". In all of our previous work we have taken this weighting function on the ISE to be $h(x)$, giving the standard weighting by $h(x)^2$, and we shall continue to do so here.

Substituting equations (4.6) and (4.7) into equation (4.8), we get an expression for the WMISE in terms of a.

$$\mathrm{MWISE}(a) = \frac{a^4(\sigma_K^2)^2}{4}\int[\eta''(x)h(x)]^2dx + \frac{R(K)}{sa}\int\frac{h(x)}{\lambda(x)[1-\lambda(x)]}dx.$$

Differentiating this with respect to a and equating to zero to find the minimum gives the WMISE-optimal bandwidth as

$$a_{OPT} = \left[\frac{R(K)I_2}{(\sigma_K^2)^2 I_1}\right]^{1/5} s^{-1/5}, \tag{4.9}$$

where

$$\begin{aligned} I_1 &= \int \left[\eta''(x)h(x)\right]^2 dx, \\ I_2 &= \int \frac{h(x)}{\lambda(x)[1-\lambda(x)]}\, dx. \end{aligned}$$

Now in equation (4.9) only the terms in these two integrals are unknown and have to be estimated from the data. Fan et al. suggest using a parametric pilot estimator for η and λ, and substitute this into the above equations to estimate a_{OPT}. The recommendation is to use a polynomial estimator of order $p+3$, which in this case would be a conventional logistic regression using a quartic polynomial in x.

Note, however, that there is another unknown quantity in the above equations in addition to λ, namely the density of all the data points $h(x)$. If, as is implied in the original publications w (and hence h) is known, then both I_1 and I_2 can be calculated directly from the parametric pilot estimate. However, as a direct consequence of the choice of weighting function, we must also estimate h. This is a practical aspect of this bandwidth selector which is not discussed in the original work, but as we shall see, the use of pilot estimates of h and related quatities can be used to give satisfactory results.

As we have previously calculated expressions for the asymptotic bias and variance for the estimators $\hat{\lambda}_a(x)$, $\hat{\lambda}_{a,c}(x)$, $\hat{\lambda}_{LLL}(x)$, we can easily adapt the above procedure to give a plug-in bandwidth selection rule for each of them. The case of $\hat{\lambda}_{LLL,2}(x)$ is, however, more complicated, and, as we shall see, this approach is not practically feasible.

4.3 Practical Issues

4.3.1 Cross-validation

Cross-validation is an intuitively appealing idea, but has certain practical limitations. Consider a single observation (Y_i, X_i) for which X_i is a distance (along the covariate axis) at least R away from any other point. Suppose we use equation (4.4) with a fixed-width kernel and the simplest case of the single-bandwidth nonparametric estimator $\hat{\lambda}_a(x)$. Then, if the bandwidth a is less than R, the leave-one-out estimate of the probability of success at this point cannot be calculated due to a lack of data and the LCV function is undefined.

Silverman [26] pointed this out in the context of density estimation, where in the absence of data the density estimate is zero and hence the log-likelihood undefined. He also noted that although the immediate problem can be solved by the use of 'infinite width' Gaussian kernels, this is a solution which apparently gives undue influence to the tails of the kernel, and intuitively seems to encourage over-smoothing of the data.

In fact, for binary regression the bounds upon the bandwidth are more complex than this. If we expand the logarithm terms in equation (4.3) in terms involving the underlying densities we get

$$\begin{aligned} L(a;X) &= \sum_{j=1}^{n} \left(Y_j \log[\hat{f}_a^{(-j)}(X_j)] + (1-Y_j) \log[\hat{g}_a^{(-j)}(X_j)] - \log[\hat{h}_a^{(-j)}(X_j)] \right) \\ &\quad + m \log(m/s) + n \log(n/s). \end{aligned}$$

Clearly, any point which results in an estimate of either f, g, or h which is zero is going to result in an undefined LCV function. Now, as f is estimated entirely from the subset of successes, and g from the failures, the single bandwidth a must be large enough to prevent the leave-one-out estimate of

either of these densities at any data point being zero. Thus the bandwidth is bounded below by the the *maximum* of the *minimum* distance between successes and the *minimum* distance between failures. Similarly, for the two bandwidth estimators, the bandwidths for a and c are bounded below separately by the minimum distance between successes and the minimum distance between failures.

There are two further drawbacks of the use of cross-validation in this situation. Firstly, as there is no closed-form expression for the bandwidth, we must implement a search algorithm to find the global maximum of the LCV function, which must be capable of coping with the aforementioned boundaries on the search space. Secondly, it would appear that calculation of the contribution of each point X_j to equation (4.4) requires a separate evaluation of the binary regression estimator for the amended dataset, with the obvious potential for time-consuming computational demands.

This second difficulty can be surmounted, however, by the use of simple formulae for the evaluation of leave-one-out estimators. It is well known that the leave-one-out density estimator $\hat{f}_a^{(-j)}(X_j)$ is related to the standard KDE by the fact that

$$\hat{f}_a^{(-j)}(X_j) = \frac{n}{n-1}\hat{f}_a(X_j) - \frac{K(0)}{(n-1)a}. \tag{4.10}$$

Thus individual terms $\hat{f}_a^{(-j)}(X_j)$ can be quickly calculated by evaluating $\hat{f}_a(x)$ on a grid of points, interpolating to get $\hat{f}_a(X_j)$ and then subtracting the constant $\frac{K(0)}{(n-1)a}$. This only involves a single full evaluation of the density estimate and so is obviously much more efficient than the naive approach.

For the single bandwidth nonparametric estimator $\hat{\lambda}_a(x)$, we simply calculate the two density estimates $\hat{f}_a$ and $\hat{h}_a$, then use the above algorithm to calculate $\hat{h}_a^{(-j)}(X_j)$ for all data points and $\hat{f}_a^{(-j)}(X_j)$ for all the successes. For failures, $\hat{f}_a^{(-j)}(X_j) = \hat{f}_a(X_j)$. Thus the calculation of $\hat{\lambda}_a^{(-j)}(x)$ is achiev-

able with approximately the same computational effort as for the standard binary regression estimator. A similar approach is taken for the two bandwidth nonparametric estimator.

For the semiparametric case, things are equally simple. Equations (2.8) and (2.9) and the corresponding expressions for the second partial derivatives are modified so that the contribution when both X_i and x are equal to X_j is zero. Note that this only involves amendments to the expressions for $\frac{\partial Q}{\partial \beta_0}$ and $\frac{\partial^2 Q}{\partial \beta_0^2}$ as the other terms involve the expression $(X_i - x)$. Thus the algorithms for the calculation of both $\hat{\lambda}_{LLL}^{(-j)}(x)$ and $\hat{\lambda}_{LLL,2}^{(-j)}(x)$ are very similar in computational terms to the standard estimates. It should be noted, however, that despite the ease with which the LCV function can be calculated for a particular bandwidth, there is still the need for a search over the bandwidth space to determine the actual bandwidth.

In the original exposition of this approach to binary regression bandwidth selection, Kappenman [13], extends the procedure to the case of bivariate regression with two covariates. As an example, he uses a dataset from the area of fish-curing, where whitefish steaks are treated with sodium chloride and high temperatures before packaging. The response variable is whether or not the steaks are toxic, and the covariates are thus sodium chloride concentration and temperature. Figure 4.1 shows the dataset, with the toxic steaks represented by filled circles, and the non-toxic steaks by open circles. It is immediately obvious that the chloride concentration during the brining process is the major determinant of the toxicity, with the temperature playing a questionable role. Despite this feature of the data, Kappenman uses a Gaussian bivariate product kernel to calculate bivariate estimates of f the density of toxic steaks, and h the overall density. The calculated LCV bandwidths, translated into the more usual standard deviation units of the

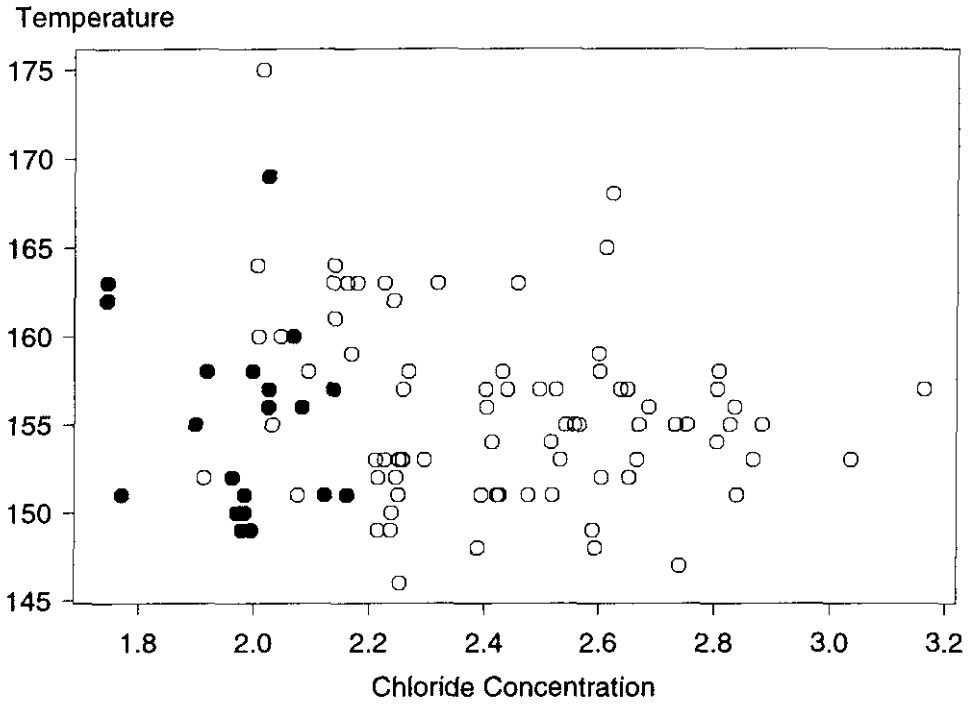

Figure 4.1: Chloride concentration vs. temperature for whitefish steak data. Filled circles represent toxic outcomes.

Gaussian density, are 0.0597 for chloride and 9.71 for temperature. We can further transform these kernels using the theory of canonical kernels (see Wand and Jones [14], p28-31) to give equivalent bandwidths of 0.157 and 25.5 for a bivariate product quartic kernel.

Thus we can see that the estimate for this dataset is driven almost entirely by the chloride variable and is almost independent of temperature. The combination of a relatively large bandwidth and no clear discrimination between toxic and non-toxic steaks on the vertical axis implies that the influence of temperature on the outcome is very limited. Note also that in terms of the chloride concentration, the spread of values for the non-toxic steaks is considerably greater than that for the the toxic one, suggesting that this may be a situation where the two bandwidth estimators may be

usefully applied. With this in mind, we shall return to this example later in this chapter, but shall only consider the relationship between chloride and toxicity, ignoring the temperature variable.

4.3.2 Plug-in Methods

As we noted above, by specifying the unknown density $h(x)$ as the weighting function, we introduce an additional unknown quantity into the estimation of the optimal bandwidth in addition to the probability function λ. Thus, in addition to the parametric pilot estimation of λ, we must also perform some density estimation.

For the simplest estimator $\hat{\lambda}_a(x)$, the WMISE can by written as

$$\begin{aligned} \text{WMISE}(a) &= \frac{a^4(\sigma_K^2)^2}{4} \int \left[\lambda''(x)h(x) + 2\lambda'(x)h'(x)\right]^2 dx \\ &\quad + \frac{R(K)}{sa} \int \lambda(x)[1-\lambda(x)]h(x)\, dx. \end{aligned} \tag{4.11}$$

Now let $\hat{p}(x)$ be the cubic polynomial logistic regression estimator (cubic since in this case $p = 0$), then the second integral term in equation (4.11) is the expectation over X of $\lambda(1-\lambda)$ and so can be estimated by

$$\hat{I}_2 = \frac{1}{s} \sum_{i=1}^{s} \hat{p}(X_i)\ [1 - \hat{p}(X_i)]\,.$$

The first integral is not so simple to estimate, as it involves both $h(x)$ and $h'(x)$. To estimate these, we replace $h(x)$ by a pilot estimate $\hat{h}(x)$ where the density is estimated using a bandwidth selected by the Sheather-Jones method [17]. As $\hat{p}(x)$ is a polynomial model, the first and second derivatives can be easily calculated using the estimated coefficients. Similarly, $h'(x)$ could be estimated directly using an appropriate kernel estimator, but for computational simplicity, we have used the numerical derivative of the estimate of $h(x)$. It should be remembered that these are pilot estimators to

derive a bandwidth for $\hat{\lambda}_a(x)$, and so absolute accuracy is not of paramount importance.

Thus to estimate the first integral, a grid of points is calculated, at each point x_g the integrand is estimated by

$$\left[\hat{p}\,''(x_g)\hat{h}(x_g) + 2\,\hat{p}\,'(x_g)\hat{h}'(x_g)\right]^2, \tag{4.12}$$

and the estimate $\hat{I}_1$ is calculated numerically by summation of these terms. Finally, we have the plug-in bandwidth estimator

$$\hat{a}_{OPT} = \left(\frac{R(K)\,\hat{I}_2}{(\sigma_K^2)^2\,\hat{I}_1}\right)^{1/5} s^{-1/5}, \tag{4.13}$$

which, as in the case of simple kernel density estimation, gives a bandwidth which is directly proportional to $s^{-1/5}$.

Although the asymptotic calculations are tedious, since this method gives a closed form expression for the data-dependent bandwidth, it is substantially faster than cross-validation, which requires a search over plausible values of a. The computational effort involved in estimating the polynomial logistic regression function, the density $h(x)$ and their derivatives is small by comparison to the optimisation algorithms necessary to derive the LCV bandwidths. In addition, the values of the plug-in bandwidth are not bounded below by an arbitrary value as a consequence of the algorithm, a very important practical consideration.

A similar approach can be taken to construct a plug-in bandwidth selector for $\lambda_{LLL}(x)$, which gives a slightly different solution to that given by Fan et.al., as we are using a different WMISE function, working in terms of λ directly rather than a transformation. Note that in this case the pilot parametric estimator $\hat{p}(x)$ is a logistic polynomial model of degree 4 as for this estimator $p = 1$.

The asymptotic variance of $\lambda_{LLL}(x)$ is identical to that of $\lambda_a(x)$, and so the estimated integral $\hat{I}_2$ is the same for both bandwidth estimators. The asymptotic bias is different, however, and from equation (2.3) we can see that the integrand term for the estimate of I_1 should be

$$\left\{ [\text{logit}(\hat{p})]''(x_g)\, \hat{p}(x_g)\, [1 - \hat{p}(x_g)]\, \hat{h}(x_g) \right\}^2. \tag{4.14}$$

Note that this is simpler to calculate than for $\lambda_a(x)$, as there is only a single estimate of the density $h(x)$, and $\text{logit}(\hat{p})$ is a polynomial of degree 4 in x, allowing the second derivative to be calculated directly from the model coefficients. With this modified estimtate of I_1, the expression for the plug-in bandwidth is identical to equation (4.13), and we proceed as before.

In some situations, the fitting of a high-degree logistic polynomial estimator may result in a numerically unstable model and an estimator $\hat{p}(x)$ which is 0 or 1 nearly everywhere. This precludes estimation of the estimated optimal bandwidth, and suggests that the data can be well fitted by a lower degree polynomial. When this occurred, a polynomial of degree $p+2$ was used instead.

For the two bandwidth estimator $\hat{\lambda}_{a,c}(x)$, the expression for the WISE involves five integral terms involving $\lambda(x)$, $h(x)$, $f''(x)$ and $g''(x)$. The first of these can be estimated using a parametric polynomial fit, and the others using standard kernel density estimation procedures and a Sheather-Jones estimate of the bandwidth. This results in an estimate of the WISE which is a polynomial with terms in a^4, a^2c^2, c^2, a^{-1} and c^{-1}. Although this is not strictly a closed-form expression for the estimates of the WISE-optimal bandwidths, standard numerical optimisation routines can be used to calculate these data-dependent estimates of the optimal bandwidths.

For $\lambda_{LLL,2}(x)$, the situation is more difficult. The bias of this estimator

can be shown to be of the form

$$\frac{\sigma_K^2}{2}\frac{1}{h^2[a^2\lambda + c^2(1-\lambda)]} \times \left(\text{terms in } a^4, a^2c^2 \text{ and } c^4\right).$$

When this bias is squared and integrated to give the WISE, the form of the inverse quadratic term in a^2 and c^2 implies that the resulting expression *cannot* be simplified to take the bandwidth terms outside the integral. Thus, rather than having to fit a parametric estimate, calculate numerically a number of integrals, and use the resulting constants as coefficients in a polynomial equation involving a and c, for this estimator we are faced with an estimate of WISE as a function of the bandwidths which must be numerically calculated for every value of a and c.

In addition, the variance of this estimator can be shown to be

$$\frac{\lambda(1-\lambda)}{sh}\int \left[a^{-1}K(u/a)\{1-\lambda(u)\} + c^{-1}K(u/c)\lambda(u)\right]^2 du.$$

Now, although the integral of this part of the WISE *can* be expressed as a polynomial in a and c with coefficients which must only be calculated once, the order of this polynomial depends upon the particular kernel function being used. Note also that this only applies if polynomial kernels are used; for the Gaussian kernel we are faced with an integral involving both exponential and inverse polynomial terms in a and c.

These difficulties, together with the unwieldy expression for the WISE in this two bandwidth case, are in direct opposition to the 'quick and simple' philosophy of plug-in bandwidth rules. Given the problems encountered in the previous chapter, where when using the optimal bandwidths this estimator tended to lower the WISE by altering the estimate of π_1 rather than changing the shape of $\hat{\lambda}$ itself, we have chosen not to pursue this particular plug-in rule for this estimator. We are still able to get some clues, however, as to the practicability or otherwise of this estimator from the LCV

approach results.

4.4 Simulation Experiment

To compare the two competing bandwidth selection approaches, and to assess whether the potential improvements that *may* be achieved by using two bandwidths instead of one can be realised, an extensive simulation experiment was performed. Where possible, for each of the four estimators $\hat{\lambda}_a(x)$, $\hat{\lambda}_{a,c}(x)$, $\hat{\lambda}_{LLL}(x)$ and $\hat{\lambda}_{LLL,2}(x)$, the cross-validated and plug-in bandwidths were calculated for each of the 100 datasets from the 24 test probability functions used in the previous chapters.

The data-dependent bandwidths were then used to calculate the WISE for each dataset, and this was compared with the previously derived best possible optimal values for that estimator, and also across estimators, with the aim of drawing some general conclusions which would be useful in the practical application of these estimators.

4.4.1 Cross-Validation

The previously discussed problem of a minimum bound on the bandwidths allowable for likelihood cross-validation (LCV) implied that calculation of the LCV function for the models involving Cauchy densities (models 13 to 16) was almost impossible. The lower bound on the bandwidth was frequently considerably larger that 10. The only practical alternative in these cases was to adopt a procedure of deleting the 'outliers', and then calculating the LCV bandwidth. This, however, would have implied that we were simulating from a truncated Cauchy distribution and could have led to an unfairly optimistic assessment of this bandwidth selection procedure, as it is clear that it is infeasible (at least with finite width kernels) for distributions

with heavy tails. For this reason, these four models were omitted from the simulation experiment, but only for these particular bandwidth selectors.

For each dataset, the LCV bandwidth estimator for each of the four binary regression estimators $\hat{\lambda}_a(x)$, $\hat{\lambda}_{a,c}(x)$, $\hat{\lambda}_{LLL}(x)$ and $\hat{\lambda}_{LLL,2}(x)$ was calculated. The WISE of the resulting estimator was then compared to both the previously derived optimal value, representing the best possible performance of that estimator for that dataset, and also to the WISE of the other estimators.

For the nonparametric bandwidth estimators Tables 4.1 and 4.2 show the median WISE values, the median percentage increase over the optimal values, and the p-value for the Wilcoxon signed rank test of the difference in WISE values for the nonparametric estimators $\hat{\lambda}_a(x)$ and $\hat{\lambda}_{a,c}(x)$.

The first point to be noted from these results is that, with the exception of Models 10, 11, 19, 20 and 22, the two bandwidth approach of $\hat{\lambda}_{a,c}(x)$ is never significantly better than the far simpler single bandwidth estimator. Indeed, the combination of the single bandwidth estimator and likelihood cross-validation gives WISE values that are on average only 10-45% larger than the ideal optimal values for all of the models except 10, 19 and 20.

The poor performance of the estimator for Models 10, 19 and 20 compared to the optimal values can be explained by the fact that for over 80% of the simulated datasets for these models, the WISE-optimal bandwidth was less than the previously described lower bound on the bandwidth. Indeed, the data-dependent bandwidths for Model 20 are hopelessly inadequate, as can be seen from the fact that the median percentage increase over the optimal WISE is measured in terms of a 600 to 1000% increase!

Thus, the two bandwidth nonparametric estimator only improves upon the single bandwidth case in the cases where the densities f and g are very

Model	λ_a		$\lambda_{a,c}$		Wilcoxon Test
	Median WISE	Increase (%) over optimal	Median WISE	Increase (%) over optimal	p-value
Linear Shift					
1 : $\mu = 0.5$	698	22.20	831	144.20	0.00010
2 : $\mu = 0.75$	799	16.32	874	88.25	0.00024
3 : $\mu = 1$	841	16.94	916	49.09	0.00473
4 : $\mu = 1.25$	591	19.56	629	47.07	0.00070
5 : $\mu = 1.5$	703	33.49	844	45.78	0.16850
Different Proportions					
6 : $\pi_1 = 0.2$	503	26.19	539	76.88	0.08040
7 : $\pi_1 = 0.4$	700	21.21	799	48.39	0.02357
8 : $\pi_1 = 0.6$	769	28.18	904	67.22	0.00002
9 : $\pi_1 = 0.8$	470	21.80	486	56.84	0.76091
Different Variance					
10 : $\sigma = 0.2$	3720	89.44	1084	89.90	0.00000
11 : $\sigma = 0.5$	1361	16.24	994	64.33	0.00119
12 : $\sigma = 0.8$	992	10.02	1017	117.84	0.05268

Table 4.1: Comparison of LCV bandwidth selection WISE values for nonparametric binary regression estimators, Models 1 to 12. WISE values are $\times 10^6$, and p-values are from a Wilcoxon signed rank test.

Model	λ_a		$\lambda_{a,c}$		Wilcoxon Test
	Median WISE	Increase (%) over optimal	Median WISE	Increase (%) over optimal	p-value
Cauchy					
13 : $\mu = 0.6$					
14 : $\mu = 0.8$					
15 : $\mu = 1$					
16 : $\mu = 1.2$					
Marron-Wand					
17 : MW(2)	1174	45.23	1131	71.53	0.68621
18 : MW(3)	488	30.41	642	82.70	0.03642
19 : MW(4)	8583	162.94	7530	207.04	0.00000
20 : MW(5)	22551	641.96	16174	1002.87	0.00000
21 : MW(6)	1563	19.44	1506	54.23	0.74004
22 : MW(7)	1236	17.34	1177	30.48	0.00210
23 : MW(8)	1117	21.66	1189	93.51	0.84597
24 : MW(9)	1398	38.00	1405	60.06	0.86756

Table 4.2: Comparison of LCV bandwidth selection WISE values for nonparametric binary regression estimators, Models 17 to 24. WISE values are $\times 10^6$, and p-values are from a Wilcoxon signed rank test.

different in terms of spread, and in other situations the added complexity actually seems to penalise estimation.

Similar tables of the results for the semiparametric estimators $\hat{\lambda}_{LLL}(x)$ and $\hat{\lambda}_{LLL,2}(x)$ are given in Tables 4.3 and 4.4.

A very similar pattern to that seen with the nonparametric estimators is evident here. The two bandwidth version with LCV bandwidth selection only improves upon the single bandwidth approach for Models 10, 11, 19 and 20, and in other situations offers no significant advantage.

What is also clear is the hugely divergent performance of these two estimators compared to the optimal results when the bandwidth was selected to minimise the WISE. For the first 9 models, which are linear on the logistic scale, the median WISE values for the single bandwidth estimator are never more than about 12% greater than the best possible, although the performance in the non-linear cases is less good. For the two bandwidth estimator, however, the WISE values are on average nearly an order of magnitude greater than the optimal. We have already seen, however, in the previous chapter that this dramatic optimal improvement in WISE was achieved in the main not by better estimation of the true probability function, but by adjusting when the proportion of successes was different from the theoretical value. Obviously cross-validation cannot achieve this, and the more realistic values we see here allow us a fairer assessment of the value of the more complex two bandwidth semiparametric estimator.

When comparisons are made between the non- and semiparametric estimators, it is clear that for likelihood cross-validation the locally linear logistic method is recommended for all models except those for which the density of failures is much more (or less) concentrated than that of the successes. For these cases (Models 10, 11, 19 and 20), it is interesting to note that

Model	λ_{LLL}		$\lambda_{LLL,2}$		Wilcoxon Test
	Median WISE	Increase (%) over optimal	Median WISE	Increase (%) over optimal	p-value
Linear Shift					
1 : $\mu = 0.5$	397	8.67	452	336.86	0.54851
2 : $\mu = 0.75$	383	6.15	379	238.10	0.47770
3 : $\mu = 1$	396	8.95	363	313.20	0.49709
4 : $\mu = 1.25$	293	3.67	296	250.37	0.98491
5 : $\mu = 1.5$	294	11.76	320	340.35	0.53259
Different Proportions					
6 : $\pi_1 = 0.2$	295	5.00	284	505.03	0.71423
7 : $\pi_1 = 0.4$	294	4.71	280	542.39	0.34879
8 : $\pi_1 = 0.6$	334	12.45	307	313.39	0.39097
9 : $\pi_1 = 0.8$	284	7.99	282	508.80	0.37411
Different Variance					
10 : $\sigma = 0.2$	2931	112.79	1406	132.75	0.00000
11 : $\sigma = 0.5$	1544	67.45	1258	115.07	0.00000
12 : $\sigma = 0.8$	843	23.53	766	185.18	0.76091

Table 4.3: Comparison of LCV bandwidth selection WISE values for semiparametric binary regression estimators, Models 1 to 12. WISE values are $\times 10^6$, and p-values are from a Wilcoxon signed rank test.

Model	λ_{LLL}		$\lambda_{LLL,2}$		Wilcoxon Test
	Median WISE	Increase (%) over optimal	Median WISE	Increase (%) over optimal	p-value
Cauchy					
13 : $\mu = 0.6$					
14 : $\mu = 0.8$					
15 : $\mu = 1$					
16 : $\mu = 1.2$					
Marron-Wand					
17 : MW(2)	826	23.63	797	110.22	0.12305
18 : MW(3)	498	57.46	471	123.97	0.41219
19 : MW(4)	13066	284.69	12286	331.00	0.00000
20 : MW(5)	28888	1212.29	19389	1142.85	0.00000
21 : MW(6)	1576	52.73	1645	91.64	0.35950
22 : MW(7)	1352	17.12	1299	63.68	0.15410
23 : MW(8)	1254	27.91	1316	74.95	0.06660
24 : MW(9)	1785	75.43	1694	83.34	0.10646

Table 4.4: Comparison of LCV bandwidth selection WISE values for semiparametric binary regression estimators, Models 17 to 24. WISE values are $\times 10^6$, and p-values are from a Wilcoxon signed rank test.

although a two bandwidth estimator must be used, it is the nonparametric one which seems to perform best.

4.4.2 Plug-in Methods

The plug-in bandwidth selection procedure gives either a closed-form expression for the bandwidth or a simple polynomial equation to be solved numerically, and so there are no constraints upon the bandwidths as there were for likelihood cross-validation. Thus, estimates of performance for the four models where g is the density of a Cauchy distribution (Models 13 to 16) can be calculated.

Tables 4.5 and 4.6 give the median WISE values, the median percentage increase over the optimal values, and the p-value for the Wilcoxon signed rank test of the difference in WISE values for the nonparametric estimators $\hat{\lambda}_a(x)$ and $\hat{\lambda}_{a,c}(x)$.

Again, the two bandwidth estimator only improves upon the single bandwidth case for the models where the spread of the distribution of failures is substantially reduced compared to the successes. Models 10, 11, 19, 20 and 22 all show significant improvements with $\hat{\lambda}_{a,c}$. If the densities are more similar however, once again the use of the two bandwidth estimator can actually make things worse, with a significant increase in the WISE observed for nearly all of these models.

The results for the single bandwidth locally linear logistic estimator $\hat{\lambda}_{LLL}$ are shown in Table 4.7.

In this case, as opposed to the the LCV approach, the differences between the non- and semiparametric estimators are not so marked. For most models $\hat{\lambda}_{LLL}$ appears to be slightly better, or at least not substantially worse than $\hat{\lambda}_a$. For the usual suspects (Models 10, 11, 19 and 20) the two bandwidth

Model	λ_a		$\lambda_{a,c}$		Wilcoxon Test
	Median WISE	Increase (%) over optimal	Median WISE	Increase (%) over optimal	p-value
Linear Shift					
1 : $\mu = 0.5$	575	8.22	755	118.92	0.00000
2 : $\mu = 0.75$	655	8.11	821	60.17	0.00000
3 : $\mu = 1$	756	6.50	906	41.66	0.00000
4 : $\mu = 1.25$	514	10.21	578	28.46	0.00000
5 : $\mu = 1.5$	569	7.88	646	25.12	0.00000
Different Proportions					
6 : $\pi_1 = 0.2$	436	9.79	481	56.37	0.00065
7 : $\pi_1 = 0.4$	564	5.65	723	35.42	0.00000
8 : $\pi_1 = 0.6$	654	6.98	739	35.98	0.00000
9 : $\pi_1 = 0.8$	391	5.24	433	40.11	0.00111
Different Variance					
10 : $\sigma = 0.2$	1696	6.08	898	53.30	0.00000
11 : $\sigma = 0.5$	1079	7.68	916	53.69	0.00088
12 : $\sigma = 0.8$	873	7.69	1054	140.24	0.00000

Table 4.5: Comparison of plug-in bandwidth selection WISE values for nonparametric binary regression estimators, Models 1 to 12. WISE values are $\times 10^6$, and p-values are from a Wilcoxon signed rank test.

Model	λ_a		$\lambda_{a,c}$		Wilcoxon Test
	Median WISE	Increase (%) over optimal	Median WISE	Increase (%) over optimal	p-value
Cauchy					
13 : $\mu = 0.6$	570	7.40	706	64.16	0.00000
14 : $\mu = 0.8$	661	8.50	938	74.08	0.00000
15 : $\mu = 1$	669	9.37	810	52.22	0.00000
16 : $\mu = 1.2$	601	7.62	825	44.63	0.00000
Marron-Wand					
17 : MW(2)	806	10.83	1015	38.52	0.00407
18 : MW(3)	485	12.10	794	185.35	0.00000
19 : MW(4)	8469	160.55	2861	19.07	0.00000
20 : MW(5)	6216	83.66	1772	18.62	0.00000
21 : MW(6)	1295	7.47	1323	39.03	0.11893
22 : MW(7)	1030	7.40	928	17.86	0.00001
23 : MW(8)	997	9.56	1061	49.80	0.02693
24 : MW(9)	1126	14.78	1227	23.79	0.46500

Table 4.6: Comparison of plug-in bandwidth selection WISE values for nonparametric binary regression estimators, Models 17 to 24. WISE values are $\times 10^6$, and p-values are from a Wilcoxon signed rank test.

Model	Median WISE	Increase (%) over optimal	Model	Median WISE	Increase (%) over optimal
Linear Shift			**Cauchy**		
1 : $\mu = 0.5$	639	82.91	**13** : $\mu = 0.6$	527	14.65
2 : $\mu = 0.75$	585	72.90	**14** : $\mu = 0.8$	667	19.92
3 : $\mu = 1$	556	53.86	**15** : $\mu = 1$	661	9.07
4 : $\mu = 1.25$	438	46.45	**16** : $\mu = 1.2$	561	20.47
5 : $\mu = 1.5$	392	54.38	**Marron-Wand**		
Different Proportions			**17** : MW(2)	772	9.54
6 : $\pi_1 = 0.2$	470	67.22	**18** : MW(3)	523	17.39
7 : $\pi_1 = 0.4$	483	71.61	**19** : MW(4)	7208	123.11
8 : $\pi_1 = 0.6$	518	60.67	**20** : MW(5)	4706	95.90
9 : $\pi_1 = 0.8$	390	37.83	**21** : MW(6)	1183	10.43
Different Variance			**22** : MW(7)	1203	5.06
10 : $\sigma = 0.2$	1498	11.22	**23** : MW(8)	1067	9.13
11 : $\sigma = 0.5$	1162	13.31	**24** : MW(9)	1197	8.75
12 : $\sigma = 0.8$	1005	20.05			

Table 4.7: Evaluation of plug-in bandwidth selection WISE values for single bandwidth semiparametric binary regression estimators, Models 1 to 24. WISE values are $\times 10^6$.

nonparametric estimator $\hat{\lambda}_{a,c}$ is clearly the best option.

4.4.3 LCV Method versus Plug-in Method

For the nonparametric estimators, in nearly every single case the plug-in approach gives substantially lower WISE values than cross-validation. This behaviour has been observed regularly in many other areas of kernel smoothing (see, for example Park and Marron [29]), and it is hardly surprising that it can be replicated in the area of binary regression.

Two reasons for the poor performance of cross-validation are the constraint of the lower bound upon the bandwidths due simply to the derivation of the algorithm, and the fact that cross-validatory procedures generally exhibit large variation. This second failing can be seen in Figure 4.2, where the data-dependent bandwidths from Model 5 selected by LCV and the plug-in methods for the simplest estimator $\hat{\lambda}_a$ are compared with the optimal values. Clearly, in this case the LCV procedure produces bandwidths with a much greater variability than the plug-in procedure, even in this model in which both successes and failures have the same standard deviation with a large distance of 1.5 standard deviations between the means of the two populations. In addition, the tendency of cross-validation methods to oversmooth the data can also be seen, with the plug-in bandwidths being much less biased.

Interestingly, this failure does not apply to the semiparametric case, at least for Models 1 to 9. Here, when using $\hat{\lambda}_{LLL}$, the LCV bandwidth procedure seems to give much better estimation than the plug-in method. Examination of the bandwidths for these models, however, gives an explanation. These models are all linear on the logistic scale, so the bandwidth a for $\hat{\lambda}_{LLL}$ should tend towards very large values since as $a \to \infty$, the estimator

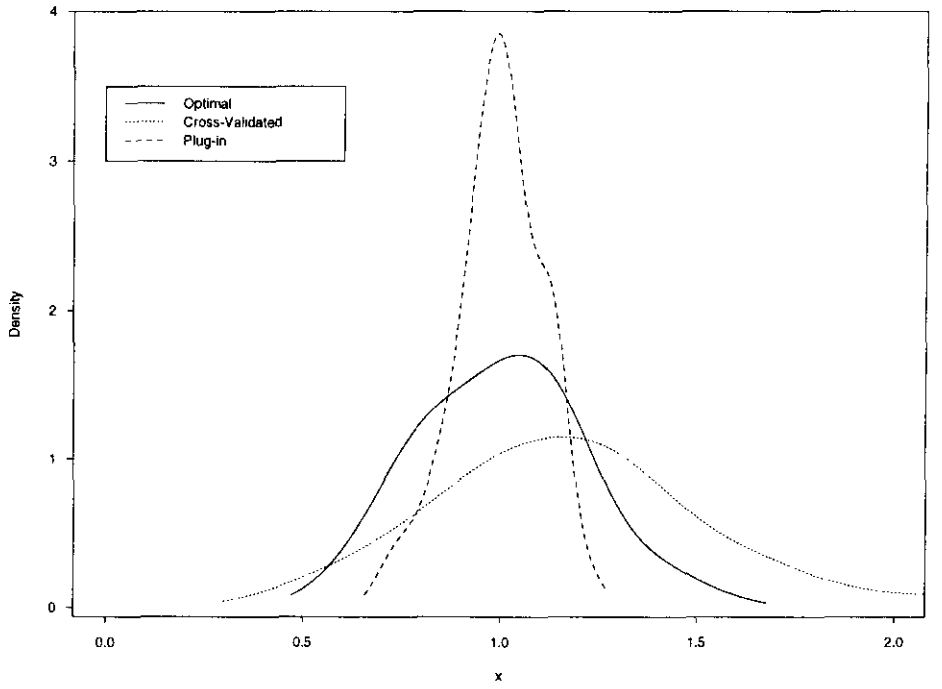

Figure 4.2: Density estimates of data-dependent bandwidths for Model 5, showing relative variance of the two approaches.

becomes a parametric logistic regression. The median optimal bandwidths for these models range from 5.4 up to 10 (which is a purely arbitrary limit which was imposed upon the experiment). For the LCV method, the data-dependent bandwidths also shows this behaviour, with the median values varying over the models from 4.6 up to 10. For the plug-in approach, however, the median bandwidths are only in the range 1.4 to 1.6, suggesting that they are considerably under-smoothing the data. This finding can be intuitively rationalised by noting that these models are a particular case where 'over-smoothing' of the estimates actually improves the situation.

It should also be noted, however, that these nine models are the cases where estimation is very easy for any of the estimators, and that the improvement offered by using LCV over the plug-in method is really only between

'very good' and 'excellent' estimation.

4.5 Conclusions

In this chapter we have finally assessed the practical performance of the various binary regression estimators previously derived and described. Two contrasting approaches to bandwidth selection have been compared, and the very real practical difficulties involved in implementing what are often very 'broad-brush' descriptions in the literature have been demonstrated.

The general recommendations from the simulation experiment would seem to be that the locally linear logistic single bandwidth estimator is superior, except for the cases in which the variances of the two densities are very different, in which case the two bandwidth nonparametric estimator is preferable. In all cases except when the true model is linear on the logistic scale, the plug-in bandwidth selection rule is considerably closer to the optimal than likelihood cross-validation, with the added advantage of computational simplicity.

This last point emphasises that the use of the term 'semiparametric' to describe $\hat{\lambda}$ is different from the normal usage of this term in density estimation. In the latter, semiparametric estimators tend to have performance very similar to fully parametric estimators *when the parametric model is correct*, but are robust to departures from this model. In our situation, even when the model is of the correct parametric form, in practice the estimator does not achieve large enough data-dependent bandwidths to approximate logistic regression.

The use of two bandwidths rather than one can be seen as a very limited form of adaptive smoothing, where the bandwidth is allowed to vary continuously with location. It is clear that two bandwidth estimators are

necessary in certain situations, where they can dramatically improve the accuracy of the estimation. It does seem, however, that these cases are those where there is a *substantial* difference in the variance of the two groups, as the use of the two bandwidth estimators actually made things worse in situations like Model 12, where the variances were 1 and 0.64 for the successes and failures respectively.

The four simulated models where the two bandwidth approach was an improvement are exactly those where the ratio of variances between successes and failures is 2 or more, so a simple practical rule-of-thumb may be to use $\hat{\lambda}_{LLL}$, when the ratio of sample variances is less than 2, to use $\hat{\lambda}_{a,c}$ otherwise, and in each case to use the appropriate plug-in bandwidth selector.

In practice, this procedure will have problems when confronted by heavy tailed distribution like those of Models 13 to 16, as the sample variance is not robust to outliers. For this reason, it may be appropriate to replace the sample variance by the more robust 'super scale' estimator $\hat{\sigma}^2_{SS}$ from Chapter 6. When applied to our battery of test models, and bearing in mind the final paragraph of Section 4.4.3, this approach should give adequate performance for the simpler models and at the same time avoid the very poor estimation of the single bandwidth approach to the differing variance case.

For each of the simulated datasets this ratio of variance estimates was calculated. For most models less than 10% of datasets resulted in a variance ratio of 2 or more. For Models 10, 19 and 20, however, the ratio was always (with a single exception) greater than the threshold and so $\hat{\lambda}_{a,c}(x)$ was used to calculate the estimate. Models 11 and 18 resulted in the use of the two bandwidth estimator in 57% and 75% of cases respectively, giving intermediate performance. Table 4.8 gives the median WISE values which result from applying this procedure to the test data, and the dramatic improvements

Model	Median WISE	Model	Median WISE
Linear Shift		**Cauchy**	
1 : $\mu = 0.5$	642	**13** : $\mu = 0.6$	543
2 : $\mu = 0.75$	585	**14** : $\mu = 0.8$	731
3 : $\mu = 1$	625	**15** : $\mu = 1$	667
4 : $\mu = 1.25$	430	**16** : $\mu = 1.2$	561
5 : $\mu = 1.5$	392	**Marron-Wand**	
Different Proportions		**17** : MW(2)	789
6 : $\pi_1 = 0.2$	470	**18** : MW(3)	782
7 : $\pi_1 = 0.4$	488	**19** : MW(4)	2861
8 : $\pi_1 = 0.6$	525	**20** : MW(5)	1772
9 : $\pi_1 = 0.8$	387	**21** : MW(6)	1183
Different Variance		**22** : MW(7)	1203
10 : $\sigma = 0.2$	898	**23** : MW(8)	1067
11 : $\sigma = 0.5$	1015	**24** : MW(9)	1211
12 : $\sigma = 0.8$	1018		

Table 4.8: WISE values resulting from empirical variance ratio rule. WISE values are $\times 10^6$.

over using $\hat{\lambda}_{LLL}(x)$ for Models 10, 19 and 20 are obvious.

Even with this practical rule, however, we are in the slightly unsatisfactory position of recommending different estimators and different bandwidth selectors in different situations, as there are situations where the use of cross-validation seems superior. It seems clear that more research is needed into the question of automatic bandwidth selection for estimators of these types to make them of practical use.

In a very recent paper Fan, Farmen and Gijbels [30] reconsider the

asymptotic behaviour of local polynomial maximum likelihood estimation, with a specific example of logistic regression. By considering higher order polynomials (essentially of order $p + 2$), they derive novel expressions for the bias and variance of the estimator, which they use to construct a data-dependent bandwidth selector using broadly similar ideas to the original plug-in approach. As they are using the likelihood, however, their estimator must avoid fitted values which are either exactly 0 or 1, and they do allude to this problem of a lower bound on the automatic bandwidth. In addition, using similar computational equipment to that used here, their estimate of the bandwidth "takes approximately 20 minutes to compute". It remains to be seen, therefore if this single bandwidth estimator and its associated bandwidth selection procedure is really an improvement upon what has gone before.

Finally, if we return to the fish-curing data of Kappenman which was discussed earlier, we can try to apply the practical advice to the relationship between chloride concentration and toxicity. As a first step, the variance of the chloride concentrations in the toxic steaks was estimated (using $\hat{\sigma}^2_{SS}$) to be 0.0062, and in the non-toxic steaks it was 0.0728, giving a ratio of approximately 12. Thus, according to our rule of thumb, we should consider using $\hat{\lambda}_{a,c}(x)$ in preference to $\hat{\lambda}_{LLL}(x)$.

If we then attempt to use the plug-in estimator for $\hat{\lambda}_{a,c}(x)$, we find that the cubic polynomial logistic regression has numerical problems and we must instead fit a quadratic model. This gives estimated bandwidths of $a = 0.066$ and $c = 0.162$, suggesting that we were wise to use two bandwidths. Figure 4.3 shows the resulting binary regression estimate, with the two bandwidth estimate of $\lambda(x)$ indicated by the dashed line. Does the dip in the probability of toxicity at a chloride concentration of 1.9 make sense? It is difficult to

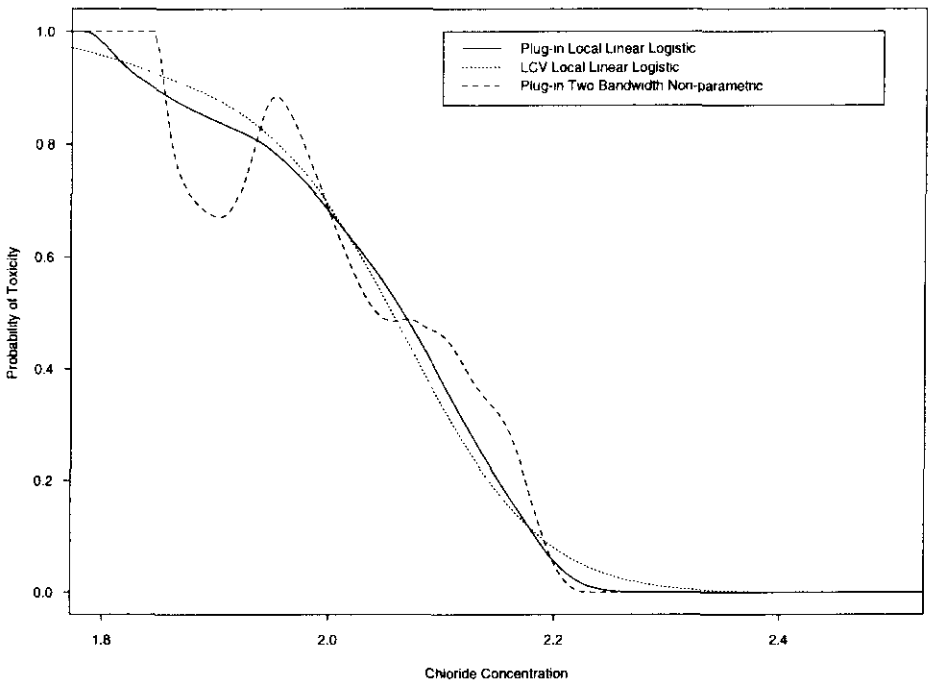

Figure 4.3: Binary regression estimates of the probability of toxicity for fish-curing data for three different combinations of estimator and bandwidth selector.

come up with a plausible biological explanation immediately, and referring back to Figure 4.1, we can see that this may be due to the single non-toxic steak with the smallest chloride concentration.

If we instead calculate the plug-in bandwidth for the single bandwidth estimate $\hat{\lambda}_{LLL}(x)$, we again cannot fit either a quartic or cubic polynomial, but must rely on the quadratic fit. This results in a bandwidth of $a = 0.19$, and the resulting estimate of λ is shown by the solid line in Figure 4.3. This is a much more scientifically reasonable estimate. For comparison, the LCV bandwidth for this estimator was calculated to be 0.40, and this estimate is plotted as the dotted line in Figure 4.3. It would seem that of the two, the LCV estimator is slightly oversmoothed in relation to the plug-in estimator,

which avoids the anomalous dip at 1.9, but is still predicting definite toxicity below 1.8 and definite non-toxicity above 2.25.

Given the apparent superiority of the single bandwidth estimator in this example, even with a variance ratio of 5, should we amend our rule-of-thumb? Probably not, as in this case the problem seems to be more related to a single outlying and highly-influential point than to a general failing of the model. It should also be noted that there were only 20 'successes' (toxic steaks) in this dataset, with the obvious implications this has for the variability of the ratio of the two variances.

Thus it would seem that the general principles of our approach are justified, and that both non- and semiparametric binary regression estimators can usefully describe pertinent features of real data. For the particular estimators recommended, $\hat{\lambda}_{LLL}$ and $\hat{\lambda}_{a,c}$, the plug-in bandwidth selection rules are relatively simple and quick to calculate, giving a very practical solution to problems of this kind.

Part II

Density Estimation

Chapter 5

Improved Kernel Density Estimation

5.1 Introduction

Kernel density estimation has become a widely-used and well accepted statistical technique. The properties of the standard kernel density estimate are well investigated and understood. Many authors have, however, tried to improve upon the simple and intuitive standard estimate. These improvements are often driven by consideration of the asymptotic behaviour as the number of data points available becomes very large. These 'higher-order' kernel density estimators (KDE) are nearly always compared only with the standard estimator and rarely with each other. The aim of this chapter is to investigate the small-sample behaviour of a wide variety of these 'improved' KDEs, over a range of distributional shapes, by means of an extensive simulation experiment.

A condensed report of the results of this experiment formed a major part of Jones and Signorini [31], which presented the asymptotic behaviour of the

higher-order estimates in a unified framework and compared most of these estimators both theoretically and in a variety of practical circumstances.

5.2 Background

We begin by describing the standard kernel density estimate and its asymptotic behaviour. Given a sample $X_1 \dots X_n$ from an unknown density $f(x)$, the standard KDE is defined by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K_2 \left(\frac{x - X_i}{h} \right), \tag{5.1}$$

where K_2 is a second-order kernel (a symmetric probability density function centred at zero) and h is the bandwidth, determining the amount of smoothing performed.

The simple and intuitive nature of the estimator defined in equation (5.1) can be seen by noting that it can be considered in two distinct ways. Firstly, we can imagine a single kernel function centred at the point x, with the contribution of each point X_i being determined by the height of the kernel function at X_i. For each different value of x at which an estimate of the density is required, the same process can be carried out. Alternatively (and this is the more obvious given the mathematical formulation), we can consider a set of n kernel functions, each centred at a data point X_i. The estimate at a point x is then the (appropriately scaled) sum of each of the kernel functions at that point.

The large-sample asymptotic behaviour of $\hat{f}_h(x)$ is well known (see, for example, Silverman [26] or Wand and Jones [14]). The asymptotic bias and variance as $n \to \infty, h \to 0$ and $nh \to \infty$ can be expressed simply as

$$\mathrm{E}\left\{ \hat{f}_h(x) - f(x) \right\} = \frac{h^2 \sigma_K^2}{2} f''(x) + o(h^2) \tag{5.2}$$

and

$$\text{var}\left\{\hat{f}_h(x)\right\} = \frac{R(K)}{nh} f(x) + o((nh)^{-1}), \tag{5.3}$$

where σ_K^2 is the variance of the particular kernel function used and $R(K) = \int K(u)^2 du$.

We can thus immediately see the trade-off between bias and variance as the bandwidth changes. A smaller value of h, and hence less smoothing, will lead to a less biased estimate but with larger variance. Conversely, a large degree of smoothing, with correspondingly large values of h, will give estimates with smaller asymptotic variance but a large degree of bias. Typically, the asymptotic mean integrated squared error (MISE) is used as a criterion for estimation which combines these two conflicting measures of estimator accuracy.

The majority of suggested improvements to the standard KDE $\hat{f}_h(x)$, have involved attempts to reduce the MISE (or simply the mean squared error) through reducing the order of the asymptotic bias from $O(h^2)$ to $O(h^4)$ or less. There are a great many ways of achieving this, so for reasons of practicality and conciseness we have chosen one or two of the more promising estimators within each general class of bias-reducing ideas. In addition, we also consider a single example of the so-called 'semiparametric' density estimation methods, which combine a fully parametric estimate with a non-parametric adjustment, with the aim of being efficient when the parametric models is correct, and robust when it is not.

5.3 Higher-Order Kernel Density Estimators

As stated above, all but one of the estimators considered fall into the class of fourth-order methods, eliminating the first term of equation (5.2) in var-

ious ways and reducing the asymptotic bias from $O(h^2)$ to $O(h^4)$. The asymptotic variance remains of order $O((nh)^{-1})$. There are a surprisingly large number of ways of achieving this reduction, and indeed some of the estimators considered are simply special cases of a general bias-reduction method. Similarly, many of the ideas for bias reduction can easily be extended to reduce the bias further, to order $O(h^6)$ and beyond. For reasons of practicality, and the belief that it is the step from order $O(h^2)$ to $O(h^4)$ which can give the largest gain over the standard KDE, we have limited the comparisons only to estimates with fourth-order asymptotic behaviour.

In this section we describe the various estimators which are examined and in the next we briefly discuss their asymptotic behaviour.

5.3.1 Fourth-order Kernel Estimators

The simplest method of eliminating the h^2 term from equation (5.2) is to use a kernel which has $\sigma_K^2 = 0$. Let $K_4(z)$ denote such a function, and define this estimator by

$$\hat{f}_4(x) = \frac{1}{nh} \sum_{i=1}^{n} K_4 \left(\frac{x - X_i}{h} \right). \tag{5.4}$$

Note, however, that the use of a symmetric function for which $\int u^2 K_4(u) du = 0$ precludes the use of probability density functions, and indeed implies that the function must actually be negative for some of its range. For regions of low probability, this can imply that the estimate of the density $\hat{f}_4(x)$ can actually be less than zero. Hall and Murison [25] discuss this problem and explore several possible adjustments to achieve a non-negative estimate, but in the context of asymptotic behaviour rather than practical application.

There are a seemingly endless number of ways in which to construct fourth-order kernel functions. Jones and Foster [32] describe a general approach, which is to use two KDEs with different kernel functions $K(u)$ and

ious ways and reducing the asymptotic bias from $O(h^2)$ to $O(h^4)$. The asymptotic variance remains of order $O((nh)^{-1})$. There are a surprisingly large number of ways of achieving this reduction, and indeed some of the estimators considered are simply special cases of a general bias-reduction method. Similarly, many of the ideas for bias reduction can easily be extended to reduce the bias further, to order $O(h^6)$ and beyond. For reasons of practicality, and the belief that it is the step from order $O(h^2)$ to $O(h^4)$ which can give the largest gain over the standard KDE, we have limited the comparisons only to estimates with fourth-order asymptotic behaviour.

In this section we describe the various estimators which are examined and in the next we briefly discuss their asymptotic behaviour.

5.3.1 Fourth-order Kernel Estimators

The simplest method of eliminating the h^2 term from equation (5.2) is to use a kernel which has $\sigma_K^2 = 0$. Let $K_4(z)$ denote such a function, and define this estimator by

$$\hat{f}_4(x) = \frac{1}{nh} \sum_{i=1}^{n} K_4 \left(\frac{x - X_i}{h} \right). \tag{5.4}$$

Note, however, that the use of a symmetric function for which $\int u^2 K_4(u) du = 0$ precludes the use of probability density functions, and indeed implies that the function must actually be negative for some of its range. For regions of low probability, this can imply that the estimate of the density $\hat{f}_4(x)$ can actually be less than zero. Hall and Murison [25] discuss this problem and explore several possible adjustments to achieve a non-negative estimate, but in the context of asymptotic behaviour rather than practical application.

There are a seemingly endless number of ways in which to construct fourth-order kernel functions. Jones and Foster [32] describe a general approach, which is to use two KDEs with different kernel functions $K(u)$ and

$L(u)$ to give a pair of simultaneous equations of the form of equation (5.2), but with differing kernel-dependent constants multiplying the $h^2 f''(x)$ term. These equations can then be solved to give a new kernel function which is a linear combination of the two kernel functions with asymptotic bias of order h^4. Two specific examples of this technique are studied, in each case beginning with a basic kernel function $K(u)$. Firstly, let the second kernel function be simply $L(u) = u^2 K(u)$, which results in the polynomial fourth order kernel,

$$K_{4P}(u) = \frac{s_4 - s_2 u^2}{s_4 - s_2^2} K(u), \tag{5.5}$$

where $s_k = \int v^k K(v) dv$, the kth moment of the kernel. Secondly, let the second kernel function be the convolution of the original kernel with itself, to give

$$K_{4C}(u) = 2K(u) - (K * K)(u), \tag{5.6}$$

where $*$ denotes convolution, such that $(K * K)(u) = \int K(u - v) K(v) dv$.

Each of these two cases produces a kernel function such that $\int u^2 K(u) du = 0$, but at the expense of having values of u for which $K(u) < 0$. Figure 5.1 shows both of these fourth-order kernels, and the corresponding second-order kernel for the case when the original kernel is the quartic function.

The portions of the curves for which $K_{4P}(u)$ and $K_{4C}(u)$ are less than zero can be clearly seen. In addition, it should be noted that although both fourth-order kernels have support $[-1, 1]$, the convolution kernel is narrower, suggesting that larger bandwidths will be required for equivalent amounts of smoothing. If the convolution kernel is 'stretched' however so that the peak is approximately the same width as for the polynomial kernel, we can see that the two functions would be very similar indeed. We shall see in the simulation experiment that the practical differences between these different types of fourth-order kernel are very small, and can be compared to the

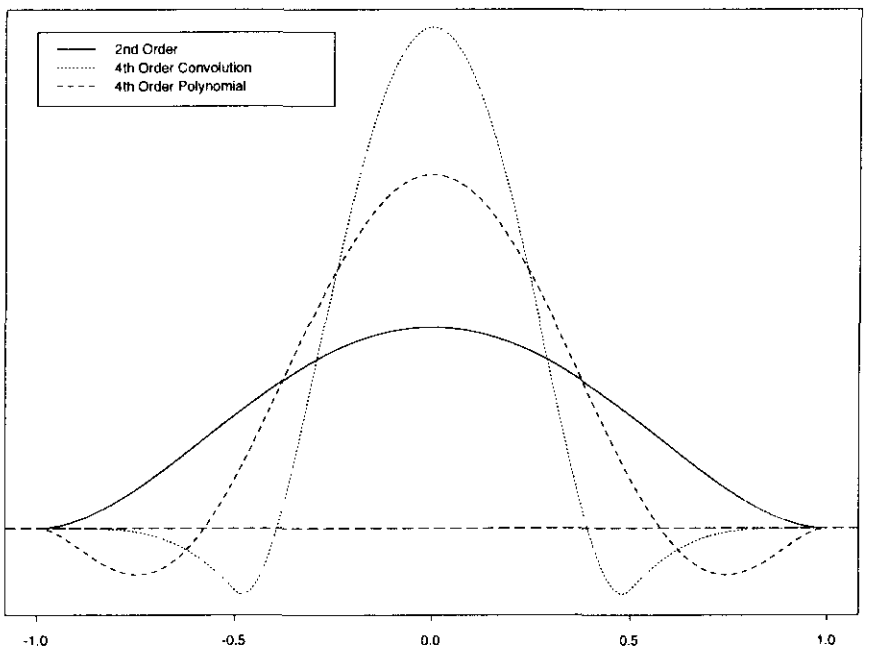

Figure 5.1: Fourth-Order Kernel Functions

differences between different kernel functions for the standard KDE. We thus define the two fourth-order estimators $\hat{f}_{4P}(x)$ and $\hat{f}_{4C}(x)$ of the form given in equation (5.4), with K_4 taken to be K_{4P} and K_{4C} respectively.

5.3.2 Multiplicative Bias-Correcting Estimators

The fourth-order kernel method may be considered as an additive bias-correction method, as the procedure essentially works by adding together two independent estimates of f with equivalent h^2 bias terms. An alternative approach is to take a standard KDE of f and an appropriate estimate of 1, and multiply them together to create a new estimate of f. The estimate of 1 is chosen to have an h^2 bias term which again cancels out that of $\hat{f}_h(x)$, leaving only bias terms of order h^4.

Probably the simplest way of doing this, described in Jones and Foster

[32], is to take the estimator

$$\hat{f}_{JF}(x) = \frac{\hat{f}_h(x)^2}{\tilde{f}_h(x)} = \hat{f}_h(x)\left[\frac{\hat{f}_h(x)}{\tilde{f}_h(x)}\right], \tag{5.7}$$

where

$$\tilde{f}_h(x) = (nh)^{-1}\sum_{i=1}^{n}(K * K)\left[\frac{X_i - x}{h}\right],$$

an estimator of f using the convolution of the kernel function with itself. It is clear in the final part of equation (5.7) that this estimator has two multiplicative components, one estimating f, the other estimating the constant value 1.

Another way of achieving this multiplicative bias correction was proposed by Jones, Linton and Neilsen [33]. This is rather similar to the previous estimator, but now the estimate of f in the denominator is moved 'inside' the second kernel density estimate in the numerator. Formally, we have

$$\hat{f}_{JLN}(x) = \hat{f}_h(x)\left[\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\hat{f}_h(X_i)}K_2\left(\frac{x - X_i}{h}\right)\right]. \tag{5.8}$$

Once again, the second term is an estimate of 1 and is very much correlated with the first term. The asymptotic bias terms in h^2 cancel out and we are left with bias of order h^4. Unfortunately this estimator is not a true density, since it does not integrate to one. Jones et al. suggest, however, that a numerical renormalisation of the final estimate can produce significant benefits in estimation, so we also consider the empirically renormalised version of this estimator, denoted by $\hat{f}^R_{JLN}(x)$.

Examination of equation (5.8) shows that it is possible to modify this estimator to have two distinct bandwidths, h and b, one for the initial pilot estimate, and one for the final estimate. Thus

$$\hat{f}_{JLN,2}(x) = \hat{f}_b(x)\left[\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\hat{f}_b(X_i)}K_2\left(\frac{x - X_i}{h}\right)\right]. \tag{5.9}$$

Provided that b is proportional to h, and that they do not differ in their *rate* of convergence to zero, the h^2 terms in the asymptotic expansion still cancel out and the higher-order behaviour is retained. This estimator can also be empirically rescaled to have unit integral, and given the dominance of the rescaled version in the single bandwidth case, we work exclusively with the renormalised two bandwidth estimator $\hat{f}^R_{JLN,2}$.

5.3.3 Transformation Estimators

A standard statistical technique when faced with data which are in some way 'difficult' to handle is to transform them. For kernel density estimation, this idea was introduced by Ruppert and Cline [34], who used an estimated form of the probability integral transform to transform the data to a uniform distribution on $[0, 1]$. This distribution is very easy to estimate well using kernel density estimation, as asymptotically f' and all higher derivatives are zero. The resulting estimate is then back-transformed to the original scale, to give another estimator with order h^4 asymptotic bias.

Formally, recall that the probability integral transform theory states that if X has cumulative distribution function (CDF) $F(x)$, then the transformed variable $Y = F(X)$ has a uniform distribution on [0,1]. Obviously when faced with an unknown distribution, F is also unknown and so must be estimated. Thus the Ruppert-Cline estimator proceeds in three stages; estimating the smoothed CDF of the data and using this to transform to an estimated uniform distribution on $[0, 1]$, calculating a smoothed KDE in this transformed space, and finally back-transforming the estimator to the original scale. Note that this immediately implies two bandwidths may be appropriate, as the CDF is estimated on the scale of the original data, whereas the density estimate in the transformed space operates in the range

from 0 to 1, and these two scales are likely to be quite different. Thus the full two-bandwidth estimator (with bandwidths h and b) can be written as

$$\hat{f}_{RC,2}(x) = \hat{f}_h(x) \left[\frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{\hat{F}_h(X_i) - \hat{F}_h(x)}{b} \right) \right], \qquad (5.10)$$

where the estimate of the CDF is given by

$$\hat{F}_h(y) = \frac{1}{n} \sum_{i=1}^{n} L\left(\frac{X_i - y}{h} \right),$$

and

$$L(z) = \int_{-\infty}^{z} K(u)du,$$

the integral of the kernel function.

The interpretation of this estimator also as a multiplicative bias correction can be seen from the decomposition of equation (5.10) into a term estimating f and one estimating the density of a uniform distribution, which is, of course, 1.

Practically, a boundary-corrected version of the density estimator in the transformed space is used, as the proportion of the range $[0, 1]$ which falls within a distance b of the boundaries is relatively large. We follow Ruppert and Cline in using the reflection method to cope with this feature. These authors also suggest that, rather than using two bandwidths, a single bandwidth scaling factor can be used, with h and b being proportional to the inter-quartile ranges (IQR) of the data and the transformed data respectively. We denote by $\hat{f}_{RC}(x)$ this single bandwidth version of $\hat{f}_{RC,2}(x)$ with

$$b = \frac{\mathrm{IQR}(X)}{\mathrm{IQR}(\hat{F}_h(X))} h.$$

Alternatively, we can also use a locally varying bandwidth in the transformed space, with $b(x) = h\hat{f}_h(x)$, to give the simplified single-bandwidth

transformation estimator

$$\hat{f}_{RC-V}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\hat{F}_h(X_i) - \hat{F}_h(x)}{h\hat{f}_h(x)}\right). \tag{5.11}$$

This estimator, however, does not integrate to 1 and so we also consider the practical performance of the numerically normalised version (using the same method as in the previous section), denoted by $\hat{f}^R_{RC-V}(x)$.

5.3.4 Variable Bandwidth Estimators

The concept of using a large amount of smoothing in the tails of a density and a small amount around the peaks is intuitively appealing, leading to modifications of the standard KDE with some form of adaptive bandwidth. The concept of thus allowing the bandwidth h in $\hat{f}_h(x)$ to vary, such that

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} h(X_i)^{-1} K_2\left(\frac{x - X_i}{h(X_i)}\right),$$

was first introduced by Victor [35] and Breiman et al. [36]. It was Abramson [37], however, who showed that by taking $h(X_i)$ proportional to $f^{-1/2}(X_i)$, the asymptotic bias is of order h^4. This must be estimated, and so we have another two-stage estimator, using an initial pilot density estimate with bandwidth b to estimate the locally modified bandwidth $h(X_i) = h\,[\hat{f}_b(X_i)]^{-1/2}$. Thus we have the full two-bandwidth version of the variable KDE

$$\hat{f}_{V,2}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h\hat{f}_b^{-1/2}(X_i)} K\left(\frac{x - X_i}{h\hat{f}_b^{-1/2}(X_i)}\right). \tag{5.12}$$

Silverman [26] simplified this estimator by taking the second bandwidth h to simply be proportional to the pilot bandwidth b, scaled by the geometric mean g so that

$$h(X_i) = b\,g\,\hat{f}_b^{-1/2}(X_i) = b\left[\prod_{j=1}^{n} \hat{f}_b^{1/2}(X_j)\right]^{1/n} \hat{f}_b^{-1/2}(X_i),$$

and hence we have the single bandwidth version of the variable KDE

$$\hat{f}_V(x) = \frac{1}{n\,b\,g}\sum_{i=1}^{n}\frac{1}{\hat{f}_b^{-1/2}(X_i)}K\left(\frac{x-X_i}{b\,g\,\hat{f}_b^{-1/2}(X_i)}\right), \tag{5.13}$$

which happily does integrate to 1.

As a final variation, we can use the local rescaling ideas of the estimator $\hat{f}_{RC-V}(x)$, so that

$$h(X_i, x) = \hat{f}_b^{-1/2}(x)\hat{f}_b^{-1/2}(X_i)\;b,$$

giving the locally rescaled variable KDE with a single bandwidth

$$\hat{f}_{V-V}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{b\hat{f}_b^{-1/2}(x)\hat{f}_b^{-1/2}(X_i)}\;K\left(\frac{x-X_i}{b\hat{f}_b^{-1/2}(x)\hat{f}_b^{-1/2}(X_i)}\right). \tag{5.14}$$

5.3.5 Variable Location Estimator

As an alternative to locally adjusting the width of the kernel functions, Samiuddin and el-Sayyad [38] demonstrate that shifting the location of each kernel by an amount proportional to f'/f is another way of achieving order h^4 bias. Once again, the size of the location shift must be estimated, requiring a pilot step, and giving the variable location estimator

$$\hat{f}_{VL}(x) = \frac{1}{nh}\sum_{i=1}^{n}K_2\left(\frac{x-X_i}{h} + \frac{h\sigma_K^2}{2}\,\frac{\hat{f}_h'(X_i)}{\hat{f}_h(X_i)}\right). \tag{5.15}$$

This estimator is effectively moving the data points in the direction of positive slope, which in practice means towards the estimated peaks of the density. Although by separating the pilot and final estimation step this estimator can also be considered in a two bandwidth way, we have restricted ourselves to the simplest single bandwidth case.

5.3.6 Semiparametric Estimator

Finally, we consider a single estimator from a completely different approach to the problem of density estimation. Semiparametric density estimation seeks to construct an estimator which combines the strengths of a parametric fit to the data when the density is of the correct form with nonparametric robustness to departures from this model. Hjort and Glad [39] suggest the semiparametric estimator

$$\hat{f}_{SP}(x) = \hat{f}(x;\hat{\theta})\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\hat{f}(X_i;\hat{\theta})}K\left(\frac{x-X_i}{h}\right), \tag{5.16}$$

where $\hat{f}(x;\hat{\theta})$ is an estimated fit of the parametric family of distributions $f(x;\theta)$ to the data. Note the similarity here with the multiplicative bias-correcting estimator $\hat{f}_{JLN}(x)$, where the pilot nonparametric estimator $\hat{f}_h(x)$ has been replaced with the parametric fit $\hat{f}(x;\hat{\theta})$.

This estimator is the only one of the improved methods which has bias of order h^2 rather than h^4, but if f is actually a member of the parametric family, then the h^2 bias term vanishes. Thus when the true density is close to the parametric target this estimator behaves like an efficient parametric estimator, yet when the density is not close to the target, the nonparametric part dominates. For the purposes of the simulation, a Gaussian distribution was used as the target distribution, with the mean and variance estimated by maximum likelihood. One of the strengths of this method, however, is that when external information about the likely shape of the unknown density is available (for example that it is skewed, or heavy-tailed) the target density for the parametric fit can be altered.

The standard form of $\hat{f}_{SP}(x)$ does not integrate to 1. This was noted by Hjort and Glad, but then dismissed by suggesting that the MISE was unaffected. For practical purposes, however, we also consider the renormalised

version of this estimator $\hat{f}_{SP}^{R}(x)$, with the correction achieved in the usual way.

5.3.7 Summary

We have by no means exhausted the list of higher order kernel density estimators. We have, however, tried to cover all of the main approaches that have been suggested to improve upon the standard KDE. It is not unreasonable to argue that the differences between variations of the same method are likely to be smaller than the between-method differences. As many of these estimators take the two-stage approach of pilot estimation followed by the final estimate, entire new families can be created by substituting one of these improved estimators for the standard KDE $\hat{f}_h(x)$ in the pilot estimation step. Jones, Signorini and Hjort [40] give a practical example of how this can be done. Furthermore, recent developments in the area of kernel density estimation seem to have concentrated more on semiparametric approaches, either directly as in $\hat{f}_{SP}(x)$ above, and in Hjort and Jones [41], or by fitting local polynomials to the log density, as in Loader [42]. Thus the methods we have considered are mainly those which are entirely nonparametric, with effectively no assumptions made about the shape of the true density.

We shall now briefly examine the asymptotic behaviour of these estimators. Note, however, that all of these estimators are 'better' than the standard KDE, in the sense that their asymptotic behaviour is improved. This is of very little comfort to the practical user of these methods, however, who wishes to know which, if any, of the improved methods should be applied to real data sets with perhaps a few hundred data points. Section 5.5 explores this important question via a simulation experiment.

5.4 Asymptotic Behaviour

With the exception of the semiparametric estimator $\hat{f}_{SP}(x)$, all of the estimators considered have asymptotic bias of order h^4 and asymptotic variance of order $(nh)^{-1}$. Jones and Signorini [31] show that the expressions for the bias and variance can all be expressed in the form

$$\mathrm{E}[\hat{f}(x) - f(x)] = \frac{1}{4!}h^4\sigma_4(\bar{K})\left[f''''(x) + \sum_{j=1}^{4} a_j g_j(x)\right], \quad (5.17)$$

$$\mathrm{var}[\hat{f}(x)] = \frac{f(x)}{nh}R(\bar{K}), \quad (5.18)$$

where $\bar{K}$ is some form of fourth order kernel, $\sigma_4(\bar{K})$ denotes the fourth moment of $\bar{K}$, the a_j are constant terms specific to each particular method and the $g_j(x)$ are functions of the true density such that

$$g_1(x) = \frac{f''(x)^2}{f(x)}, \quad g_2(x) = \frac{f'''(x)f'(x)}{f(x)},$$
$$g_3(x) = \frac{f'(x)^2 f''(x)}{f(x)^2}, \quad \text{and} \quad g_4(x) = \frac{f'(x)^4}{f(x)^3}.$$

The simplest expressions for the asymptotic bias are those for the fourth-order kernel estimators $\hat{f}_{4C}(x)$ and $\hat{f}_{4P}(x)$, where $a_1 = a_2 = a_3 = a_4 = 0$, $\bar{K} = K_{4C}$ or K_{4P}, and the bias depends on $f(x)$ only through its fourth derivative. For the other estimators the coefficients a_j are non-zero, constant for some estimators and functions of the kernel $\bar{K}$ for others, and the more complicated functions of the true density f are included. Details can be found in Jones and Signorini, but they are not repeated here, partly because the focus of this work is on the practical performance, and partly because the complex dependency upon f and its derivatives make theoretical comparison difficult unless the form of f is specified. Moreover, it may be argued that the concentration upon the asymptotic behaviour is one of the reasons why there are so many different improved KDEs and no clear recommendations

for small-sample applications, and so we proceed more or less directly to the simulation experiment.

For the semiparametric estimator, the asymptotic variance is simply the same as that for $\hat{f}_h(x)$ and the bias is given by

$$\mathrm{E}[\hat{f}_{SP}(x) - f(x)] = \frac{h^2\sigma_K^2}{2} f_0(x) \left(\frac{f}{f_0}\right)''(x) + O(h^4),$$

where $f_0(x)$ is the parametric density of the form $f(x;\theta)$ which is 'closest' in terms of the Kullback-Leibler distance metric to the true density $f(x)$. Thus the asymptotic bias is similar to that of the standard KDE, but when the parametric model can achieve a good fit to the data, the multiplier of h^2 will be small, giving reduced bias, and $o(h^2)$ bias if $f_0(x)$ is actually equal to $f(x)$ and the parametric model is true.

5.5 Simulation Experiment

We have described seven different approaches to improving upon the standard KDE, giving 13 possible single-bandwidth estimators and 3 two-bandwidth variations. Their practical performance was assessed using simulated data of size $n = 100$ and $n = 500$ from the first ten densities of Marron and Wand [16], shown in Figure 5.2. These densities, which can be called respectively, 'Gaussian', 'Skewed unimodal', 'Strongly skewed', 'Kurtotic unimodal', 'Outlier', 'Bimodal', 'Separated bimodal', 'Skewed bimodal', 'Trimodal' and 'Claw', are all mixtures of Gaussian distributions, and provide a wide ranging spectrum of densities on which to test the estimators.

Global accuracy of an estimator was measured by the integrated squared error (ISE), defined by

$$\mathrm{ISE}(\hat{f}) = \int \left[\hat{f}(x) - f(x)\right]^2 dx. \tag{5.19}$$

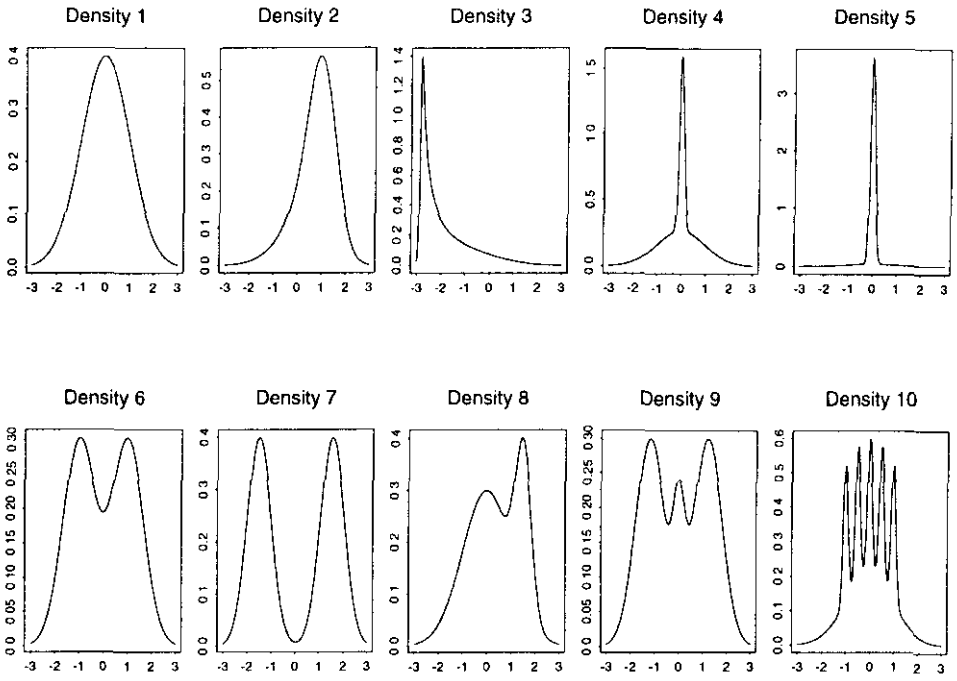

Figure 5.2: The Ten Test Densities

By choosing the bandwidth(s) h (and if necessary b) to minimise this quantity for each simulated dataset, a 'best-case' scenario was created. This decoupling of the choice of estimator from the choice of bandwidth is exactly the same approach as was taken for the binary regression simulation experiments of Part I.

Each density estimate was calculated on a grid of 301 points on the fixed range $[-3, 3]$, and the ISE numerically approximated by

$$\mathrm{ISE}^*(\hat{f}) = \frac{1}{50} \sum_{j=1}^{301} \left[\hat{f}(x_j) - f(x_j) \right]^2. \tag{5.20}$$

Whenever possible, linear binning algorithms of the type described by Fan and Marron [43] were used for computational efficiency. The number of grid points chosen was designed to minimise the potential errors due both to numerical integration over the grid and from the binning algorithms, but not to such an extent that the time required to complete the estimation

became prohibitive. For similar reasons of efficiency, a biweight kernel was used, of the form

$$K(u) = \frac{15}{16}(1-u^2)^2\, I(|u| < 1),$$

which has a finite domain implying that only points within a distance defined by the bandwidth contribute to the estimate at a point x. For all types of kernel smoothing differences in performance caused by using different kernels are very small.

For each of the ten test distributions, 1000 samples of size 100 and 500 were generated. A grid-search algorithm was used to find the bandwidth(s) which minimised the ISE for each particular estimator, ensuring that global and not local minima were found. Examples of cases where the ISE function has multiple local minima are given later.

Within each test distribution, the mean and standard error of the minimum ISE for each of the estimates were calculated over all simulations and compared. Each of the estimators was also compared to the second-order case $\hat{f}_h(x)$ in terms of the percentage reduction in ISE achieved by using the more complex estimator, that is

$$\frac{ISE[\hat{f}_h] - ISE[\hat{f}^*]}{ISE[\hat{f}_h]} \times 100,$$

where $\hat{f}^*(x)$ is one of the improved estimators. This quantity was summarised by using the median percentage reduction for each estimator over the 1000 simulated datasets, as the distribution of relative reductions was often skewed.

5.6 Results

The main results are presented in Tables 5.1 to 5.10, which show the mean ISE (with standard errors) and median reduction in ISE over the standard

KDE for 1000 simulated datasets from each model.

Before comparing the estimators with each other, we shall discuss the merits of the variations within each particular class of estimators on a case-by-case basis and in relation to the standard KDE. Results are discussed initially in terms of the $n = 100$ case, and differences in conclusions when $n = 500$ are noted.

5.6.1 Fourth-order Kernel Estimators

There is clearly very little difference between the two particular fourth-order kernels which were chosen. The mean ISE for the polynomial kernel K_{4P} is *always* less than that of the convolution kernel K_{4C}, but the difference is at most approximately 1%. This fits well with the finding that the choice of kernel in the standard KDE situation has a very limited effect upon the final estimate.

Despite its asymptotic attractions, these estimators are only significantly better than the standard method in the $n = 100$ case for Densities 1, 2, 5 and 7. There are only marginal changes for Models 4, 6 and 8, and actually significant increases in the ISE for Models 3, 9 and 10. For $n = 500$, a similar pattern emerges, but in this case the fourth-order estimators are never worse than the standard KDE, and the proportional reduction in the ISE compared to $\hat{f}_h(x)$ is larger. This disappointing practical performance of these estimators was previously noted by Marron and Wand [16].

The cases in which this method seems to work are those models which have very clear modes, with a failure to extract the fine detail of the more multimodal densities of Models 8, 9 and 10. When compared to the standard KDE, this estimator does seems to emphasise peaks, but because of the nature of the kernel, there is also a lowering of the troughs. This can be seen in

Model 1	**n=100**			**n=500**		
	Mean ISE	SE	ISE Reduction (%)	Mean ISE	SE	ISE Reduction (%)
$\hat{f}_h$	462	12	-	154	3	-
$\hat{f}_{4P}$	358	10	25.1	104	3	35.7
$\hat{f}_{4C}$	362	10	24.3	105	3	35.0
$\hat{f}_{JF}$	358	10	25.6	105	3	35.9
$\hat{f}_{JLN}$	319	8	31.4	93	2	42.3
$\hat{f}^R_{JLN}$	219	7	58.1	58	2	67.4
$\hat{f}^R_{JLN,2}$	168	6	71.7	44	1	77.7
$\hat{f}_{RC}$	365	10	21.5	115	3	26.3
$\hat{f}_{RC-V}$	383	9	17.8	116	2	29.2
$\hat{f}^R_{RC-V}$	252	8	51.1	60	2	65.6
$\hat{f}_{RC,2}$	343	10	25.7	107	3	29.5
$\hat{f}_V$	347	9	21.4	117	2	25.5
$\hat{f}_{V-V}$	268	8	46.2	58	2	67.1
$\hat{f}_{V,2}$	317	9	32.0	89	2	40.8
$\hat{f}_{VL}$	312	9	38.7	90	3	46.7
$\hat{f}_{SP}$	226	7	54.3	47	1	72.9
$\hat{f}^R_{SP}$	220	7	53.5	46	1	72.2

Table 5.1: Mean minimum achievable ISE and median percentage reduction in minimum achievable ISE compared to the standard estimator $\hat{f}_h(x)$ for Density 1. ISE values are $\times 10^5$.

Model 2	n=100			n=500		
	Mean ISE	SE	ISE Reduction (%)	Mean ISE	SE	ISE Reduction (%)
$\hat{f}_h$	755	17	-	234	5	-
$\hat{f}_{4P}$	642	15	18.3	176	4	27.5
$\hat{f}_{4C}$	645	15	17.8	177	4	26.9
$\hat{f}_{JF}$	628	15	20.3	174	4	28.3
$\hat{f}_{JLN}$	551	14	31.3	157	4	36.9
$\hat{f}^R_{JLN}$	477	13	43.1	135	4	48.7
$\hat{f}^R_{JLN,2}$	398	12	51.7	122	3	51.7
$\hat{f}_{RC}$	595	15	24.1	176	4	27.8
$\hat{f}_{RC-V}$	629	14	21.3	182	4	25.6
$\hat{f}^R_{RC-V}$	531	14	36.6	149	4	43.2
$\hat{f}_{RC,2}$	538	14	32.6	163	4	31.5
$\hat{f}_V$	556	14	29.8	172	4	30.9
$\hat{f}_{V-V}$	539	13	31.4	151	4	38.5
$\hat{f}_{V,2}$	520	14	34.3	143	4	40.8
$\hat{f}_{VL}$	569	14	28.7	158	4	36.4
$\hat{f}_{SP}$	605	14	23.2	176	4	26.7
$\hat{f}^R_{SP}$	604	14	23.2	177	4	27.0

Table 5.2: Mean minimum achievable ISE and median percentage reduction in minimum achievable ISE compared to the standard estimator $\hat{f}_h(x)$ for Density 2. ISE values are $\times 10^5$.

Model 3	**n=100**			**n=500**		
	Mean ISE	SE	ISE Reduction (%)	Mean ISE	SE	ISE Reduction (%)
$\hat{f}_h$	4227	53	-	1345	15	-
$\hat{f}_{4P}$	4471	51	-5.1	1340	14	2.0
$\hat{f}_{4C}$	4478	51	-5.3	1340	14	2.0
$\hat{f}_{JF}$	4415	50	-3.2	1337	14	2.7
$\hat{f}_{JLN}$	4369	61	-2.4	1317	16	2.6
$\hat{f}^R_{JLN}$	4470	61	-3.9	1319	16	3.4
$\hat{f}^R_{JLN,2}$	3984	54	3.4	1217	15	7.7
$\hat{f}_{RC}$	3588	61	19.4	1047	16	24.2
$\hat{f}_{RC-V}$	4506	60	-6.3	1405	17	-3.8
$\hat{f}^R_{RC-V}$	4597	60	-8.4	1399	17	-3.2
$\hat{f}_{RC,2}$	3297	57	25.2	981	15	29.1
$\hat{f}_V$	3689	58	15.3	1071	15	22.5
$\hat{f}_{V-V}$	4626	57	-9.5	1437	17	-6.5
$\hat{f}_{V,2}$	3425	56	21.4	1031	15	24.9
$\hat{f}_{VL}$	4342	53	-2.2	1306	15	4.0
$\hat{f}_{SP}$	4253	55	-0.2	1343	15	0.1
$\hat{f}^R_{SP}$	4266	55	-0.6	1343	15	0.2

Table 5.3: Mean minimum achievable ISE and median percentage reduction in minimum achievable ISE compared to the standard estimator $\hat{f}_h(x)$ for Density 3. ISE values are $\times 10^5$.

Model 4	n=100			n=500		
	Mean ISE	SE	ISE Reduction (%)	Mean ISE	SE	ISE Reduction (%)
$\hat{f}_h$	4152	59	-	1193	16	-
$\hat{f}_{4P}$	4168	55	1.5	1086	13	11.5
$\hat{f}_{4C}$	4170	55	1.4	1090	13	11.2
$\hat{f}_{JF}$	4035	55	5.5	1064	13	13.7
$\hat{f}_{JLN}$	3855	56	7.0	1007	13	15.1
$\hat{f}^R_{JLN}$	3882	55	5.7	994	13	16.2
$\hat{f}^R_{JLN,2}$	3652	53	10.4	978	13	17.2
$\hat{f}_{RC}$	3299	54	23.0	870	13	29.6
$\hat{f}_{RC-V}$	4003	57	3.7	1056	14	11.7
$\hat{f}^R_{RC-V}$	4042	55	3.0	1041	13	13.0
$\hat{f}_{RC,2}$	2814	55	35.3	748	13	39.3
$\hat{f}_V$	3035	55	29.5	770	13	37.3
$\hat{f}_{V-V}$	3999	54	4.0	985	13	17.7
$\hat{f}_{V,2}$	2698	54	39.3	644	12	48.9
$\hat{f}_{VL}$	3929	55	5.8	1035	13	14.3
$\hat{f}_{SP}$	4125	59	0.4	1183	16	0.8
$\hat{f}^R_{SP}$	4125	59	0.5	1182	15	0.9

Table 5.4: Mean minimum achievable ISE and median percentage reduction in minimum achievable ISE compared to the standard estimator $\hat{f}_h(x)$ for Density 4. ISE values are $\times 10^5$.

Model 5	**n=100**			**n=500**		
	Mean ISE	SE	ISE Reduction (%)	Mean ISE	SE	ISE Reduction (%)
$\hat{f}_h$	4908	110	-	1542	32	-
$\hat{f}_{4P}$	4039	94	19.3	1125	27	29.8
$\hat{f}_{4C}$	4063	95	18.8	1133	27	29.2
$\hat{f}_{JF}$	3967	94	21.1	1120	27	30.3
$\hat{f}_{JLN}$	3410	80	32.7	965	21	39.2
$\hat{f}^R_{JLN}$	2701	71	48.0	737	18	55.0
$\hat{f}^R_{JLN,2}$	2168	61	59.0	660	17	60.0
$\hat{f}_{RC}$	3770	95	25.6	1133	26	26.9
$\hat{f}_{RC-V}$	3847	91	23.3	1116	24	28.4
$\hat{f}^R_{RC-V}$	3355	84	35.0	931	22	41.7
$\hat{f}_{RC,2}$	3634	93	28.1	1108	26	28.2
$\hat{f}_V$	3271	87	35.4	982	21	36.1
$\hat{f}_{V-V}$	3118	76	37.9	710	17	54.9
$\hat{f}_{V,2}$	3138	86	40.5	846	21	45.9
$\hat{f}_{VL}$	3514	89	31.2	999	25	38.8
$\hat{f}_{SP}$	4775	102	2.8	1464	30	4.9
$\hat{f}^R_{SP}$	4776	102	2.7	1466	30	4.7

Table 5.5: Mean minimum achievable ISE and median percentage reduction in minimum achievable ISE compared to the standard estimator $\hat{f}_h(x)$ for Density 5. ISE values are $\times 10^5$.

Model 6	n=100			n=500		
	Mean ISE	SE	ISE Reduction (%)	Mean ISE	SE	ISE Reduction (%)
$\hat{f}_h$	717	13	-	223	4	-
$\hat{f}_{4P}$	704	14	3.4	190	4	17.1
$\hat{f}_{4C}$	707	14	3.1	191	4	16.3
$\hat{f}_{JF}$	712	14	2.5	191	4	16.4
$\hat{f}_{JLN}$	664	14	9.7	177	4	23.4
$\hat{f}^R_{JLN}$	658	15	11.2	171	4	26.4
$\hat{f}^R_{JLN,2}$	608	14	15.0	165	4	27.7
$\hat{f}_{RC}$	746	16	0.1	199	4	13.2
$\hat{f}_{RC-V}$	710	15	3.3	190	4	17.1
$\hat{f}^R_{RC-V}$	700	16	5.4	178	4	22.0
$\hat{f}_{RC,2}$	668	13	3.3	193	4	14.9
$\hat{f}_V$	742	16	-0.8	197	4	13.9
$\hat{f}_{V-V}$	724	16	2.3	182	4	20.7
$\hat{f}_{V,2}$	683	14	4.3	185	4	18.1
$\hat{f}_{VL}$	672	14	9.3	181	4	21.6
$\hat{f}_{SP}$	702	15	0.9	209	4	5.4
$\hat{f}^R_{SP}$	704	15	0.7	209	4	5.2

Table 5.6: Mean minimum achievable ISE and median percentage reduction in minimum achievable ISE compared to the standard estimator $\hat{f}_h(x)$ for Density 6. ISE values are $\times 10^5$.

Model 7	n=100			n=500		
	Mean ISE	SE	ISE Reduction (%)	Mean ISE	SE	ISE Reduction (%)
$\hat{f}_h$	1053	19	-	313	5	-
$\hat{f}_{4P}$	930	18	12.9	244	4	23.0
$\hat{f}_{4C}$	934	18	12.5	246	4	22.4
$\hat{f}_{JF}$	929	18	13.7	245	5	23.0
$\hat{f}_{JLN}$	841	16	20.9	220	4	31.4
$\hat{f}^R_{JLN}$	711	16	35.0	169	4	48.8
$\hat{f}^R_{JLN,2}$	641	15	42.3	152	3	54.8
$\hat{f}_{RC}$	1009	20	6.8	280	4	11.8
$\hat{f}_{RC-V}$	952	16	9.4	257	4	18.1
$\hat{f}^R_{RC-V}$	804	17	26.2	182	4	44.7
$\hat{f}_{RC,2}$	939	18	11.1	258	4	17.2
$\hat{f}_V$	966	19	8.5	271	4	12.0
$\hat{f}_{V-V}$	831	17	22.2	183	4	44.0
$\hat{f}_{V,2}$	925	19	13.7	239	4	23.0
$\hat{f}_{VL}$	846	17	21.6	219	4	32.9
$\hat{f}_{SP}$	1021	19	2.4	303	5	3.2
$\hat{f}^R_{SP}$	1018	19	2.9	302	5	3.9

Table 5.7: Mean minimum achievable ISE and median percentage reduction in minimum achievable ISE compared to the standard estimator $\hat{f}_h(x)$ for Density 7. ISE values are $\times 10^5$.

Model 8	n=100			n=500		
	Mean ISE	SE	ISE Reduction (%)	Mean ISE	SE	ISE Reduction (%)
$\hat{f}_h$	934	15	-	299	5	-
$\hat{f}_{4P}$	966	15	-2.5	281	5	7.8
$\hat{f}_{4C}$	968	16	-2.7	282	5	7.5
$\hat{f}_{JF}$	966	16	-1.9	279	5	8.8
$\hat{f}_{JLN}$	930	16	0.8	270	5	12.1
$\hat{f}^R_{JLN}$	924	16	1.8	269	5	12.2
$\hat{f}^R_{JLN,2}$	826	15	9.0	257	5	15.3
$\hat{f}_{RC}$	981	17	-2.9	275	5	10.3
$\hat{f}_{RC-V}$	966	16	-3.5	283	5	6.2
$\hat{f}^R_{RC-V}$	961	16	-2.4	281	5	7.5
$\hat{f}_{RC,2}$	873	15	2.7	263	5	11.8
$\hat{f}_V$	976	18	-2.2	272	5	12.5
$\hat{f}_{V-V}$	968	15	-3.3	291	5	4.3
$\hat{f}_{V,2}$	889	16	4.2	259	5	14.7
$\hat{f}_{VL}$	934	15	1.0	273	5	10.8
$\hat{f}_{SP}$	950	17	-1.2	298	5	0.1
$\hat{f}^R_{SP}$	951	17	-1.3	298	5	0.0

Table 5.8: Mean minimum achievable ISE and median percentage reduction in minimum achievable ISE compared to the standard estimator $\hat{f}_h(x)$ for Density 8. ISE values are $\times 10^5$.

Model 9	n=100			n=500		
	Mean ISE	SE	ISE Reduction (%)	Mean ISE	SE	ISE Reduction (%)
$\hat{f}_h$	864	13	-	284	4	-
$\hat{f}_{4P}$	876	14	-0.9	280	4	2.9
$\hat{f}_{4C}$	879	14	-1.1	280	4	2.8
$\hat{f}_{JF}$	879	14	-1.2	280	4	3.2
$\hat{f}_{JLN}$	827	13	4.5	269	4	6.2
$\hat{f}^R_{JLN}$	813	13	5.6	268	4	6.8
$\hat{f}^R_{JLN,2}$	744	13	11.9	248	4	11.7
$\hat{f}_{RC}$	906	15	-3.4	284	4	1.8
$\hat{f}_{RC-V}$	873	14	-0.6	275	4	4.5
$\hat{f}^R_{RC-V}$	861	14	1.0	273	4	5.6
$\hat{f}_{RC,2}$	816	13	1.9	265	4	4.9
$\hat{f}_V$	903	15	-3.4	279	4	3.2
$\hat{f}_{V-V}$	885	14	-1.4	271	4	6.3
$\hat{f}_{V,2}$	834	14	3.1	256	4	9.3
$\hat{f}_{VL}$	838	13	3.8	269	4	6.3
$\hat{f}_{SP}$	852	15	0.6	280	4	0.7
$\hat{f}^R_{SP}$	854	15	0.5	281	4	0.8

Table 5.9: Mean minimum achievable ISE and median percentage reduction in minimum achievable ISE compared to the standard estimator $\hat{f}_h(x)$ for Density 9. ISE values are $\times 10^5$.

Model 10	n=100			n=500		
	Mean ISE	SE	ISE Reduction (%)	Mean ISE	SE	ISE Reduction (%)
$\hat{f}_h$	3652	36	-	1110	11	-
$\hat{f}_{4P}$	3790	37	-3.3	1024	11	8.1
$\hat{f}_{4C}$	3801	37	-3.6	1028	11	7.6
$\hat{f}_{JF}$	3779	37	-3.0	1021	11	8.6
$\hat{f}_{JLN}$	3684	37	-0.3	996	11	10.4
$\hat{f}^R_{JLN}$	3754	36	-2.0	994	11	10.6
$\hat{f}^R_{JLN,2}$	3569	36	1.6	987	11	11.1
$\hat{f}_{RC}$	3834	39	-3.8	1019	11	9.1
$\hat{f}_{RC-V}$	3755	37	-2.4	1018	10	8.8
$\hat{f}^R_{RC-V}$	3833	36	-4.3	1012	11	9.1
$\hat{f}_{RC,2}$	3513	36	2.6	1008	11	9.5
$\hat{f}_V$	3708	39	-0.8	977	11	12.8
$\hat{f}_{V-V}$	3813	35	-4.2	1017	11	8.6
$\hat{f}_{V,2}$	3531	38	3.1	963	11	13.9
$\hat{f}_{VL}$	3697	37	-1.0	1013	11	8.9
$\hat{f}_{SP}$	3666	36	-0.1	1108	11	0.2
$\hat{f}^R_{SP}$	3667	36	-0.2	1108	11	0.3

Table 5.10: Mean minimum achievable ISE and median percentage reduction in minimum achievable ISE compared to the standard estimator $\hat{f}_h(x)$ for Density 10. ISE values are $\times 10^5$.

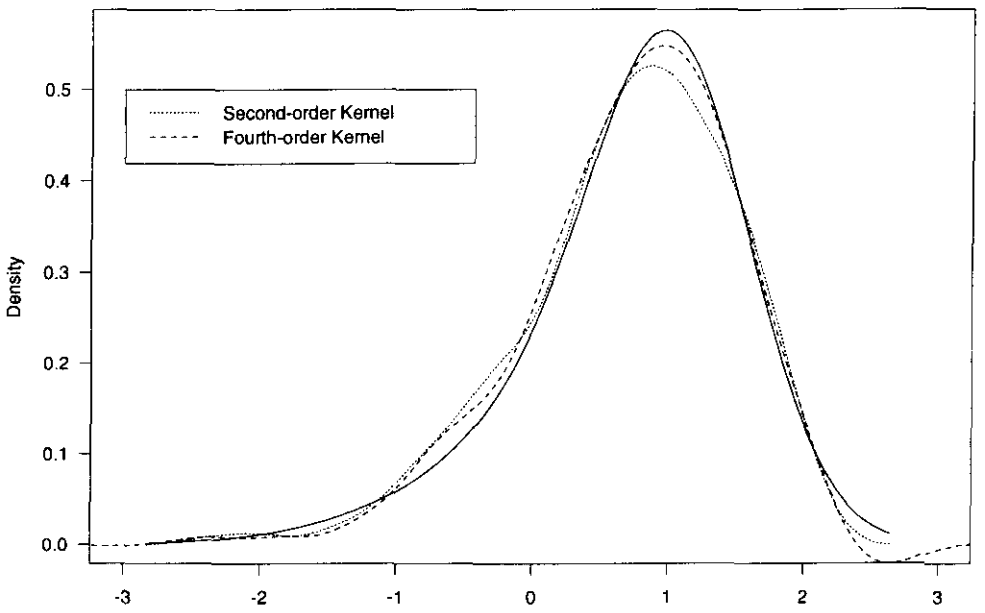

Figure 5.3: Example of fourth-order kernel estimator for Model 2. The ISEs are 246×10^{-5} and 163×10^{-5} for the $\hat{f}_h(x)$ and $\hat{f}_{4P}(x)$ estimators respectively. The true density is shown by the solid line.

Figure 5.3, which is an example taken from the skewed unimodal Density 2, showing both the ISE-optimal estimates for $\hat{f}_h(x)$ and $\hat{f}_{4P}(x)$. The enhancement of the peak by the fourth-order estimator without a corresponding loss of smoothness in the left hand tail can clearly be seen. Also apparent is the negativity of the estimate in the right hand tail, a feature which can be particularly undesirable when the density estimation is a component of a more complex statistical procedure such as discriminant analysis.

5.6.2 Multiplicative Bias-Correcting Estimators

The performance of the simple multiplicative bias-correcting estimator $\hat{f}_{JF}(x)$ is almost identical to that of the fourth-order kernel methods. This connection between the methods can be argued heuristically by noting that taking

the logarithm of $\hat{f}_{JF}(x)$ gives

$$2\log[\hat{f}_h(x)] - \log[\tilde{f}_h(x)],$$

where $\hat{f}_h(x)$ and $\tilde{f}_h(x)$ are estimated using $K(u)$ and $(K * K)(u)$ respectively. Comparison with equation (5.6) suggests a possible link between the methods. In common with the fourth-order kernel estimators, this method gives higher peaks and smoother tails, although unlike $\hat{f}_{4P}(x)$ and $\hat{f}_{4C}(x)$, the estimator is always positive.

For the other multiplicative estimators of Jones, Linton and Nielsen, the single bandwidth cases are identical to the results published in the original report [33]. The empirical renormalisation to have unit integral is nearly always beneficial, with $\hat{f}^R_{JLN}(x)$ giving large improvements over $\hat{f}_{JLN}(x)$ for Densities 1, 2, 5 and 7. Renormalisation is only worse for the case of $n = 100$ and Densities 3, 8 and 10, and even then the increase in error caused by adjusting the estimate is marginal.

When compared to the standard KDE, $\hat{f}^R_{JLN}(x)$ and $\hat{f}_{JLN}(x)$ are clearly the best performing estimators so far. Only for the strongly skewed (3) and claw (10) densities in the $n = 100$ case is $\hat{f}^R_{JLN}(x)$ worse than the basic estimator, and quite often the median relative improvement in ISE is over 30%.

To examine qualitatively how this estimator achieves such good estimation consider Figure 5.4, which shows an example for a dataset from Density 2 where the relative improvement in ISE is about 43%, close to the median value. Although $\hat{f}^R_{JLN}(x)$ works by tightening peaks, it also seems to alter their location when compared to $\hat{f}_h(x)$. In Figure 5.4 we can see that the latter estimate is shifted to the right of the true peak, but is approximately the right height. The optimal estimate using $\hat{f}^R_{JLN}(x)$, however, is able to centre the peak appropriately.

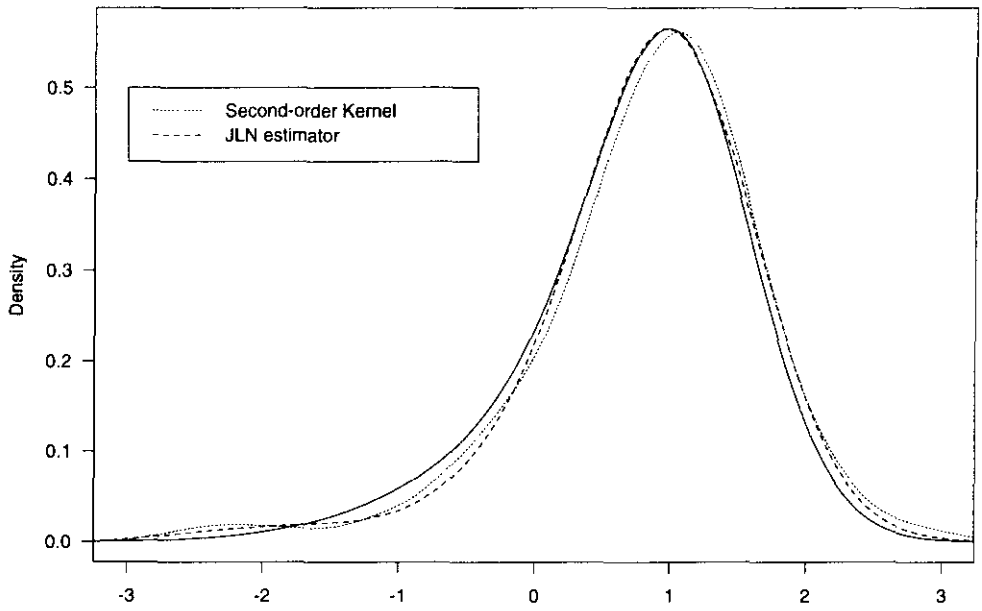

Figure 5.4: Example of $\hat{f}^R_{JLN}(x)$ kernel estimator for Model 2. The ISEs are 269×10^{-5} and 153×10^{-5} for the $\hat{f}_h(x)$ and $\hat{f}^R_{JLN}(x)$ estimators respectively. The true density is shown by the solid line.

The reasons for the failure of this estimator for the strongly skewed density of Model 3 are unclear. Both densities 4 and 5 are similar, but produce opposite results, although one could argue that density 3 has a less sharply defined peak. The minimum ISE in this case is obtained with a relatively small bandwidth, which balances the height of the main mode against spurious modes caused by undersmoothing the tail. Although $\hat{f}^R_{JLN}(x)$ can improve the main peak, it does not seem to reduce these spurious modes, and in some cases can even enhance them, as shown in Figure 5.5, which is an example from density 3 where the more sophisticated estimator offers no improvement. The problem of the 'peak-enhancement' property operating both on the true mode and the spurious ones simultaneously can be seen, with the improvement in the main peak with $\hat{f}^R_{JLN}(x)$ being counterbal-

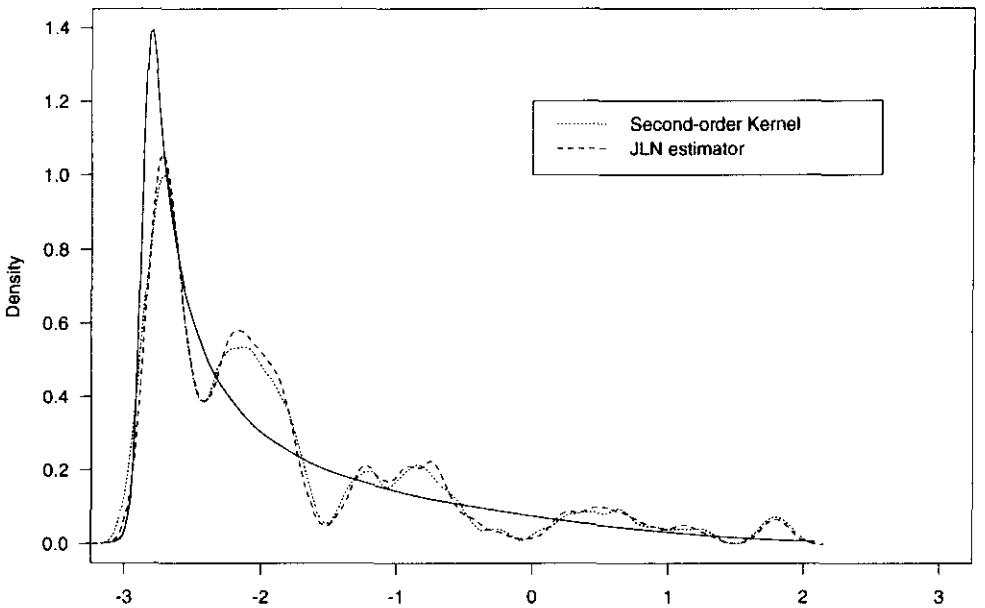

Figure 5.5: Example of $\hat{f}^R_{JLN}(x)$ kernel estimator for Model 3. The ISEs are 5785×10^{-5} and 6831×10^{-5} for the $\hat{f}_h(x)$ and $\hat{f}^R_{JLN}(x)$ estimators respectively. The true density is shown by the solid line.

anced by the raising of the false mode at -2. Despite these problems, this renormalised estimator is exceedingly promising.

For the two-bandwidth case the results are simpler, as $\hat{f}^R_{JLN,2}(x)$ is uniformly superior to both the single bandwidth multiplicative estimators and the standard KDE, although the benefits over $\hat{f}^R_{JLN}(x)$ are modest except in the easiest models.

5.6.3 Transformation Estimators

For local bandwidth variation forms of the transformation estimators, once again renormalisation to give proper densities is beneficial, with $\hat{f}^R_{RC-V}(x)$ proving to be superior, or at least nearly as good as, $\hat{f}_{RC-V}(x)$ in all cases. Indeed $\hat{f}^R_{RC-V}(x)$ turns out to be superior to the original Ruppert-Cline

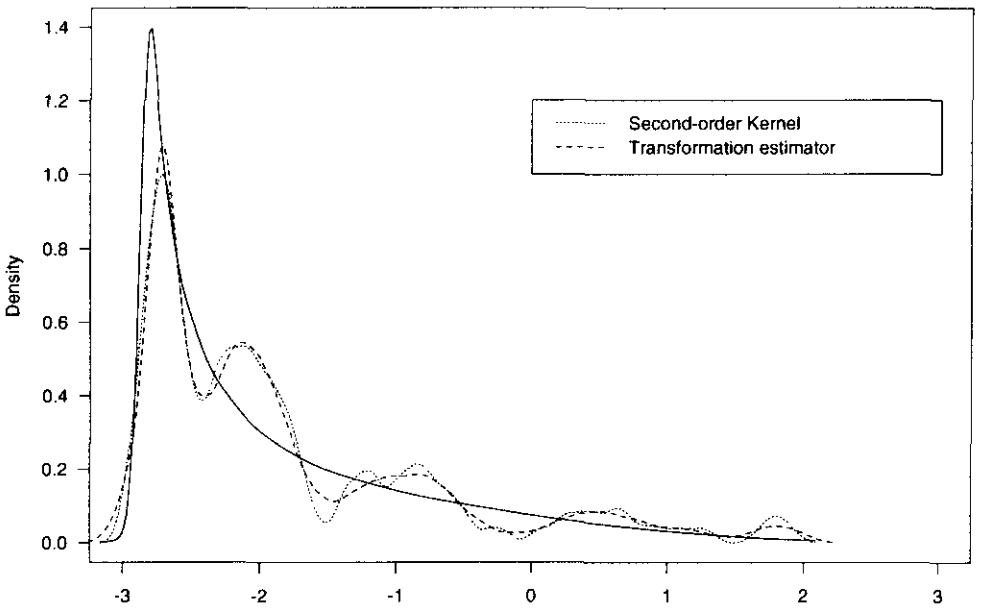

Figure 5.6: Example of $\hat{f}_{RC}(x)$ kernel estimator for Model 3, using the data from Figure 5.5. The true density is shown by the solid line.

estimator $\hat{f}_{RC}(x)$ in all situations except for Densities 3 and 4, at both sample sizes.

Comparing $\hat{f}^{R}_{RC-V}(x)$ to $\hat{f}_{h}(x)$, we can see that with the exception of the highly skewed Density 3, the former gives great improvement in ISE for the unimodal densities (of which Density 7 can almost be considered as the peaks are almost disjoint) and marginal improvement for the multimodal densities. Model 3 is exactly the situation in which the simple transformation estimator $\hat{f}_{RC}(x)$ works. This is hardly surprising, as this is where we would expect transformation of the data to improve matters. Figure 5.6 shows the ISE-optimal estimate using $\hat{f}_{RC}(x)$ for exactly the same dataset as Figure 5.5. It is clear that this estimator can enhance the leftmost mode, but can still retain a greater degree of smoothness in the right tail of the density.

The advantages of using two bandwidths over one are unclear, as $\hat{f}_{RC-2}(x)$

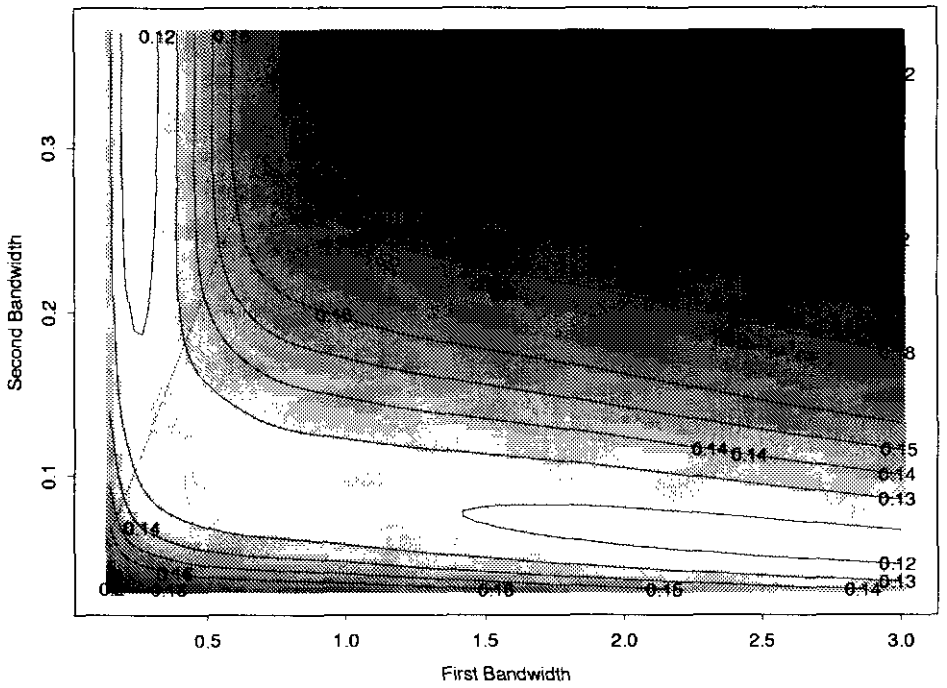

Figure 5.7: ISE function for two bandwidth transformation estimator for a dataset from the strongly skewed density

is only a slight improvement upon $\hat{f}_{RC}(x)$. Note that the latter is merely a constrained version of the former, so that two bandwidths will always improve estimation, and the question becomes *by how much*. When compared to the better performing locally rescaled estimator $\hat{f}^{R}_{RC-V}(x)$, the two bandwidth estimator is actually poorer in terms of median ISE for the easy to estimate Models 1, 2, 5 and 7.

In addition to this disappointing conclusion regarding the two bandwidth estimator, there are also practical difficulties in finding the ISE-optimal bandwidths. If we examine the ISE as a bivariate function of the bandwidths for a typical dataset from Density 3, as shown in Figure 5.7, we can see that there is an L-shaped valley with local minima in both the horizontal and vertical arms. Although this density is an extreme example, this is a

typical feature for all models considered. The dotted line shows the possible values for the single bandwidth case, which always intersect with the vertical arm. It would appear from these observations that the estimator can operate in two different ways; with a small initial bandwidth and a smooth estimate of the uniform density, or conversely, with a smooth estimate initially and the estimate of the transformed data providing the detail. For the unimodal densities the first case nearly always gives the overall minimum, with the result that $\hat{f}_{RC}(x)$ performs well, whereas the multimodal densities can produce minima corresponding to the second case.

Thus of all the transformation estimators, it would appear that $\hat{f}^R_{RC-V}(x)$ is, if not always best, at least not too far away from being so.

5.6.4 Variable Bandwidth and Location Estimators

In the previous section the locally modified bandwidth estimator seemed to decrease the ISE relative to the basic transformation estimator, and a similar pattern emerges for the variable kernel methods. Here, $\hat{f}_V(x)$ is only better than $\hat{f}_{V-V}(x)$ for Densities 3, 4 and possibly 10. In the first two of these cases, however, the improvement is relatively large, in contrast to the more modest differences in favour of $\hat{f}_{V-V}(x)$ for the easier densities. Furthermore, the performance of $\hat{f}_V(x)$ is between 15% and 35% better that the standard KDE for *all* of the first five unimodal models, leading us to prefer this original variable kernel to the more complicated locally rescaled version. Qualitatively, this estimator operates in a very similar way to the Ruppert-Cline transformation estimator $\hat{f}_{RC}(x)$, although the mean ISE is almost always better.

The variable location estimator, which has received very little attention to date, performs more than adequately in comparison to $\hat{f}_h(x)$. Its be-

haviour is similar to that of $\hat{f}_{V-V}(x)$, and so again suffers from an inability to cope well with the sharp peak and heavy tails of Densities 3 and 4.

The two bandwidth variable kernel estimator $\hat{f}_{V,2}(x)$ is better that the single bandwidth version, but not dramatically so, mirroring the uninspiring results for the transformation estimators. Typically, when moving from one bandwidth to two, the median ISE reduction only increases by at most 10%, again suggesting that the extra work involved in finding two data-dependent bandwidths may not be worthwhile.

5.6.5 Semiparametric Estimator

For the only order h^2 bias estimators we consider, the effect of renormalisation seems be reduced, with little or no difference between the two forms. Interestingly, however, these estimators are better, or a least as good, for all models considered, even the previously intractable strongly skewed model. Examination of the actual ISE-optimal estimates in these cases shows that the estimator is behaving almost exactly like the standard KDE. This is in line with the motivation behind the development of this method; better estimation when (in this case) the Gaussian model is approximately correct, and order h^2 nonparametric behaviour elsewhere. Thus for Density 3 the parametric estimates for the Gaussian target will give a mean around -2.8, with an estimated standard deviation of approximately 1. This implies that for points X_i which are above about -2, the values of $\hat{f}(X_i;\hat{\theta})$ from Equation (5.16) will be in the right tail of this Gaussian distribution, and thus will be approximately equal. This forces the estimator to behave like $\hat{f}_h(x)$ in this region, and so this estimator does not suffer from the enhancement of false peaks which reduces the effectiveness of many of the order h^4 estimators.

5.6.6 Comparisons between Estimators

Is it possible to derive general conclusions about the relative merits of the various estimators from the extensive results reported above? It is clear that it is possible to improve upon the standard KDE for estimating many different shapes of density. Indeed, it is only for small datasets of size $n = 100$ that none of the estimators are able to resolve the fine structure of Models 8, 9 and 10.

It would seem that the renormalised estimator of Jones, Linton and Nielsen [33] is amongst the best of the proposed single bandwidth estimators, except for Models 3 and 4, which each exhibit a strong peak with large 'shoulders' (as opposed to Model 5 which decays rapidly to zero away from the peak). In these cases it would appear that either the transformation or variable kernel methods should be used in preference. Given that both $\hat{f}^R_{JLN}(x)$ and $\hat{f}_V(x)$ involve the use of a pilot estimator of $\hat{f}(X_i)$ it may be possible to combine these estimators in the manner pursued by Jones, Signorini and Hjort [40] to give an estimator with order h^4 asymptotic bias and good performance in all situations.

To give a numerical summary of the relative merits of the general approaches, the most promising of each family and the standard KDE were selected, and for each dataset the minimum ISEs were ranked. Table 5.11 shows the mean ranking over 1000 datasets for each model and sample size.

These results confirm our previous conclusions; $\hat{f}^R_{JLN}(x)$ is on average the best estimator in most circumstances, and even for Models 3 and 4 it is nearly always second best to $\hat{f}_V(x)$. Table 5.11 also shows how difficult it is to estimate Densities 8, 9 and 10 with only 100 data points, as there is very little to choose between the average rankings of all the estimators.

The performance of the two bandwidth estimators, at least in comparison

Density	Mean Rank of Minimum ISE					
	$\hat{f}_h$	$\hat{f}_{4P}$	$\hat{f}^R_{JLN}$	$\hat{f}^R_{RC-V}$	$\hat{f}_V$	$\hat{f}^R_{SP}$
n=100						
MW(1)	5.2	4.0	1.9	2.5	4.3	3.0
MW(2)	4.9	3.9	2.1	2.9	3.2	4.0
MW(3)	3.3	4.2	3.9	4.8	1.4	3.4
MW(4)	4.5	4.3	3.1	3.8	1.1	4.1
MW(5)	5.4	3.9	1.7	2.6	2.3	5.1
MW(6)	3.9	3.4	2.4	3.4	4.1	3.8
MW(7)	5.0	3.6	1.6	2.4	4.1	4.4
MW(8)	3.3	3.7	3.0	3.8	3.6	3.6
MW(9)	3.5	3.6	2.6	3.6	4.1	3.5
MW(10)	2.8	4.0	3.6	4.5	3.1	3.0
n=500						
MW(1)	5.5	4.1	2.1	2.1	4.8	2.4
MW(2)	5.1	3.5	1.9	2.7	3.7	4.1
MW(3)	4.0	3.7	3.4	4.7	1.2	4.0
MW(4)	5.5	4.0	2.5	3.3	1.0	4.7
MW(5)	5.6	3.5	1.4	2.6	3.2	4.8
MW(6)	4.9	3.1	2.0	2.6	3.8	4.6
MW(7)	5.2	3.4	1.4	1.8	4.4	4.8
MW(8)	4.4	3.2	2.5	3.7	2.6	4.5
MW(9)	4.0	3.6	2.8	3.2	3.6	3.8
MW(10)	5.3	3.5	2.3	3.0	1.8	5.1

Table 5.11: Mean ranking of the minimum ISE within each model and sample size for the five 'best' representatives of each approach and the standard KDE.

to the improved single bandwidth estimators, was disappointing. Although they did seem to consistently produce lower optimal ISE values, the gains were modest. Given that the problem of choosing two data-dependent bandwidths is much more complicated that choosing one, and the fact that for binary regression the gains in the optimal setting were not carried through to practice, it is difficult to see any merit in pursuing the development of these particular estimators when nearly as good single bandwidth alternatives are available.

5.7 Conclusions

In this chapter we have described and investigated a great many suggested improvements to the standard kernel density estimator. The focus has been on small-sample performance rather than asymptotic behaviour, and the major conclusions were derived from an extensive simulation experiment.

It would seem that, except in a few special cases, the multiplicative bias-correcting estimator $\hat{f}^R_{JLN}(x)$ is generally the best of the estimators considered. For densities which are strongly skewed or have very heavy tails, however, the variable kernel method $\hat{f}_V(x)$ may be more appropriate. A suggested way of combining these estimators is also given.

The use of two bandwidths, although novel, did not radically improve estimation. The very real complications that two bandwidths entail suggest that, at least in practical terms, these estimators should not be favoured over the best of the single bandwidth cases.

As in the study of binary regression, it is worth noting that these findings are based upon using the bandwidth which minimises the chosen error function in every case. In practice, a data-dependent bandwidth must be chosen, and it is by no means certain that the encouraging 'best-case' perfor-

mance of $\hat{f}^R_{JLN}(x)$ will be reproduced. With this in mind, the next chapter explores a very simple bandwidth selection rule, both for $\hat{f}_h(x)$ and the most promising of these improved estimators.

Chapter 6

An Evaluation of Some Rule-of-Thumb Bandwidth Selectors for Density Estimation

6.1 Introduction

The selection of an appropriate bandwidth for density estimation is not a trivial task, and there is a vast and still expanding literature on the subject. Comprehensive recent reviews of this area are given by Wand and Jones [14], and by Jones, Marron and Sheather [44]. The practical performance of the most popular are assessed by Sheather [45], and Park and Turlach [46].

Many bandwidth selection methods that have been suggested all stem from the standard expression for the asymptotic MISE of the second order

KDE for some unknown density f, namely

$$AMISE(\hat{f}) = \frac{1}{4}h^4(\sigma_K^2)^2 R(f'') + \frac{1}{nh}R(K), \tag{6.1}$$

where $R(g) = \int g^2$, and σ_K^2 is the variance of the kernel i.e. $\int u^2 K(u)du$.

Minimising equation (6.1) over h gives the asymptotically MISE-optimal bandwidth as

$$h_{\mathrm{OPT}} = \left[\frac{R(K)}{(\sigma_K^2)^2 R(f'')}\right]^{1/5} n^{-1/5}. \tag{6.2}$$

Now both σ_K^2 and $R(K)$ are known, as is n. Therefore the only unknown quantity in equation (6.2) is $R(f'')$. So-called 'plug-in' bandwidth estimators rely upon various methods of estimating this functional. This adds another level of complexity to the problem, in that this estimate of $R(f'')$ will often itself require a pilot bandwidth g to be chosen, which itself depends on functions of higher derivatives and so on *ad infinitum*. Many of the proposed methods deal with the number of iterations of this process beyond which the gains in estimation are negligible. For good examples of this kind of iterative bandwidth selection see Sheather and Jones [17], and Park and Marron [47].

A simple alternative to these computationally expensive methods, often called 'quick and dirty' (QAD) methods, is to substitute for f in equation (6.2) a normal density with standard deviation σ. This gives

$$R(f'') = \frac{3}{8\sqrt{\pi}}\ \sigma^{-5},$$

and hence equation (6.2) becomes

$$h_{\mathrm{OPT}} = \left[\frac{8\sqrt{\pi}R(K)}{3(\sigma_K^2)^2}\right]^{1/5} n^{-1/5}\,\sigma. \tag{6.3}$$

This reduces the problem to that of finding a reasonable estimate of the scale σ, for which there are many alternatives. Note that it is always possible to replace the Gaussian reference density we have used here by any

other distribution in cases where information about the likely shape of the unknown density is available.

In this chapter we summarise the properties of some simple QAD bandwidth estimators (differing only in their methods of estimating σ) for the second-order KDE case, then extend the method to provide simple bandwidth estimators for the most promising fourth-order density estimator from the previous chapter. Simulation experiments complementing those previously carried out are used to determine whether the theoretical and idealised advantages of these higher order estimators can be achieved in practice.

In all that follows, we shall use a quartic kernel, so $\sigma_K^2 = 1/7$ and $R(K) = 5/7$, and equation (6.3) can be written simply as

$$h_{\mathrm{OPT}} = \left[\frac{280\sqrt{\pi}}{3}\right]^{1/5} n^{-1/5}\,\sigma. \tag{6.4}$$

6.2 Scale Estimation

The obvious choice for an estimator of σ for use in equation (6.4) is simply the sample standard deviation, denoted by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Silverman [26] considers both this estimate and the alternative based upon the normalised sample inter-quartile range ($\widehat{IQR}$),

$$t = \frac{\widehat{IQR}}{\Phi^{-1}(\frac{3}{4}) - \Phi^{-1}(\frac{1}{4})} = \frac{\widehat{IQR}}{1.348},$$

where Φ^{-1} is the inverse of the cumulative distribution function for the Gaussian distribution with mean 0 and variance 1. Silverman suggested that a useful rule-of-thumb (ROT) estimate of scale was to take 90% of the

minimum of these two measurements, defined as

$$\hat{\sigma}_{\mathrm{ROT}} = 0.9 \min(s, t).$$

The arbitrary figure of 0.9 is taken because of an observed tendency of the simple minimum of these two estimates to oversmooth the density.

Janssen et al.[48] construct another scale estimator based upon the minimum value of first differences

$$RD1_j = X_{[j+q]} - X_{[j-q]}, \tag{6.5}$$

where $X_{[k]}$ is the kth order statistic of the data and q is an integer governing the span of the differences. By dividing the minimum over j of these values by its expected value μ_1, a reasonable estimate of the scale, denoted by $\hat{\sigma}_{\mathrm{D1}}$ is calculated, given by

$$\hat{\sigma}_{\mathrm{D1}} = \min_j RD1_j/\hat{\mu}_1,$$

where $\hat{\mu}_1$ is a function of β and n defined by

$$\hat{\mu}_1 = 2\Phi^{-1}(\beta + 1/2) - 4^{-7/8}(n/25)^{-2/3}.$$

Notice that the choice of q is itself a smoothing problem, but although Janssen et al. gloss over this fact, they do provide some simulation evidence that a choice of q equal to the largest integer value not larger than βn where $\beta = 0.2$ works well, and that the final estimate of the bandwidth is robust to variations in this rather arbitrary value of β.

However, by considering non-Gaussian target densities, any of which can be approximated arbitrarily closely by a mixture of Gaussian densities, it can be shown that what is really needed is an estimate of the scale of the dominant component and its relevant weight.

To see this, let f be of the form

$$f(x) = \sum_{k=1}^{m} w_k \, \phi_{\sigma_k}(x - \mu_k)$$

where $\phi_\sigma(x - \mu)$ is the density of an $N(\mu, \sigma^2)$ distribution. Without loss of generality, assume that the dominant component of f is the first one i.e. $k = 1$. In this case 'dominant' is taken to mean that the modal value of f is at $x = \mu_1$. Then $\hat{\sigma}_{\mathrm{D1}}$ estimates σ_1, but is biased by a factor involving w_1. In a similar vein to equation (6.5), third differences of smoothed order statistics are calculated and are used to estimate the curvature of the density and through this w_1. Finally, the bias-corrected estimate of scale is defined by

$$\hat{\sigma}_{\mathrm{D3}} = \hat{w}_1^{4/5} \hat{\sigma}_{\mathrm{D1}}.$$

This estimator can, however, fail in two distinct ways. Firstly, the smoothed third difference is essentially a function of the third derivative of the inverse distribution function F^{-1}, and so is related to the curvature of the density f''. In fact, it can be shown that, asymptotically, the smoothed third difference is essentially dominated by a term involving $-f''$. At the dominant peak the curvature should be negative and the third difference positive. However, it is not unusual for the estimate to be negative, implying a positive curvature and in this case (designated a type A error), the estimate of w_1 is infeasible. Alternatively, even when the third difference is positive, the estimate of w_1 can be greater than 1 (designated a type B error). In either case, in the absence of a sensible estimate of w_1, $\hat{\sigma}_{\mathrm{D3}}$ is replaced by $\hat{\sigma}_{\mathrm{D1}}$.

Following Section 6 of Janssen et al., the final scale estimate (referred to in the original paper as the 'super scale' estimator) is defined to be the minimum of $\hat{\sigma}_{\mathrm{D3}}$ and the sample standard deviation s, which, as stated in

the paper, addresses the problem when the underlying density f has several tall peaks close together. Thus

$$\hat{\sigma}_{\mathrm{SS}} = \min(s, \hat{\sigma}_{\mathrm{D3}}).$$

This gives us four different estimates of scale to substitute for σ in equation (6.4). The performance of these estimates was assessed by means of a simulation experiment, using the same ten test densities from the previous chapter.

6.3 Fourth Order Bias Kernel Density Estimation

The original motivation for this study of some simple bandwidth selection rules was to find out if some of the apparent gains in accuracy observed by higher-order kernel density estimation could be translated into improved practical performance. The methods of the previous chapter decoupled the problem of bandwidth selection from the choice of fourth-order estimator by selecting the ISE-optimal bandwidth in each case. This is obviously not possible in reality, as the true density f remains unknown. The key question is whether the impressive performance of the two versions of the multplicative bias-correction estimator of Jones, Linton and Nielsen [33], $\hat{f}_{JLN}$ and $\hat{f}^{R}_{JLN}$, is maintained when a data-driven bandwidth is used.

Furthermore, given that the minimum achievable ISE for the higher-order estimators is smaller than that of the standard KDE, it can be argued that rather than a single optimum, there is a range of bandwidths which will all achieve smaller ISEs than the simple case. This can be thought of as a type of robustness to the estimation of the bandwidth for these higher-order estimators.

For the fourth-order estimator $\hat{f}_{JLN}$ and considering equations (5.17)

and (5.18), we can express the asymptotic MISE in a form similar to equation (6.1), namely

$$AMISE(\hat{f}_{JLN}) = \left(\frac{1}{4!}\right)^2 (\sigma_L^2)^2 R(B_f) h^8 + \frac{1}{nh} R(L), \tag{6.6}$$

where B_f is the bias term involving f for this estimator and $L(u) = K_{4C}(u)$, the fourth-order convolution kernel as defined in equation (5.6). Equation (6.6) can be applied to any higher order estimator from the previous chapter, with differing values for B_f and L depending upon which particular estimator is used. Details of the various kernel functions L and bias terms B_f are unified and summarised in Jones and Signorini [31].

Minimising equation (6.6) over h gives

$$h^*_{\mathrm{OPT}} = \left[\frac{72\, R(L)}{(\sigma_L^2)^2 R(B_f)}\right]^{1/9} n^{-1/9}. \tag{6.7}$$

Now, for L based upon the quartic kernel, tedious algebra leads to the results $R(L) = 1.00663$ and $\sigma_L^2 = 6/49$. Finally, we have that

$$B_f = f\left[\frac{f''}{f}\right]'',$$

so if we replace f by a Gaussian reference distribution with variance σ^2, it can be shown that

$$R(B_f) = \frac{2}{\sqrt{\pi}}\, \sigma^{-9}.$$

Thus QAD bandwidth estimators for $\hat{f}_{JLN}$ can be based on the simple formula

$$\widehat{h^*_{\mathrm{OPT}}} = 2.53243\, n^{-1/9}\, \hat{\sigma}. \tag{6.8}$$

This can be compared with equation (6.4) which can be rewritten as

$$\widehat{h_{\mathrm{OPT}}} = 2.77794\, n^{-1/5}\, \hat{\sigma}. \tag{6.9}$$

Similar expressions can be calculated for other fourth-order estimators, but we shall focus solely on the most promising of these in the simulation experiment.

6.4 Simulation Experiment

For each of the ten test densities used in the previous chapter and shown in Figure 5.2, the same 1000 datasets of size $n = 100$ and 1000 datasets of size $n = 500$ were used to test the measures of scale.

For each dataset, the four measures of scale, $\hat{\sigma}_{\mathrm{ROT}}$, $\hat{\sigma}_{\mathrm{D1}}$, $\hat{\sigma}_{\mathrm{D3}}$ and $\hat{\sigma}_{\mathrm{SS}}$ were calculated. These were used in equations (6.4) and (6.8) to calculate QAD bandwidths for the standard kernel density estimator $\hat{f}$ and both the standard and rescaled multiplicative bias-correcting estimates of Jones, Linton and Neilsen, $\hat{f}_{JLN}$ and $\hat{f}^{R}_{JLN}$. The achieved ISE using these bandwidths for each estimator could then be compared to the optimal ISE values derived from the work of the previous chapter. Janssen et al. [48] studied the performance of their scale estimators only in the simple second order case, but came to the conclusion that the super scale estimator $\hat{\sigma}_{\mathrm{SS}}$ was superior in nearly all circumstances. The aim of this study was to confirm these results and extend and evaluate the bandwidth selection rules using improved estimators.

In the discussion of the papers by Sheather [45] and Park and Turlach [46], which evaluate several cross-validatory and plug-in bandwidth estimators, Terrell proposes the use of a simple rule-of-thumb estimator applied with a simple fourth order kernel. This is shown by Sheather in his rejoinder to compare unfavourably with the Sheather-Jones plug-in bandwidth estimator [17], but this is hardly surprising given the poor performance when compared with $\hat{f}$, of the simple fourth order KDE demonstrated in Chapter 5. We would hope that the more promising $\hat{f}_{JLN}$ estimators would be more competitive. Before considering the performance of the density estimators themselves, however, it is useful to consider the performance of the various scale estimators.

6.4.1 Estimates of Scale

It was alluded to above that it is not always possible to calculate $\hat{\sigma}_{D3}$. The weight of the dominant component is constrained to be between 0 and 1. The estimate, however, can be outside these limits, and in these cases we must revert to the simpler scale estimate $\hat{\sigma}_{D1}$. To see how often this is a problem, consider Table 6.1, which shows, for the number of times from the 1000 simulated datasets for $n = 100$ where either $\hat{\sigma}_{D3}$ could be estimated successfully, where $\hat{w}_1 < 0$ (a type A error), or where $\hat{w}_1 > 1$ (a type B error). The results do not qualitatively change for the $n = 500$ case.

The most obvious conclusion to be drawn from this table is that the actual calculation of $\hat{\sigma}_{D3}$ is frequently not possible. For the case of a simple Gaussian density, Density 1, in over 70% of cases the estimate of curvature

Density	$\hat{\sigma}_{D3}$ Successful	Type A Error	Type B Error
MW(1)	171	716	113
MW(2)	207	685	108
MW(3)	524	375	101
MW(4)	825	104	71
MW(5)	177	720	103
MW(6)	323	550	127
MW(7)	599	283	118
MW(8)	313	583	104
MW(9)	364	544	92
MW(10)	116	818	66

Table 6.1: Observed number of cases from 1000 simulated datasets that $\hat{\sigma}_{D3}$ could not be estimated

at the estimated location of the peak was positive. Similarly, for Density 2 the problem discussed by Janssen et al. of a combination of several large peaks close together also produces a positive estimate of curvature in the majority of simulated datasets. Note that in these cases, however, the sample standard deviation should provide an adequate estimate of σ, as there is really only a single strong mode.

Density 4, however, is clearly a situation in which the sample standard deviation could lead to oversmoothing and inability to resolve the central peak accurately, $\hat{\sigma}_{\mathrm{D3}}$ is estimable in over 80% of cases. Looking at the definition of this density

$$f(x) = \frac{1}{3}\phi_{1/10}(x) + \frac{2}{3}\phi_1(x),$$

we see that the highest peak or 'dominant' component has $w_1 = 1/3$ and $\sigma_1 = 1/10$. Taking only those 825 cases where $\hat{\sigma}_{\mathrm{D3}}$ was successfully calculated, Figures 6.1 and 6.2 respectively show kernel density estimates of the calculated values of w_1 and σ_1. The true values of the estimated quantity are shown by the dotted vertical lines.

Thus we can see that although this method is giving good estimates of w_1, the final estimate of scale is still averaging about double what we would like it to be ideally. This is still, however a considerable improvement upon the sample standard deviation, which in this case had a *mean* of 0.82, and a *minimum* of 0.52.

In general, for Densities 3,4 and 5, all of which have a high, narrow peak, the difference-based scale measures always result in smaller estimates of σ_1 than the sample standard deviation. A similar situation exists for Density 7, with two separated peaks. The sample standard deviation gives estimates of σ_1 based on the whole sample, not just the points around a single peak.

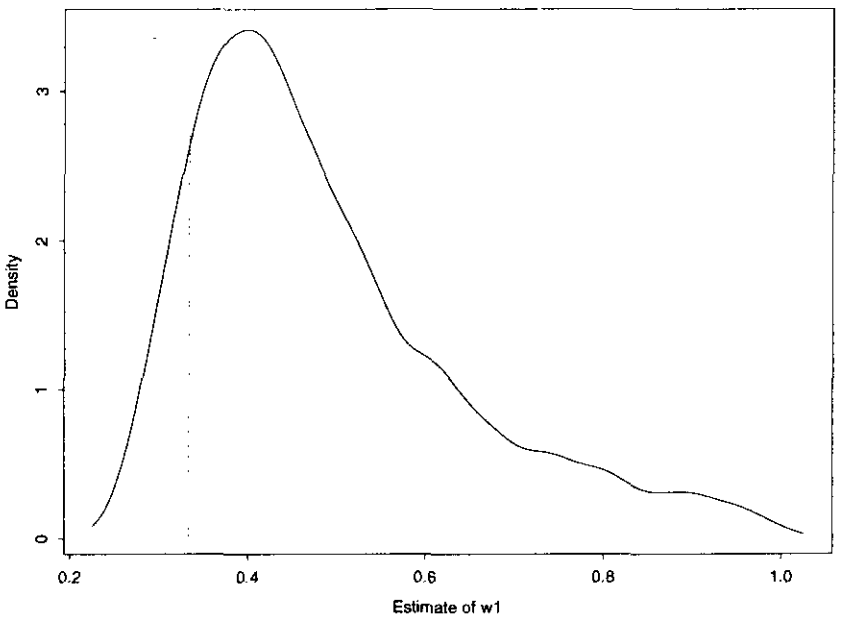

Figure 6.1: Estimated values for w_1 from Density 4, $n = 100$.

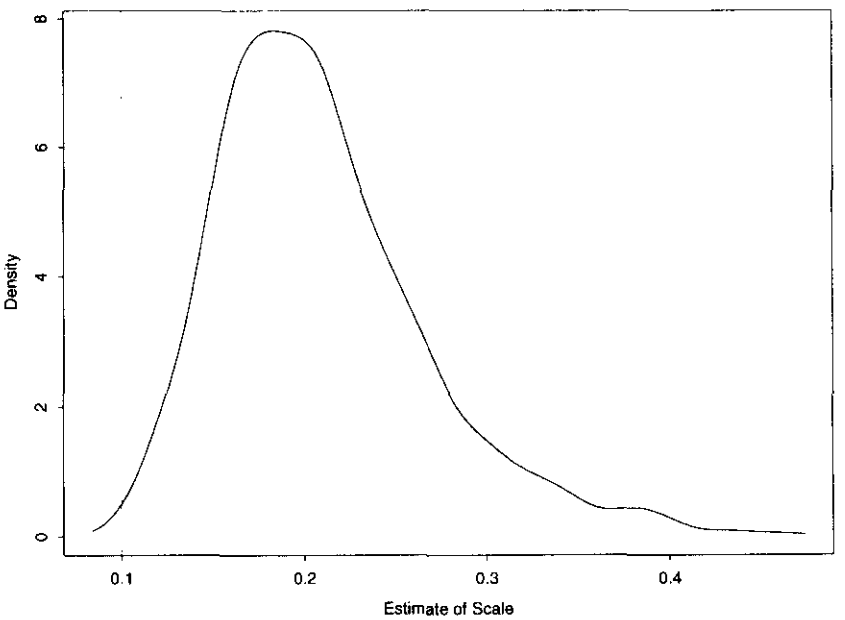

Figure 6.2: Estimated values for σ_1 from Density 4, $n = 100$.

For the other multimodal densities (6, 8, 9 and 10), however, in approximately 60% of cases (for both $n = 100$ and $n = 500$) the sample standard deviation was smaller than $\hat{\sigma}_{\mathrm{D3}}$, suggesting that even 500 data points are insufficient to resolve the small peaks in these cases.

To assess the actual performance of the most complex, and thus hopefully most practically useful scale estimator $\hat{\sigma}_{\mathrm{SS}}$, the true values of σ_1 were derived from the formulae for each of the densities, where the dominant component was the one with the maximum value of w_i/σ_i. For all five multimodal densities, this leads to some confusion, in that there is no single dominant component. Only for Density 7 does this lead to different values of σ. Table 6.2 shows various percentiles of the values of $\hat{\sigma}_{\mathrm{SS}}$ from the simulated datasets, together with the 'true' value.

Clearly, there is little difference for the results in Table 6.2 between the $n = 100$ and $n = 500$ case. Both the median and the spread of estimated values remain more or less the same when the sample size increases.

For the first five (unimodal) densities, the scale estimates are what could be termed 'reasonable', in that they are quite close to the ideal value. For the last five models however, which of course are multimodal, $\hat{\sigma}_{\mathrm{SS}}$ gives inflated estimates of scale, often failing to recognise the dominant component with small variance. This suggests, therefore, that a bandwidth which is too large to resolve the multiple modes present will be used in these cases. However, it is important to note that the same difficulties will also be encountered by other more complex estimators.

Although the performance of the scale estimators is of interest, the real question is whether the use of them in a QAD bandwidth selection procedure can improve the overall accuracy of the density estimation, which is what we now examine.

Density	Quantiles of $\hat{\sigma}_{SS}$					True Value
	10%	25%	50%	75%	90%	
n = 100						
MW(1)	0.736	0.865	0.948	1.012	1.065	1.000
MW(2)	0.489	0.613	0.693	0.767	0.818	0.556
MW(3)	0.198	0.251	0.344	0.432	0.515	0.059
MW(4)	0.152	0.175	0.213	0.269	0.328	0.100
MW(5)	0.081	0.097	0.110	0.121	0.132	0.100
MW(6)	0.821	1.070	1.166	1.218	1.260	0.667
MW(7)	0.545	0.658	0.876	1.069	1.203	0.500
MW(8)	0.666	0.930	1.055	1.108	1.152	0.333 or 1
MW(9)	0.773	1.060	1.232	1.288	1.330	0.600
MW(10)	0.720	0.789	0.843	0.888	0.934	0.100
n = 500						
MW(1)	0.881	0.942	0.977	1.007	1.030	1
MW(2)	0.594	0.661	0.703	0.734	0.763	0.556
MW(3)	0.210	0.235	0.277	0.345	0.408	0.059
MW(4)	0.149	0.158	0.170	0.184	0.199	0.100
MW(5)	0.093	0.103	0.110	0.115	0.119	0.100
MW(6)	0.848	0.965	1.156	1.201	1.225	0.667
MW(7)	0.588	0.638	0.701	0.797	0.922	0.500
MW(8)	0.738	0.899	1.068	1.100	1.120	0.333 or 1
MW(9)	0.906	1.120	1.254	1.280	1.299	0.600
MW(10)	0.816	0.843	0.863	0.885	0.903	0.100

Table 6.2: Quantiles of the 'Super Scale' estimator $\hat{\sigma}_{SS}$.

6.4.2 Density Estimates

Using equations (6.4) and (6.8) with the four estimates of scale $\hat{\sigma}_{\mathrm{ROT}}$, $\hat{\sigma}_{\mathrm{D1}}$, $\hat{\sigma}_{\mathrm{D3}}$ and $\hat{\sigma}_{\mathrm{SS}}$, we can now easily calculate QAD bandwidths for both standard second-order kernel density estimates, and for $\hat{f}_{JLN}$ and $\hat{f}^{R}_{JLN}$. We use the rescaled version of this estimator because of its apparent better performance in the simulations of the previous chapter, but do not amend the formula for $\widehat{h^{*}_{OPT}}$. This is because, when using a Gaussian reference distribution, the additional $O(h^4)$ term introduced by the rescaling is exactly the same as the original term and thus the resulting value of B_f is zero. To avoid expanding these expressions to terms involving h^6, therefore, we have used the formula given by equation (6.8).

Thus we have twelve bandwidth estimators, denoted by $\hat{h}$ for $\hat{f}$ and by $\hat{h}^{*}$ for $\hat{f}_{JLN}$ or $\hat{f}^{R}_{JLN}$, and subscripted by either ROT, D1, D3 or SS depending on which particular measure of scale was applied.

For both samples sizes, $n = 100$ and $n = 500$, the 1000 simulated datasets were used to calculate the actual ISE achieved by each QAD bandwidth selection procedure. These were compared with the minimum achievable ISE-optimal value from the previous chapter.

Finally, we also considered what several authors have suggested is one of the most generally applicable of the more 'hi-tech' estimators, that of Sheather and Jones [17]. This is a plug-in bandwidth selection procedure which starts from equation (6.2), but estimates $R(f'')$ directly rather than substituting a normal density for f. This implies that it is only applicable to the standard case, and not to the more complicated fourth order KDEs. It does, however, allow us to compare the combination of a simple KDE and a complex bandwidth selector ($\hat{h}_{SJ}$) to the combination of a complex KDE and a simple bandwidth selector (e.g. $\hat{h}^{*}_{SS}$).

The mean and standard error of the estimated ISEs over 1000 simulated datasets are presented in Tables 6.3, 6.4, 6.5 and 6.6.

Some of these estimators have previously been compared in the standard KDE case. Janssen et al. [48] consider $\hat{\sigma}_{\mathrm{ROT}}$, $\hat{\sigma}_{\mathrm{D1}}$, $\hat{\sigma}_{\mathrm{D3}}$ and $\hat{\sigma}_{\mathrm{SS}}$, and show that $\hat{h}_{ROT}$ performs approximately as well as $\hat{h}_{SS}$ for all densities except 3, 4 and 7, where the latter estimator is clearly better. Jones, Marron and Sheather [44] assess a large number of sophisticated bandwidth selectors, but include both $\hat{h}_{ROT}$ and $\hat{h}_{SJ}$ in one comparison, suggesting that $\hat{h}_{SJ}$ is amongst the best of these methods, although $\hat{h}_{ROT}$ is nearly as good for definitely bimodal densities such as Models 6, 8 and 9. Jones et al. then go on to compare the same scale estimators as Janssen et al., coming to similar conclusions. They do not, however, directly compare $\hat{h}_{SS}$ with $\hat{h}_{SJ}$, which we do here.

When considering the results of the simulations, the first point to note is that in every single case except the pure Gaussian density the renormalised estimator $\hat{f}^R_{JLN}$ leads to a smaller mean ISE than the unrenormalised version for the same model and bandwidth selector. This confirms that the previous result on the idealised superiority of the renormalisation can in fact be carried through to practice.

Recall that from Chapter 5, on the basis of minimum achievable ISEs, $\hat{f}^R_{JLN}$ was the best estimator for Models 1, 2, 4, 5, 7, 8 and 9, although for the last two the benefit was marginal. Ignoring for the moment differences in performance between the estimates of scale, if we just consider the minimum over all four scale estimators, for $n = 100$ we have that $\hat{f}^R_{JLN}$ is superior for Models 1, 2, 5 and 7. The only difference for $n = 500$ is that $\hat{f}^R_{JLN}$ is superior for Model 4 also. These results parallel the ISE-optimal results, and we can thus conclude that the choice of bandwidth selector is less important than

Estimator	Estimated Mean (SE) ISE				
	MW(1)	MW(2)	MW(3)	MW(4)	MW(5)
$\hat{f}$					
$\hat{h}_{ROT}$	581 (13)	918 (19)	11894 (76)	10417 (126)	5826 (117)
$\hat{h}_{D1}$	568 (14)	927 (20)	7158 (79)	8481 (123)	5946 (123)
$\hat{h}_{D3}$	625 (16)	1016 (24)	6159 (84)	5281 (85)	6341 (136)
$\hat{h}_{SS}$	617 (16)	1011 (24)	6159 (84)	5281 (85)	6341 (136)
Optimal	462 (12)	755 (17)	4227 (53)	4152 (59)	4908 (110)
$\hat{h}_{SJ}$	632 (16)	1019 (23)	6002 (67)	5323 (86)	6357 (134)
$\hat{f}_{JLN}$					
$\hat{h}^*_{ROT}$	475 (12)	745 (18)	15085 (84)	12194 (137)	4515 (102)
$\hat{h}^*_{D1}$	425 (11)	727 (17)	9027 (99)	10004 (134)	4388 (98)
$\hat{h}^*_{D3}$	486 (14)	824 (23)	7526 (108)	5640 (96)	4836 (120)
$\hat{h}^*_{SS}$	490 (14)	821 (23)	7526 (108)	5640 (96)	4836 (120)
Optimal	319 (8)	551 (14)	4369 (61)	3855 (56)	3410 (80)
$\hat{f}^R_{JLN}$					
$\hat{h}^*_{ROT}$	481 (13)	741 (19)	13947 (78)	12044 (135)	4506 (111)
$\hat{h}^*_{D1}$	405 (12)	697 (18)	8662 (95)	9852 (132)	4067 (101)
$\hat{h}^*_{D3}$	473 (15)	806 (24)	7353 (104)	5627 (95)	4624 (128)
$\hat{h}^*_{SS}$	486 (15)	806 (24)	7353 (104)	5627 (95)	4624 (128)
Optimal	219 (7)	477 (13)	4470 (61)	3882 (55)	2701 (71)

Table 6.3: Estimated mean and standard error of ISE over 1000 simulations for QAD and Sheather-Jones bandwidth estimators, compared to minimum achievable ISE. Models 1 to 5, n=100.

Estimator	Estimated Mean (SE) ISE				
	MW(6)	MW(7)	MW(8)	MW(9)	MW(10)
$\hat{f}$					
$\hat{h}_{ROT}$	819 (12)	4226 (16)	1149 (13)	1073 (12)	5251 (16)
$\hat{h}_{D1}$	1080 (10)	2501 (30)	1396 (13)	1402 (10)	5404 (14)
$\hat{h}_{D3}$	1017 (13)	1798 (30)	1326 (15)	1302 (13)	5415 (16)
$\hat{h}_{SS}$	895 (13)	1798 (30)	1220 (15)	1146 (13)	5332 (16)
Optimal	717 (13)	1053 (19)	934 (15)	864 (13)	3652 (36)
$\hat{h}_{SJ}$	836 (14)	1188 (20)	1119 (16)	1031 (14)	5374 (18)
$\hat{f}_{JLN}$					
$\hat{h}^*_{ROT}$	913 (13)	5861 (11)	1400 (14)	1205 (12)	5408 (15)
$\hat{h}^*_{D1}$	1444 (11)	3004 (38)	1730 (14)	1856 (12)	5512 (13)
$\hat{h}^*_{D3}$	1279 (15)	1876 (37)	1598 (18)	1633 (18)	5517 (15)
$\hat{h}^*_{SS}$	1042 (14)	1876 (37)	1464 (17)	1331 (14)	5444 (16)
Optimal	664 (14)	841 (16)	930 (16)	827 (13)	3684 (37)
$\hat{f}^R_{JLN}$					
$\hat{h}^*_{ROT}$	883 (15)	4755 (14)	1383 (16)	1153 (15)	5410 (17)
$\hat{h}^*_{D1}$	1397 (13)	2076 (36)	1692 (15)	1790 (15)	5478 (15)
$\hat{h}^*_{D3}$	1249 (17)	1375 (30)	1578 (19)	1596 (19)	5489 (17)
$\hat{h}^*_{SS}$	1018 (16)	1375 (30)	1455 (18)	1296 (16)	5435 (18)
Optimal	658 (15)	711 (16)	924 (16)	813 (13)	3754 (36)

Table 6.4: Estimated mean and standard error of ISE over 1000 simulations for QAD and Sheather-Jones bandwidth estimators, compared to minimum achievable ISE. Models 6 to 10, n=100.

Estimator	Estimated Mean (SE) ISE				
	MW(1)	MW(2)	MW(3)	MW(4)	MW(5)
$\hat{f}$					
$\hat{h}_{ROT}$	179 (4)	264 (5)	8911 (32)	5272 (56)	1730 (33)
$\hat{h}_{D1}$	174 (4)	265 (5)	3944 (29)	3358 (44)	1764 (35)
$\hat{h}_{D3}$	180 (4)	275 (5)	2538 (38)	1297 (17)	1804 (36)
$\hat{h}_{SS}$	178 (4)	275 (5)	2538 (38)	1297 (17)	1804 (36)
Optimal	154 (3)	234 (5)	1345 (15)	1193 (16)	1542 (32)
$\hat{h}_{SJ}$	183 (4)	276 (5)	2093 (22)	1362 (20)	1787 (35)
$\hat{f}_{JLN}$					
$\hat{h}^*_{ROT}$	123 (3)	186 (4)	13146 (39)	8555 (75)	1151 (24)
$\hat{h}^*_{D1}$	111 (3)	186 (4)	6276 (41)	5567 (62)	1132 (24)
$\hat{h}^*_{D3}$	116 (3)	194 (4)	3877 (59)	1213 (19)	1184 (25)
$\hat{h}^*_{SS}$	117 (3)	194 (4)	3877 (59)	1213 (19)	1184 (25)
Optimal	93 (2)	157 (4)	1317 (16)	1007 (13)	965 (21)
$\hat{f}^R_{JLN}$					
$\hat{h}^*_{ROT}$	121 (3)	178 (4)	12143 (36)	8357 (74)	1091 (24)
$\hat{h}^*_{D1}$	102 (3)	171 (4)	5929 (40)	5374 (61)	989 (22)
$\hat{h}^*_{D3}$	108 (3)	181 (4)	3699 (56)	1182 (18)	1061 (24)
$\hat{h}^*_{SS}$	110 (3)	181 (4)	3699 (56)	1182 (18)	1061 (24)
Optimal	58 (2)	135 (4)	1319 (16)	994 (13)	737 (18)

Table 6.5: Estimated mean and standard error of ISE over 1000 simulations for QAD and Sheather-Jones bandwidth estimators, compared to minimum achievable ISE. Models 1 to 5, n=500.

Estimator	Estimated Mean (SE) ISE				
	MW(6)	MW(7)	MW(8)	MW(9)	MW(10)
$\hat{f}$					
$\hat{h}_{ROT}$	275 (4)	1757 (7)	449 (5)	447 (4)	4831 (4)
$\hat{h}_{D1}$	447 (4)	964 (9)	650 (5)	716 (4)	4746 (4)
$\hat{h}_{D3}$	348 (6)	413 (8)	562 (7)	616 (7)	4752 (4)
$\hat{h}_{SS}$	295 (5)	413 (8)	485 (6)	487 (5)	4777 (4)
Optimal	223 (4)	313 (5)	299 (5)	284 (4)	1110 (11)
$\hat{h}_{SJ}$	250 (4)	339 (5)	345 (5)	333 (4)	1362 (12)
$\hat{f}_{JLN}$					
$\hat{h}^*_{ROT}$	339 (4)	3311 (5)	742 (5)	620 (4)	4817 (4)
$\hat{h}^*_{D1}$	837 (5)	1506 (12)	1149 (7)	1211 (5)	5012 (4)
$\hat{h}^*_{D3}$	528 (10)	373 (10)	932 (13)	982 (13)	4994 (4)
$\hat{h}^*_{SS}$	382 (6)	373 (10)	790 (10)	697 (7)	4943 (5)
Optimal	177 (4)	220 (4)	270 (5)	269 (4)	996 (11)
$\hat{f}^R_{JLN}$					
$\hat{h}^*_{ROT}$	308 (5)	2146 (6)	717 (6)	578 (5)	4808 (5)
$\hat{h}^*_{D1}$	778 (6)	744 (9)	1111 (7)	1132 (6)	4986 (4)
$\hat{h}^*_{D3}$	492 (10)	233 (6)	906 (13)	927 (13)	4969 (5)
$\hat{h}^*_{SS}$	352 (6)	233 (6)	768 (10)	653 (7)	4929 (5)
Optimal	171 (4)	169 (4)	269 (5)	268 (4)	994 (11)

Table 6.6: Estimated mean and standard error of ISE over 1000 simulations for QAD and Sheather-Jones bandwidth estimators, compared to minimum achievable ISE. Models 6 to 10, n=500.

the choice of estimator.

When looking in detail at the various scale estimators, the picture is less clear. For Densities 1, 2 and 5, when $\hat{f}^R_{JLN}$ is superior, it would seem that $\hat{h}^*_{D1}$ is best. Note however that these three models are those where it was impossible to calculate $\hat{\sigma}_{\mathrm{D3}}$ in approximately 70% of cases, suggesting that most of the poorer performance of $\hat{h}^*_{SS}$ in these models was due to values of s being used, although the ISEs achieved were still an improvement over the standard KDEs. For the separated bimodal Density 7 and the strongly skewed Density 3, however, which are exactly the situations that $\hat{\sigma}_{\mathrm{SS}}$ was designed to cope with, the use of $\hat{\sigma}_{\mathrm{SS}}$ gave clear benefits for all estimators.

For the multimodal densities 6, 8, 9 and 10, it is interesting to note that the original suggestion of $\hat{h}_{ROT}$ gives such good performance, even when compared to a 'state-of-the-art' method $\hat{h}_{SJ}$. It also seems that for Model 10 in most cases the choice of estimator or bandwidth selector is irrelevant; 500 sample points are not enough to resolve the very fine structure apparent in this density. The clear exception to this rule is $\hat{h}_{SJ}$, in the $n = 500$ case, however, with greatly improved performance over all other methods considered.

Comparing the actual ISE values obtained to their ISE-optimal minimum value can also provide some insights. In the cases where $\hat{f}^R_{JLN}$ works (Densities 1, 2, 5 and 7) the ISEs achieved are close to and often less than the best that can be attained with a second order KDE, especially for the larger sample size.

Can we draw some useful practical conclusions from these simulation results? Given the extensive empirical evidence that $\hat{h}_{SJ}$ is amongst the best of the complex bandwidth selectors, the matching performance of $\hat{h}_{SS}$ across nearly all models (except Density 7) suggests that the latter would be

a very good place to start. If the density however is close to normal (Models 1 and 2) or has very heavy tails (Models 4 and 5) then real practical gains can be made by using $\hat{f}^R_{JLN}$ and $\hat{h}^*_{SS}$. In both cases, the value of the scale estimator $\hat{\sigma}_{\mathrm{SS}}$ is clear.

6.5 Conclusions

In this chapter we have studied in detail several proposed scale estimators. These form the basis for some very simple bandwidth selection procedures for standard kernel density estimation, and these methods have previously been studied. It is relatively simple to extend these methods to provide analogous procedures for use with the higher order kernel density estimators.

An extensive simulation experiment has shown that the superiority of these higher-order KDEs when used with an ideal bandwidth can be translated into improved practical performance with very simple bandwidth selection, at least for unimodal densities.

The most interesting result of this chapter, however, is the fact that in practice the gain is greater by using a more complex estimator than by using a more complex bandwidth selector. This has obvious implications for future research in the area of density estimation, and for smoothing in general.

Part III

Poisson Regression

Chapter 7

Power and Sample Size for Poisson Regression

7.1 Introduction

Logistic regression modelling is a well-established statistical technique for analysing relationships between binary outcomes (e.g. alive/dead, yes/no) and a set of (possibly multivariate) covariates. The technique is particularly prevalent in the areas of biometrics, epidemiology and medical statistics.

Unfortunately, all too often in practice, data which can be collected as counts, e.g. number of migraine attacks per month, number of moths of a particular species collected per hour, are summarised into either presence or absence, and then analysed accordingly, typically with logistic regression. This obviously entails a loss of information, since all counts greater than zero are pooled. For such outcomes [49, 50] however, Poisson regression can be used to give results which are superior in terms of power and sample size. The following work allows us in some sense to quantify this loss of information.

In this chapter we discuss the calculation of power and sample size for both logistic and Poisson regression models, using asymptotic techniques based upon the Fisher information matrix, and demonstrate that substantial savings in sample size, or conversely, gains in power, can be extracted from the uncondensed data. Some of the theoretical development in this chapter was published in 1991 [51].

7.2 Asymptotic Theory

Suppose we have N individuals, each observed for a, possibly constant, 'exposure time' t_i, $(i = 1, \ldots, N)$. Let Y_i be the Poisson distributed response, and x_i the corresponding covariate p-vector. The natural parameterisation for a standard Poisson regression model defines the rate of events λ_i of the ith individual as

$$\lambda_i = \exp(\beta_0 + \beta^T x_i),$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$. Thus, assuming time-homogeneity, the expected value of Y_i will be

$$E(Y_i) = t_i \lambda_i.$$

Now, consider both x_i and t_i as realisations of random variables X and T, with probability density functions $f_X(x)$ and $f_T(t)$ respectively. If we assume that the exposure time t_i is independent of x_i for each individual, then the likelihood function of a sample from the joint distribution of Y, T and X will be

$$L(\beta_0, \beta) = \prod_{i=1}^{n} f_X(x_i) f_T(t_i) (t_i \lambda_i)^{y_i} \exp(-t_i \lambda_i)/y_i!. \tag{7.1}$$

Consider the maximum likelihood estimators of β_0 and β in equation (7.1). As N increases, standard asymptotic theory [3] can be used to show

that these converge in distribution to a multivariate normal distribution, mean $(\beta_0, \beta^T)^T = \beta^*$ and variance-covariance matrix I^{-1}, where I is the Fisher information matrix with elements

$$I_{[jk]} = -E\left(\frac{\partial^2 \log L}{\partial\beta_j \partial\beta_k}\right) \qquad (j,k = 0,\ldots,p),$$

where $I_{[jk]}$ denotes the element of I in row j and column k. So, from equation (7.1) we have

$$\log L(\beta_0,\beta) = \sum_{i=1}^{n} \left\{ y_i(\beta_0 + \beta^T x_i) - t_i \exp(\beta_0 + \beta^T x_i) + F(x_i,t_i,y_i) \right\},$$

where $F(x,t,y)$ is independent of $\beta_0,\ldots,\beta_p$. Thus, when differentiated twice with respect to β, only the second term contributes, and we have,

$$I_{[jk]} = N\ E[T X_j X_k \exp(\beta_0 + \beta^T X)] \qquad (j,k = 0,\ldots,p),$$

where $X_j,\ \ j = 1,\ldots,p$ are the elements of X and we define $X_0 = 1$.

But, by the independence of T and X,

$$I_{[jk]} = N\ \mu_T\ e^{\beta_0} E[X_j X_k \exp(\beta^T X)], \tag{7.2}$$

where μ_T is the mean exposure time.

This development is very similar to that of Whittemore [52] who used this technique to estimate power and sample size for logistic regression. Unfortunately those results rely upon an approximation which is only valid when the underlying success probability is small. This approximation is unnecessary here.

Define the moment generating function (MGF) of the covariates X by $m(s) = E\left[\exp(s^T X)\right]$. Define the first and second partial derivatives of this function as

$$m_i = \frac{\partial m}{\partial s_i},\ m_{ij} = \frac{\partial^2 m}{\partial s_i \partial s_j},\ (i,j = 1,\ldots,p).$$

Let $m_0 = m_{00} = m$, (the 'zeroth' derivatives), and similarly $m_{i0} = m_{0i} = m_i$. Now form the $(p+1)$ by $(p+1)$ matrix $\mathbf{M}(s) = (m_{ij}), \quad i,j = 0,\ldots,p.$

Thus, we can now express equation (7.2) as

$$I_{[jk]} = N\,\mu_T\,\exp(\beta_0)\,M(\beta)_{[jk]}.$$

Hence the maximum likelihood estimate $\hat{\beta}^*$ of β^* is asymptotically, as $N \to \infty$, multivariate normal with mean β^* and covariance matrix

$$(N\,\mu_T)^{-1}\,\exp(-\beta_0)\,M^{-1}(\beta). \tag{7.3}$$

So, assuming that the parameter of interest (for example the contrast between two treatment effects) is β_1, suppose we wish to test the null hypothesis $H_0 : \beta = \beta_N = (0, \beta_2, \ldots, \beta_p)$ against the alternative hypothesis $H_1 : \beta = \beta_A = (\tilde{\beta}, \beta_2, \ldots, \beta_p)$ for $\tilde{\beta} > 0$. Set the significance level to be α and the power to be at least $1-\gamma$. Assuming N is large enough to apply the asymptotic results derived above, the asymptotic variance of the maximum likelihood estimator $\hat{\beta}_1$ is then given by the second diagonal term of I^{-1}, giving the Wald statistic $\hat{\beta}_1 / \hat{I}^{-1}_{[11]}$.

Begin therefore from the requirement that

$$Pr[\text{Reject}H_0 \mid H_1\text{true}] \geq 1-\gamma,$$

and define Z_δ such that $Pr[X < Z_\delta] = 1-\delta$, where X follows a standard Gaussian distribution. Then we require

$$Pr\left[\frac{\hat{\beta}_1}{\sqrt{\text{var}_N(\hat{\beta}_1)}} \leq Z_\alpha \mid H_1\text{true}\right] \leq \gamma$$

where $\text{var}_N(\hat{\beta}_1)$ is the asymptotic variance of $\hat{\beta}_1$ under the null hypothesis. Similarly, let $\text{var}_A(\hat{\beta}_1)$ be the variance under the alternative hypothesis, then some basic algebraic manipulation gives

$$Pr\left[\frac{\hat{\beta}_1 - \tilde{\beta}}{\sqrt{\text{var}_A(\hat{\beta}_1)}} \leq \frac{Z_\alpha\sqrt{\text{var}_N(\hat{\beta}_1)} - \tilde{\beta}}{\sqrt{\text{var}_A(\hat{\beta}_1)}} \mid H_1\text{true}\right] \leq \gamma \tag{7.4}$$

which implies that

$$\frac{Z_\alpha\sqrt{\text{var}_N(\hat{\beta}_1)} - \tilde{\beta}}{\sqrt{\text{var}_A(\hat{\beta}_1)}} \leq Z_{1-\gamma} = -Z_\gamma.$$

But equation (7.3) can be used to calculate the variances of the ML estimate $\hat{\beta}_1$ under both null and alternative hypotheses, and thus it can be shown that to achieve the required power, we need

$$N\mu_T e^{\beta_0} \geq \left[Z_\alpha V^{1/2}(\beta_N) + Z_\gamma V^{1/2}(\beta_A)\right]^2 / \tilde{\beta}^2 \tag{7.5}$$

where $V(\beta) = \{M^{-1}(\beta)\}_{[22]}$, the second diagonal term of M^{-1} evaluated at β. This can be easily generalised to the case where H_0 involves a non-zero value of β_1.

Alternatively, given a value of $N\mu_T e^{\beta_0}$, the power of the test is calculated by a simple rearrangement of (7.4) to give

$$\text{Power} = 1 - \gamma = 1 - \Phi\left(\frac{Z_\alpha V^{1/2}(\beta_N) - \tilde{\beta}\sqrt{N\mu_T e^{\beta_0}}}{V^{1/2}(\beta_A)}\right) \tag{7.6}$$

where Φ is the cumulative distribution function of the standard Gaussian.

Similarly, it can be shown that for a two-sided hypothesis test, the power of the test becomes

$$\begin{aligned}\text{Power} \quad = \quad & 1 - \Phi\left(\frac{Z_{\alpha/2} V^{1/2}(\beta_N) - \tilde{\beta}\sqrt{N\mu_T e^{\beta_0}}}{V^{1/2}(\beta_A)}\right) \\ & + \Phi\left(\frac{-Z_{\alpha/2} V^{1/2}(\beta_N) - \tilde{\beta}\sqrt{N\mu_T e^{\beta_0}}}{V^{1/2}(\beta_A)}\right),\end{aligned}$$

If we assume however, without loss of generality, that $\tilde{\beta} > 0$, then the third term in the above equation may be considered negligible for $N\mu_T e^{\beta_0}$ sufficiently large. Thus, the power of a two-sided test may be approximated by that for a one-tailed test at half the size.

So, as always in sample size calculations, by specifying the null and alternative hypotheses (which determine the expected difference between

groups), the required significance and power and, crucially, the distribution of the covariates X, we can readily calculate the minimum sample size required to reject the null when it is in fact false.

7.3 Over-dispersion

A common problem encountered when assessing the fit of a Poisson model to data is the phenomenon of over-dispersion [53, 54]. This can arise in several ways, as discussed by McCullagh and Nelder [3] (pp.199-200). It may be simply parameterised through the relationship between the mean and the variance such that

$$\text{var}(Y_i) = \sigma^2 E(Y_i),$$

with $\sigma^2 > 1$. In this case the maximum likelihood parameter estimates are identical, but the variance-covariance matrix becomes $\sigma^2 I^{-1}$. Thus the calculated sample size should be increased by a factor of σ^2, which must be estimated prior to the study.

More complicated scenarios, in which the mechanism of over-dispersion can be modelled in some way, such as by using a gamma-distributed random effect or by more explicit means, are beyond the scope of this work, and there is some evidence to suggest that they are, in cases where the over-dispersion is not great, unnecessary; see Yanez and Wilson [54] for details. Moreover, the various approximations and assumptions necessary to the implementation of prospective sample size calculations render any resulting figure indicative only and very detailed modelling techniques which modify these figures only slightly are, in the view of this author, unnecessary.

7.4 The Univariate Case

For a single covariate, $p = 1$, $\mathbf{M}$ is a 2×2 matrix and the second diagonal term of $\mathbf{M}^{-1}$ becomes

$$V(\beta) = m(\beta)/[m(\beta)m''(\beta) - m'(\beta)^2]. \tag{7.7}$$

This is especially useful in testing the simple hypothesis $H_N : \beta_1 = 0$ against $H_A : \beta_1 = \tilde{\beta} > 0$, since then $m(0) = 1, m'(0) = E(X)$ and $m''(0) = E(X^2)$. Thus equation (7.5) becomes

$$N\mu_T e^{\beta_0} \geq \left[Z_\alpha(\text{var}(X)^{-1/2}) + Z_\gamma V^{1/2}(\tilde{\beta})\right]^2 / \tilde{\beta}^2. \tag{7.8}$$

Certain intuitively obvious factors related to the covariate can be seen at work in this equation. The greater the expected size of the effect (as measured by $\tilde{\beta}$), the smaller the required sample size, although the simple inverse quadratic relationship is modified by the presence of $V^{1/2}(\tilde{\beta})$ in the numerator. Similarly, the greater the variability of the covariate X, again the smaller the sample size required.

Take the simplest case, that of comparing two homogenous groups of size n_1 and n_2, which can be parameterised by defining a Bernoulli covariate X such that $P[X = 1] = \pi$, naturally estimated by $n_1/(n_1 + n_2)$.

Let $\exp(\tilde{\beta})$ be the rate ratio for the presence versus the absence of the study factor. Thus, for $X = 0$, the control group, the rate of events is equal to $\lambda_0 = \exp(\beta_0)$, and for $X = 1$ it is equal to $\lambda_1 = \exp(\tilde{\beta})\lambda_0 = \exp(\beta_0 + \tilde{\beta})$.

Simple calculations allow us to derive the MGF of this distribution, which is $(1 - \pi) + \pi e^t$, and hence show that

$$V(\beta) = (\pi \exp(\beta))^{-1} + (1 - \pi)^{-1}.$$

Substituting this into equation (7.8) and performing some algebraic manip-

ulation gives the inequality

$$N\mu_T e^{\beta_0} \geq \frac{1}{\tilde{\beta}^2 \pi(1-\pi)} \left[Z_\alpha + Z_\gamma \sqrt{e^{-\tilde{\beta}}(1-\pi) + \pi} \right]^2 . \tag{7.9}$$

Similar calculations can be used to show that in this case equation (7.6) becomes

$$\text{Power} = 1 - \Phi \left(\frac{Z_\alpha - \tilde{\beta}\sqrt{N\mu_T e^{\beta_0}\pi(1-\pi)}}{\sqrt{e^{-\tilde{\beta}}(1-\pi) + \pi}} \right) .$$

For a fixed significance level $\alpha = 0.05$, constant exposure time $\mu_T = 1$ and baseline event rate $e^{\beta_0} = 1$, Figure 7.1 shows how the power achieved varies with $\tilde{\beta}$ for several values of N and π. Clearly the power is greatest (for fixed N) when the groups are of equal size, as this corresponds to the maximum possible variance of X.

Moment generating functions are easily calculated for many other common distributions, both discrete and continuous. Table 7.1 shows the func-

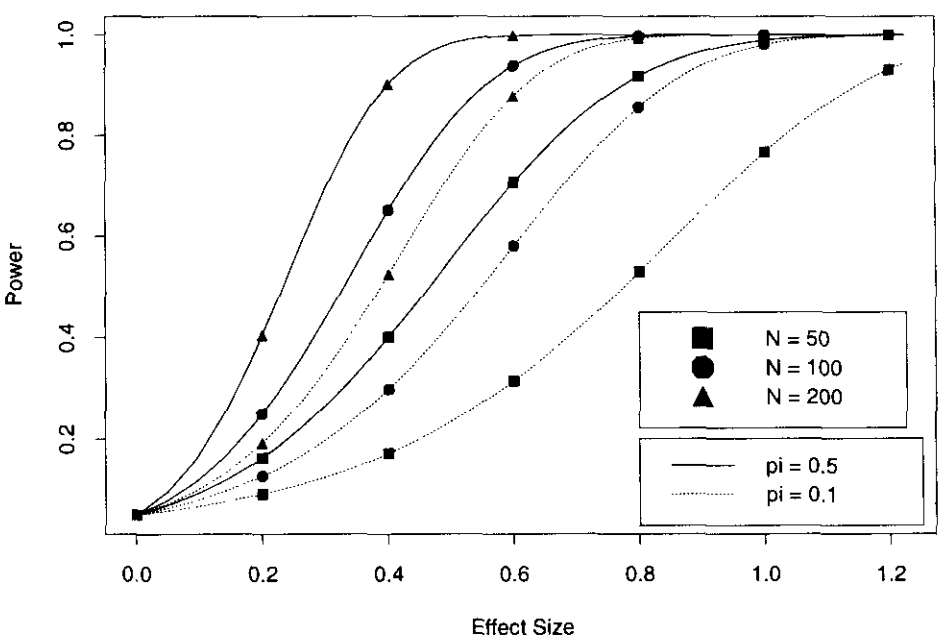

Figure 7.1: Power functions for a Bernoulli covariate, $N = 50, 100$ and 200

tion $V(\beta)$ for several of these. The final column shows the value for a standardised covariate; that is one with the mean value subtracted and divided by the standard deviation. This allows direct comparison between differing covariate distributions. The second last column tabulates the value of $V(\beta)$ for the untransformed variables.

Let R denote the ratio of λ under H_A to λ under H_N; that is $R(x) = \exp(\tilde{\beta}x)$. Thus for standardised distributions having mean zero and variance one, $\exp(\tilde{\beta})$ is the rate ratio for a value of X one standard deviation above the mean. For such standardised covariates, equation (7.6) gives

$$\text{Power} = 1 - \Phi\left(\frac{Z_\alpha - \tilde{\beta}\sqrt{N\mu_T e^{\beta_0}}}{\sqrt{V(\tilde{\beta})}}\right).$$

This power function is plotted for the various distributions given in Table 7.1 in Figure 7.2. Note that to achieve 50% power, $Z_\gamma = 0$ and from (7.8) we have that the minimum sample size required is dependent on the covariate

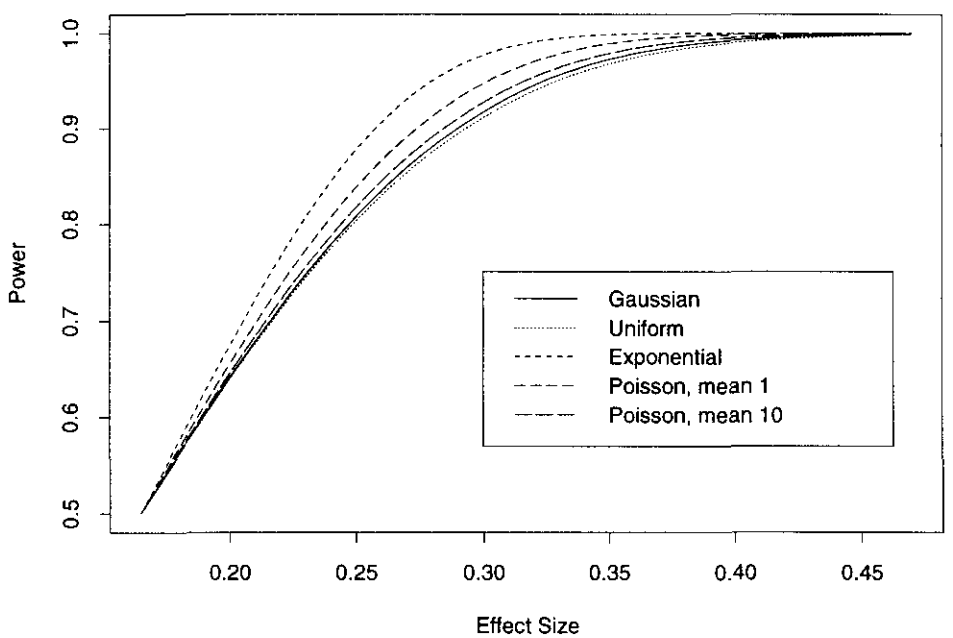

Figure 7.2: Comparison of $V(t)$ for standardised covariate disributions

Distribution	Mean (Var)	$V(t)$	Standardised $V(t)$
Bernoulli	π	$(\pi e^t)^{-1} + (1-\pi)^{-1}$	—
Poisson	λ	$\lambda^{-1}\exp(\lambda - \lambda e^t - t)$	$\exp\left(\lambda + \sqrt{\lambda}t - \frac{1}{\sqrt{\lambda}}t - \lambda e^{t/\sqrt{\lambda}}\right)$
Exponential	λ^{-1}	$(\lambda - t)^3/\lambda \quad (t < \lambda)$	$(1-t)^3 e^t \quad (t < 1)$
Gaussian	$\mu \quad (\sigma^2)$	$\frac{1}{\sigma^2}\exp\left(-\mu t - \frac{\sigma^2 t^2}{2}\right)$	$\exp(-t^2/2)$
Uniform $[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)t^3}{(e^{bt}-e^{at}) - \frac{t^2(a-b)^2 e^{(a+b)t}}{e^{bt}-e^{at}}}$	$\frac{\sqrt{3}t^3\sinh(\sqrt{3}t)}{\sinh^2(\sqrt{3}t) - 3t^2}$

Table 7.1: Values of $V(\beta)$ for raw and standardised covariate distributions

distribution only through its variance, thus

$$N\mu_T e^{\beta_0} \geq \frac{Z_\alpha^2}{\tilde{\beta}^2 \mathrm{var}(X)}.$$

The range of power and effect size has been focused upon the region most likely to apply in practice i.e. high power and moderate effect size, and it can clearly be seen that non-normality of the covariate can have an effect upon the power of a study. However, the differences evident in estimated power are never greater than 10%. In any practical situation, necessary simplifying assumptions will have been made to allow such calculations to take place, and the margin for error introduced by these assumptions imply that such estimates should not be considered definitive. The careful consideration of covariate distributions is a topic which is pursued in the next chapter.

7.5 Simulation Experiment

To test the validity of this asymptotic method, a simulation experiment was used to generate covariates and outcomes in a variety of situations, and empirical estimates of the power were compared to the calculated values.

Beginning once again with the Bernoulli covariate, mean π, i.e. two homogenous groups, for a given sample size N, baseline rate e^{β_0}, and rate ratio $R = e^{\tilde{\beta}}$, the group sizes n_1 and n_2 were generated by sampling from an appropriate binomial distibution. Given n_1 and n_2, and using the homogeneity of the groups, two Poisson distributed values were drawn from distributions with means $n_1\, e^{\beta_0}$ and $R\, n_2\, e^{\beta_0}$ respectively. These sums are sufficient statistics and allow the Wald test statistic to be calculated. This process was repeated 10000 times and the empirical power was calculated by counting the proportion of samples for which this Wald statistic was greater than the critical value of $\Phi^{-1}(0.95)$, corresponding to a significance

Rate Ratio	Nominal Power	$\pi = 0.1$	$\pi = 0.5$	$\pi = 0.9$
1.2	70	74 (0.45)	72 (0.45)	72 (0.44)
	80	82 (0.38)	82 (0.38)	82 (0.38)
	90	91 (0.29)	91 (0.28)	92 (0.28)
1.4	70	76 (0.43)	74 (0.44)	74 (0.44)
	80	84 (0.38)	83 (0.37)	84 (0.37)
	90	91 (0.28)	92 (0.26)	93 (0.26)
1.6	70	77 (0.42)	77 (0.42)	74 (0.43)
	80	85 (0.36)	86 (0.35)	85 (0.35)
	90	92 (0.28)	93 (0.24)	94 (0.24)

Table 7.2: Estimated power (%) for the two group comparison problem (with standard errors in parentheses).

level of 5%. Table 7.2 shows the estimated power for several rate ratios and sample sizes which give rise to the nominal power values in the second column. There appears to be a clear trend towards mildly conservative under-estimation of the true power of the test at all rate ratios and study factor prevalences.

This experiment was repeated with other covariate distributions, as shown in Table 7.3. The four distributions considered are the exponential with parameter 1, the standard Gaussian, the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$, and the Poisson with parameter 1.

Both the Gaussian and uniform, i.e. the symmetric distributions show a very good degree of agreement with the nominal power, with only one result for the uniform case being significantly different (greater than 2 standard errors) from that estimated (rate ratio 1.6, power 80%). The Gaussian appears to slightly over-estimate power, with a trend to decreasing accuracy

Rate Ratio	Nominal Power	Exponential	Gaussian	Uniform	Poisson, $\lambda = 1$
1.2	70	78.2 (0.41)	69.7 (0.46)	69.6 (0.46)	77.9 (0.41)
	80	85.9 (0.35)	79.8 (0.40)	79.9 (0.40)	85.6 (0.35)
	90	91.7 (0.28)	89.9 (0.30)	90.4 (0.29)	92.7 (0.26)
1.4	70	78.8 (0.41)	68.8 (0.46)	69.3 (0.46)	80.8 (0.39)
	80	83.2 (0.37)	79.0 (0.41)	80.2 (0.40)	86.4 (0.34)
	90	89.1 (0.31)	88.6 (0.32)	89.9 (0.30)	92.4 (0.26)
1.6	70	72.6 (0.45)	67.2 (0.47)	69.4 (0.46)	78.8 (0.41)
	80	78.7 (0.41)	77.5 (0.42)	78.5 (0.41)	85.5 (0.35)
	90	83.4 (0.37)	88.2 (0.32)	89.7 (0.30)	92.2 (0.27)

Table 7.3: Estimated power (%) for common covariate distributions (with standard errors in parentheses).

as the effect size increases.

For the skewed distributions (exponential and Poisson) the method appears to be less accurate. The Poisson experiment shows clear underestimations of the power, although this does decrease at the higher power. The exponential is unusual in that for the small treatment effect sizes it is conservative with the reverse being true for a large effect.

Overall, it would seem that this method of calculating power is reasonably accurate, with the possible exception of a highly skewed covariate distribution and a large effect size. This result is less surprising when we realise that large effect sizes correspond to small sample sizes with the result that the asymptotic results may not hold. Table 7.4 shows the sample sizes used in the simulation experiment for the largest rate ratio. These figures explain the relatively poor performance of the method for the exponential and Poisson distribution, but perhaps the most startling thing about them is the very low values of N for which the asymptotic approximation is valid.

Nominal Power	Exponential	Gaussian	Uniform	Poisson, $\lambda = 1$
70	15	21	22	17
80	18	27	28	21
90	21	37	39	26

Table 7.4: Calculated sample sizes for simulation experiment, RR=1.6

7.6 The Multivariate Case

Recalling that the original derivation of the inequality in equation (7.5) did not specifically refer to a univariate distribution of X, the extension to the

multivariate case is simple.

Consider the special case where X has a multivariate exponential family distribution of s dimensions, as defined by Barndorff-Nielsen [55] (p.139), with density

$$f(x,\theta) = h(x)\exp\left\{x^T\theta - q(\theta)\right\},$$

and moment generating function

$$m(t) = \exp\left\{q(\theta+t) - q(\theta)\right\},$$

where $q(\theta)$ is a bounded analytic function of θ independent of X.

As in Section 7.2, let M be the $(p+1)\times(p+1)$ partitioned matrix

$$M = \begin{pmatrix} m(t) & m^{(1)}(t)^T \\ m^{(1)}(t) & m^{(2)}(t) \end{pmatrix}$$

where $m^{(1)}$ denotes the p-vector of first partial derivatives of m, and $m^{(2)}$ denotes the $p\times p$ matrix of second partial derivatives. Then, since

$$\begin{aligned} m^{(1)}(t) &= m(t)\ q^{(1)}(\theta+t), \\ m^{(2)}(t) &= m(t)\left[q^{(1)}(\theta+t)\ q^{(1)}(\theta+t)^T + q^{(2)}(\theta+t)\right], \end{aligned}$$

and using standard results for the inverse of partitioned matrices (e.g Mardia, Kent and Bibby [56], p.459) we can show that

$$M^{-1} = e^{q(\theta)-q(\theta+t)}\begin{pmatrix} 1+q^{(1)\ T}q^{(2)\ -1}q^{(1)}(\theta+t) & -q^{(1)\ T}q^{(2)\ -1}(\theta+t) \\ -q^{(2)\ -1}q^{(1)}(\theta+t) & q^{(2)\ -1} \end{pmatrix}$$

and hence that

$$v(\beta) = \exp\left\{q(\theta) - q(\theta+\beta)\right\}\left\{q^{(2)}(\theta+\beta)\right\}_{11}^{-1}, \tag{7.10}$$

where $\left\{q^{(2)}(\theta+\beta)\right\}_{11}^{-1}$ is the first diagonal term of $q^{(2)\ -1}(\theta+\beta)$. This development follows almost exactly that of Theorem 1 in Whittemore [52], and equation (7.10) is the natural multivariate analogue of (7.7).

To illustrate this derivation, consider the case of the multivariate normal distribution with mean μ and positive definite covariance Σ. For this distribution, the MGF is

$$m(t) = \exp\left(t^T\mu + t^T\Sigma t/2\right)$$

and so we have $q(\theta) = \theta^T\Sigma\theta/2$ where the vector of parameters θ is equal to $\Sigma^{-1}\mu$. Thus in this case $q^{(2)} = \Sigma$ and

$$v(\beta) = \exp\left(-\beta^T\mu - \beta^T\Sigma\beta/2\right)\Sigma^{-1}_{[11]}.$$

However, it can be easily be shown (Mardia et al.[56], p.182) that $\Sigma^{-1}_{[11]} = \left[\text{var}(X_1)(1-\rho^2_{1.2\ldots p})\right]^{-1}$ where $\rho_{1.2\ldots s}$ is the multiple correlation coefficient relating X_1 to $X_2 \ldots X_s$, (see Kleinbaum, Kupper and Muller pp.146-149 [57]).

Thus, for multivariate normal covariates, we have

$$v(\beta) = \frac{(1-\rho_{1.2\ldots s})^{-1}}{\text{var}(X_1)}\exp(-\beta^T\mu - \frac{1}{2}\beta^T\Sigma\beta),$$

and hence for standardised covariates

$$v(\beta) = (1-\rho_{1.2\ldots s})^{-1}\exp(-\frac{1}{2}\beta^T R\beta),$$

where R is now the correlation matrix. This result demonstrates that the asymptotic variance of $\hat{\beta}_1$ is minimized when X_1 is independent of $X_2 \ldots X_s$, a result which parallels classical multiple regression theory.

To extend this example, consider the case where X_1 is Bernoulli, parameter π, independent of $X_2 \ldots X_s$, and $(X_2 \ldots X_s) \sim MN(\mu, \Sigma)$. One practical application of this is a randomized controlled clinical trial, where subjects are assigned at random to the treatment ($X_1 = 1$) or control ($X_1 = 0$) groups. Then if $\theta = (\theta_1, \ldots, \theta_s)^T$, where $\theta_1 = \text{logit}(\pi)$ and $\tilde{\theta} = (\theta_2, \ldots, \theta_s) = \Sigma^{-1}\mu$,

$$q(\theta) = \log(1 + e^{\theta_1}) + \tilde{\theta}^T\Sigma\tilde{\theta}.$$

Hence using the independence between X_1 and $X_2, \ldots X_s$, and defining $\beta = (\beta_1, \ldots \beta_s)^T$ and $\tilde{\beta} = (\beta_2, \ldots \beta_s)^T$ we have that

$$v(\beta) = \left\{ \frac{1}{1-\pi} + \frac{1}{\pi} \exp(-\beta_1) \right\} \exp(-\tilde{\beta}^T \mu - \frac{1}{2} \tilde{\beta}^T \Sigma \tilde{\beta}).$$

Thus to test $H_0 : \beta = (0, \beta_2, \ldots, \beta_s)$ against $H_1 : \beta = (\beta_1, \beta_2, \ldots, \beta_s)$, the sample size can be calculated from the univariate Bernoulli case, adjusting for the Gaussian covariates by multiplying by a factor of $\exp(-\tilde{\beta}^T \mu - \frac{1}{2} \tilde{\beta}^T \Sigma \tilde{\beta})$. For standardised confounding covariates, the first term of this expression will vanish, and Σ is replaced by the correlation matrix R. By definition this will be positive semi-definite, and hence the adjusting factor will always be less than or equal to 1. In other words, adjusting for known factors which influence the outcome will *always* decrease the required sample size or increase the power.

7.7 Examples

In this section we consider two cases where Poisson regression may be used to analyse medical data. A search of the MEDLINE [58] database of medical abstracts from January 1992 to June 1996 on the terms 'Poisson' and 'regression' resulted in a total of 176 articles, although these do not always apply the technique. This can be contrasted with a search on the terms 'logistic' and 'regression' which results in 4753 articles. Those articles which do apply Poisson regression can be crudely categorised in one of two ways; clinical studies and epidemiological surveys.

In the first case, the unit of analysis is normally an individual patient, and the primary outcome measure is the number of events e.g. epileptic seizures or migraines in a particular time interval. In the second the unit of analysis is typically a group of patients with similar exposure patterns, or a

geographical area, with the response variable being the number of cases of a relatively rare condition such as leukaemia, suicide, etc.

7.7.1 A Randomised Trial

To take an example of a clinical trial where the primary outcome is a count, McMahon et al. [59] describe the pilot phase of the Asymptomatic Cardiac Ischemia Pilot (ACIP) Study. To quote from their paper

> The purpose of ACIP is to compare treatments designed to suppress episodes of transient myocardial ischemia (reduced blood flow to heart muscle) ...

These transient events may be modelled in the first instance by a Poisson process, leading to an comparison of treatment groups using Poisson regression. Screening data consisting of the number of transient ischemic attacks (TIAs) in a 48 hour period was available on 325 patients. This resulted in an estimate of the baseline event rate as $R = 1.41$ episodes per patient. The results of Section 7.4 can be applied to calculate the required number of patients per group for comparing a new therapy to the existing standard for a variety of potential rate reductions and power.

Suppose we consider a relative rate reduction of 20% (from 1.41 to 1.13 episodes per patient) as the minimum clinically significant improvement necessary to change clinical practice, then this corresponds to $\tilde{\beta} = \log(0.8)$. Using equation (7.9), and assuming equal sized groups ($\pi = 0.5$) we calculate that we require 155 patients for 50% power, 369 patients for 80% power and 518 patients for 90% power.

One of the stated primary outcomes of the ACIP trial was to be the number of patients with zero episodes i.e. a binary response. Using the Poisson distribution to calculate the probability of zero events, we can calculate the

required sample size to detect a change in proportions from $0.323 = e^{-1.41}$ to $0.244 = e^{-1.13}$ using logistic regression. The methods of Whittemore [52] only apply when the response probability is small, so we shall use the standard formula for comparing two proportions, as given in Fleiss [60]. This gives required patient numbers of 550 for 50% power, 1070 for 80% power and 1414 patients for 90% power. Clearly the loss of information incurred by ignoring the number of episodes and merely recording their presence or absence is substantial.

In their analysis McMahon et al. detect over-dispersion resulting from large patient variability. They proceed to model the pilot data using a gamma-Poisson mixture model, which practically implies a generalised linear model with negative binomial response. For our illustrative purposes, however, it suffices to use the estimate of the overdispersion given by the ratio of the sample variance to the sample mean, which for this case was $\hat{\sigma}^2 = 6.5/1.4 = 4.6$. Thus, simply inflating each of the calculated sample sizes from the above paragraph by this factor, gives a total of 719, 1713 and 2405 patients for 50%, 80% and 90% power respectively.

These figures are not directly comparable to those quoted in the original paper, where the study design was to use a screening process to only enroll patients with one or more attacks in the initial period. However, Figure 3 of the paper does estimate the power to detect differences in mean number of episodes for a rate reduction of 50%.

Our method, taking into account the aforementioned over-dispersion, calculates that we require 73 patients for 50% power, 194 patients for 80% power and 280 patients for 90% power, whereas the corresponding estimates for ACIP with only patients with at least one episode admitted are approximately 56, 120 and 144 respectively. This reduction in patient numbers can

be explained by the fact that admitting only those patients with more than one episode during the screening phase will both reduce the over-dispersion and increase the baseline rate e^{β_0}. Crudely, the mean number of events in the zero-truncated population will be $1.41/(1 - e^{-1.41}) = 1.87$, so if we decrease the calculated sample sizes by a factor $1.41/1.87 = 0.76$, we get 55 patients for 50% power, 147 patients for 80% power and 211 patients for 90% power, results which are surprisingly close to the much more sophisticated method.

7.7.2 An Epidemiological Survey

Schwartz [61] describes a study into the association between airborne particles and/or ozone concentrations and hospital admissions with respiratory disease for patients aged 65 and over. From the American city of Birmingham, Alabama, daily counts of the total number of hospital admissions for both pneumonia and chronic obstructive pulmonary disease (COPD) were recorded and correlated with daily measurements of ozone concentration in parts per billion (ppb) and the concentration of airborne particles with a diameter of less than 10 μm denoted by PM_{10}. The study ran for four years from January 1, 1986 to December 31, 1989, giving $n = 1461$ sample points for use in a Poisson regression model. Confounding factors also considered were daily temperature, humidity, and seasonal variations.

What level of association does this study have a reasonable chance of detecting?

Quantiles of the distribution of PM_{10} are given in the paper, and we can use them to approximate the daily distribution of this factor by a Gaussian random variable with mean 45 and standard deviation 22. Over the period of the study, the number of patients admitted to hospital with pneumonia

averaged approximately 5.9 per day.

If we thus assume that in the Poisson model we fix a single covariate which is a standardised transform of PM_{10}, then we can use the previous work to calculate the power of this study to detect various effect sizes. Figure 7.3 shows the power as a function of the rate ratio corresponding to a 10 unit increase in PM_{10}.

The dotted line shows that to detect an effect which produces a 5% increase in pneumonia admission rates for each 10 unit rise in PM_{10}, the study has a power of 69%. If the study had been stopped after three years ($N = 1096$) then this would have dropped to 47%. Conversely, extending the study by another year would have increased the power to 83% for this particular combination of effect size and significance level.

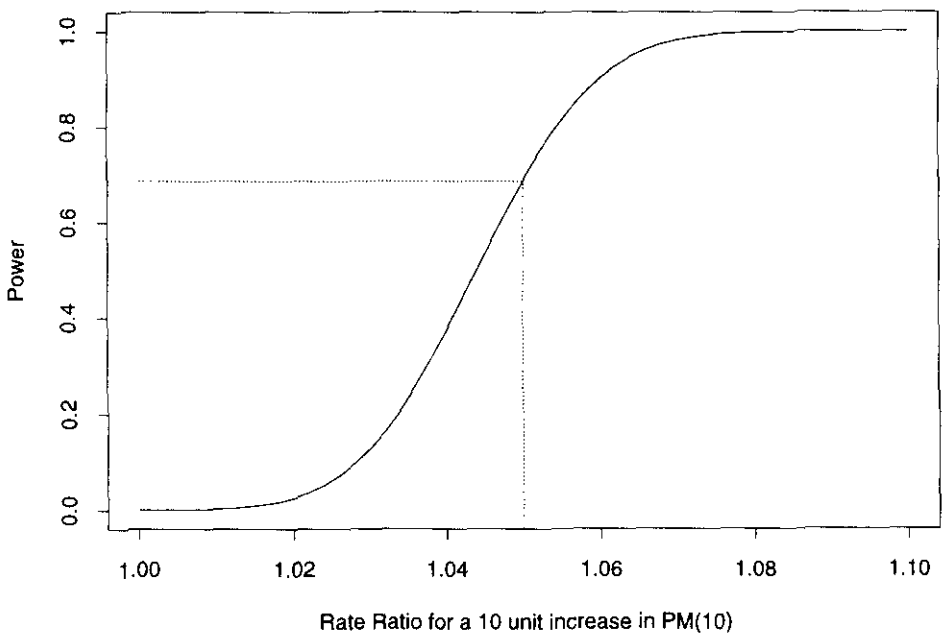

Figure 7.3: Power function for the Alabama air quality study

7.8 Comparison with Alternative Methods

The method of moment generating functions is not the first way of calculating power proposed for generalised linear models. For the likelihood ratio test to compare hierarchical models, the distribution of the test statistic under the null hypothesis has long been known [62]. The distribution under alternative hypotheses, however, has naturally been more difficult to determine.

Self and Mauritsen [63] calculated an asymptotic approximation to the power of a general score test for the parameters of interest. Their method is based upon approximating the distribution of the score statistic T_n by a non-central chi-square. Unfortunately this method entails some tricky calculations and is consequently difficult to implement.

In a later paper, however, Self, Mauritsen and Ohara [64] work with the likelihood ratio test statistic, again approximating the distribution with a non-central χ^2. In this formulation of the problem, however, the calculations are much more practical.

The heart of their argument is to expand the standard log-likelihood ratio test statistic

$$D = 2\left[l_A - l_N\right]$$

where l_A denotes the maximum of the log-likelihood under the model defined by the alternative hypothesis, and l_N is the same quantity, but under the model defined by the null hypothesis, e.g. $\beta_1 = 0$.

The expansion proceeds by writing D as three separate parts. Each part is considered separately, expanded as a Taylor series, and the expectation taken up to terms of order n^{-1}. The statistic D is distributed asymptotically as a χ^2, but this is non-central under the alternative hypothesis. Equating

expectations (i.e. the method of moments) is used to estimate the non-centrality parameter. The degrees of freedom of the distribution is equal to the number of parameters in the hypothesis being tested.

Specialising their argument to the Poisson regression case, and using the previous notation, assume that we wish to test $H_0 : \beta^* = (\beta_0, 0, \beta_2, \ldots, \beta_p)$ against $H_1 : \beta^* = (\beta_0, \tilde{\beta}, \beta_2, \ldots, \beta_p)$, that is, a univariate hypothesis. Define $\eta_i = X_i\hat{\beta}_U^*$, and $\eta_i^* = X_i\hat{\beta}_C^*$, where $\hat{\beta}_U^*$ is the maximum likelihood estimator of β^* under H_1, and $\hat{\beta}_C^*$ is the maximum likelihood estimator of β^* under H_1 subject to the constraint that $\beta_1 = 0$. Thus η_i^* is the maximum likelihood estimator which would result is we assumed H_0 to be true when in fact H_1 was true.

Then, from [64], the non-centrality parameter can be calculated as

$$\gamma = 1 - \mathrm{Tr}\,(M) + \Delta$$

where

$$M = \left[\sum_{i=1}^{n} \exp(\eta_i^*)\mathbf{X_i}\mathbf{X_i}^T\right]^{-1} \left[\sum_{i=1}^{n} \exp(\eta_i)\mathbf{X_i}\mathbf{X_i}^T\right],$$

and

$$\Delta = 2E\left\{\sum_{i=1}^{n} [\exp(\eta_i)(\eta_i - \eta_i^*) - \exp(\eta_i) + \exp(\eta_i^*)]\right\}, \tag{7.11}$$

with the expectation taken under the alternative hypothesis.

However, it can be shown that for the case of Poisson regression with a canonical link $\mu_i = \exp(\eta_i) = \exp(\beta_0 + X^T\beta)$ and a univariate hypothesis that M is always equal to 1, thus simplifying the non-centrality parameter to Δ, and the degrees of freedom to 1.

Note that in this section the individual exposure times t_i have been assumed equal and supressed for notational clarity. The extension to non-equal exposure times is straightforward.

Given Δ and the other parameters, we may then express the power of the test as

$$\text{Power} = 1 - \Psi_1\left(\chi_\alpha | \Delta\right), \tag{7.12}$$

where $\Psi(x|c)$ is the cumulative distribution function of a non-central χ^2 variable with 1 degree of freedom and non-centrality parameter c, and χ_α is the $(1-\alpha)$th percentile of a central χ^2 distribution with 1 degree of freedom.

When there is only a single covariate X, equation (7.11) reduces to

$$\Delta = 2E\left\{\sum_{i=1}^{n}\left[\exp(\beta_0 + X_i\tilde{\beta})(\beta_0 + X_i\tilde{\beta} - \beta_0^*) - \exp(\beta_0 + X_i\tilde{\beta}) + \exp(\beta_0^*)\right]\right\}, \tag{7.13}$$

where

$$\exp(\beta_0^*) = E\left(\frac{1}{N}\sum_{i=1}^{n}\exp(\beta_0 + X_i\tilde{\beta})\right).$$

This result can be seen by noting that assuming H_0 to be true implies that the covariate has no effect and that we are dealing with a homogenous Poisson process observed for time $N\mu_T$. Thus the maximum likelihood estimator of β_0 is the natural logarithm of the sample mean. However, the expected value of this quantity *under the alternative hypothesis* is given by the above quantity.

To link this with the previous work we can evaluate the expectations in equation (7.13) to get

$$\Delta = 2Ne^{\beta_0}\left\{\tilde{\beta}M_X'(\tilde{\beta}) - M_X(\tilde{\beta})\log[M_X(\tilde{\beta})]\right\}.$$

where once again $M_X(t)$ denotes the MGF of the covariate distribution.

Furthermore, if C is distributed as a non-central χ^2 with ν degrees of freedom,and non-centrality parameter λ, then Sankaran [65] showed that it is reasonable to use the approximation that

$$\left(\frac{C}{\nu+\lambda}\right)^{1/2} \sim N\left(\left[1 - \frac{\nu-1}{3(\nu+\lambda)}\right]^{1/2}, (\nu+\lambda)^{-1}\right),$$

and so in this case, we can approximate equation (7.12) by

$$\begin{aligned} \text{Power} &= 1 - \Phi\left(Z_\alpha - (1+\Delta)^{1/2}\right) \\ &= 1 - \Phi\left(Z_\alpha - \sqrt{1 + 2Ne^{\beta_0}\left[\tilde{\beta}M'_X(\tilde{\beta}) - M_X(\tilde{\beta})\log(M_X(\tilde{\beta})\right]}\right) \end{aligned}$$

This can be compared with equation (7.6), where there are obvious similarities.

For a simple practical comparison of the two methods, once again taking the case of a Bernoulli covariate with parameter π, we can see that

$$\beta_0^* = (1-\pi)\exp(\beta_0) + \pi\exp(\beta_0 + \tilde{\beta})$$

and so

$$\Delta = 2N\mu_T e^{\beta_0}\left[(\pi + (1-\pi)e^{\tilde{\beta}})\log\left(\frac{1}{(\pi + (1-\pi)e^{\tilde{\beta}})}\right) + (1-\pi)\tilde{\beta}e^{\tilde{\beta}}\right]. \tag{7.14}$$

Using equation (7.12), for $\pi = 0.5, \alpha = 0.05$, and $N\mu_T e^{\beta_0} = 100$ we can then plot the calculated power and compare it with that calculated by the method outlined in Section 7.4. Figure 7.4 shows the two estimated curves over the complete range of power.

We can clearly see that the curves differ, indicating that the calculated power with depend upon the method used. However, it is important to note that the two methods are calculating the power for *two different tests*. The MGF method applies to the Wald test of the maximum likelihood, whereas Self's method begins from the deviance statistic. That these tests differ is well-known, and has been explored theoretically by Chandra and Joshi [66], and Chandra and Mukerjee [67]. They conclude that although both tests have the same asymptotic limiting distribution, and hence the same Pitman efficiency, Wald's test is less locally powerful. This contradicts the relative

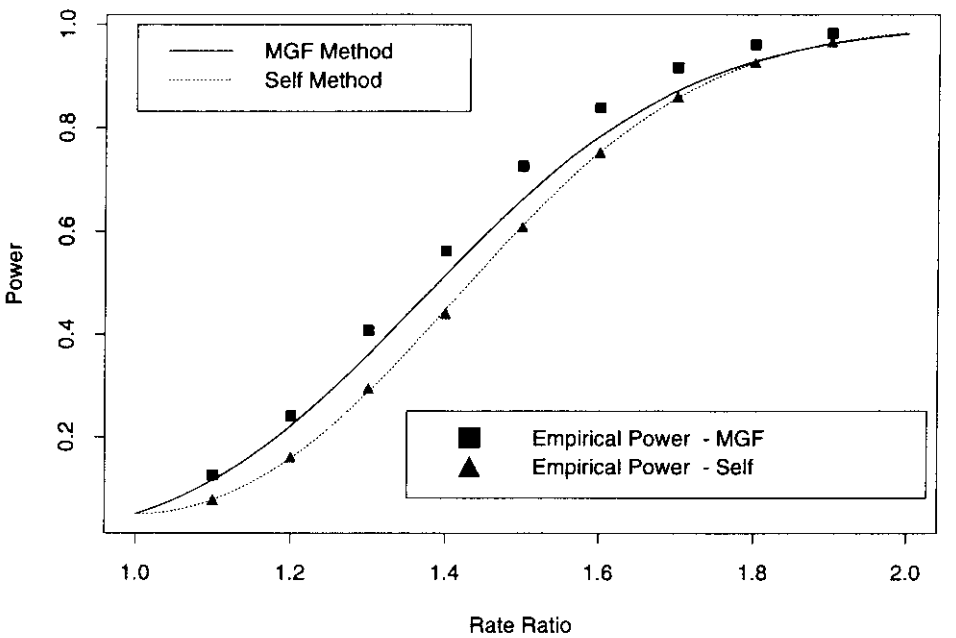

Figure 7.4: Comparison of power functions and empirical results

positions of the two curves, which imply that Wald's test is more powerful in this case.

To examine whether this discrepancy carries through to practical performance, empirical power values from a simulation study very similar to that carried out in Section 7.5 were plotted upon the graph. As can be clearly seen, Self's method performs very well for the deviance test, and the MGF method is slightly conservative for the Wald test. The order of the two tests is unchanged however, suggesting that perhaps the small sample behaviour of these two tests does not follow the theoretical pattern.

7.9 Conclusions

This chapter has introduced a simple and practical way to calculate sample sizes and/or power for Poisson regression models. The dependence of power upon the distribution of covariate(s) is explicitly demonstrated. Simulation experiments suggest that the method is slightly conservative in nature and that the asymptotic approximations are applicable even when the sample size is as small as 30.

Comparisons with another recent method of power calculation for generalised linear models highlights differences between the formulae, but simulations suggest that this is more because they apply to different asymptotic tests rather than any inherent inaccuracy and in the two group comparison case it appears that the Wald test is superior to the likelihood ratio test.

Chapter 8

Estimating the Moment Generating Function

8.1 Introduction

In the previous chapter methods for calculating the power and sample size for Poisson and logistic regression were examined. They are based on expressions for Fisher's information matrix which involve the moment generating function of the distribution of the covariates. One immediately obvious drawback of these methods is, of course, that this distribution is often unknown.

In this chapter we address the problem of estimating the moment generating function (MGF) and various associated functionals non-parametrically, from a univariate sample $X_1, \ldots, X_n$. Comparisons of unsmoothed or empirical estimates are compared with kernel smoothed estimates in terms of the MSE of estimation.

There is a small amount of previous work in this area, centering around consideration of the empirical moment generating function (EMGF), defined

by

$$\hat{m}_0(s) = n^{-1} \sum_{i=1}^{n} \exp(sX_i). \tag{8.1}$$

In a series of papers Maiboroda [68, 69, 70] derives central limit theorems and other asymptotic properties of $\hat{m}_0(s)$, including confidence bands. Epps, Singleton and Pulley [71] use similar properties to construct a test for differing distributions based upon the EMGF.

The most relevant work in this case, however, was done by Gbur and Collins [72]. They compared the EMGF to parametric models fitted by both maximum likelihood and the method of moments. By means of both asymptotic calculation and simulation they demonstrate, perhaps not surprisingly, that if the assumed model is correct the the parametric model is best otherwise

"The empirical MGF is the better estimator in *some* cases."

Even with such equivocal results, however, our interest is in the function $V(s)$ derived from the specially constructed matrix $\mathbf{M}$ of 'zeroth', first and second derivatives of the MGF and thus 'good' estimation of $m_X(s)$ may not necessarily correspond to 'good' estimation of $V(s)$.

8.2 The Univariate Case

Let $\mathbf{X}$ be a continuous random variable. Consider equation (8.1). The empirical MGF can be written as

$$\hat{m}_0(s) = n^{-1} \int \exp(sx) \sum_{i=1}^{n} \delta(x - X_i) dx, \tag{8.2}$$

where $\delta(z)$ is the Dirac delta function. Thus we may consider the EMGF as being derived from a density estimate consisting of a series of probability spikes at the data values, corresponding to a bandwidth $h = 0$.

This leads naturally to the smoothed MGF, replacing the point density by a kernel density estimate $\hat{f}_h(x)$, to get the following (assuming that we use a kernel with domain $[-1,1]$; the extension to Gaussian kernels is trivial)

$$\begin{aligned}\hat{m}_h(s) &= (nh)^{-1}\sum_{i=1}^{n}\int_{X_i-h}^{X_i+h}\exp(sx)K\left(\frac{x-X_i}{h}\right)dx \\ &= n^{-1}\sum_{i=1}^{n}\exp(sX_i)\int_{-1}^{1}\exp(hsu)K(u)du \\ &= \hat{m}_0(s)m_K(hs), \end{aligned} \tag{8.3}$$

where $m_K(v)$ is the MGF of the probability distribution defined by the kernel.

This result is a consequence of the fact that moment generating functions are really Laplace transforms in disguise (see Grimmet and Welsh [73], p.114) and that one interpretation of kernel smoothers is as a convolution of the kernel and the empirical density function.

Trivially, $\hat{m}_0(s)$ is unbiased, as the X_i's are i.i.d., so that

$$E\left[\hat{m}_0(s)\right] = nE\left[n^{-1}e^{sX}\right] = m_X(s),$$

implying that the empirical estimate has MSE equal to the variance. Calculation of this variance gives

$$\text{var}\left[\hat{m}_0(s)\right] = \frac{1}{n}\left[m_X(2s) - m_X(s)^2\right]. \tag{8.4}$$

That this quantity is always positive can be confirmed by an application of the Cauchy-Schwarz inequality

$$E(UV)^2 \leq E(U^2)E(V^2),$$

with $U = e^{sX}$ and $V = 1$.

Returning to equation (8.3), we can see that this implies that the kernel smoothed estimate of the MGF always has a positive asymptotic bias, since

for symmetric kernels

$$E\left[\hat{m}_h(s)\right] = m_X(s)\left[1 + \sum_{i=1}^{\infty} \frac{h^{2i}s^{2i}}{2i!}\mu_{2i}^K\right] > m_X(s), \tag{8.5}$$

where μ_p^K is the pth central moment of the kernel. The fact that only even central moments contribute determines that the summation term is necessarily positive. The smoothed estimate will only be unbiased if kernels of infinite order are used [74] to allow $\mu_p^K = 0$ for all p, a result which parallels the classical bias arguments of kernel density estimation.

Considering the variance of $\hat{m}_h(s)$, we have

$$\text{var}\left[\hat{m}_h(s)\right] = \text{var}\left[\hat{m}_0(s)\right]\left[m_K(hs)\right]^2,$$

resulting in the MSE of the smoothed estimate as

$$\begin{aligned}\text{MSE}\left[m_h(s)\right] &= (m_K(hs) - 1)^2 m_X^2(s) + \\ &\quad \frac{1}{n}\left\{\left[m_X(2s) - m_X^2(s)\right] m_K(hs)^2\right\}\end{aligned} \tag{8.6}$$

Thus, kernel smoothing will *always* increase the MSE of the estimated MGF, introducing positive bias and increasing the variance.

8.3 Derivatives of the MGF

As before, define the empirical estimate of the kth derivative of the MGF by

$$\hat{m}_0^{(k)}(s) = n^{-1}\sum_{i=1}^{n} X_i^k \exp(sX_i), \tag{8.7}$$

and the smoothed estimates as

$$\begin{aligned}\hat{m}_h^{(k)}(s) &= \int x^k e^{sx}\hat{f}_h(x)dx \\ &= (nh)^{-1}\sum_{i=1}^{n}\int_{X_i-h}^{X_i+h} x^k e^{sx} K\left(\frac{x - X_i}{h}\right)dx\end{aligned}$$

$$= n^{-1}\sum_{i=1}^{n} e^{sX_i}\int_{-1}^{1}(X_i+uh)^k e^{u(sh)}K(u)du$$

$$= \sum_{j=0}^{k}\binom{k}{j}h^j\hat{m}_0^{(k-j)}(s)m_K^{(j)}(sh). \qquad (8.8)$$

Once again we see that all empirical derivative estimates are unbiased. Our particular interest focuses however upon the first and second derivatives only. From equation (8.8) smoothed estimates can be written in terms of the empirical estimates as follows:

$$\hat{m}_h'(s) = \hat{m}_0'(s)m_K(hs) + h\hat{m}_0(s)m_K'(hs)$$

$$\hat{m}_h''(s) = \hat{m}_0''(s)m_K(hs) + 2h\hat{m}_0'(s)m_K'(hs) + h^2\hat{m}_0(s)m_K''(hs),$$

suggesting that in these cases the asymptotic bias may not always be positive.

Two useful general results concerning the variance and covariances (and hence the MSE) of the empirical estimates are that

$$\text{var}\left[\hat{m}_0^{(p)}(s)\right] = \frac{1}{n}\left[m_X^{(2p)}(2s) - m_X^{(p)}(s)^2\right], \quad p = 0, 1, \ldots, \qquad (8.9)$$

and

$$\text{cov}\left[\hat{m}_0^{(p)}(s), \hat{m}_0^{(q)}(s)\right] = \frac{1}{n}\left[m_X^{(p+q)}(2s) - m_X^{(p)}(s)m_X^{(q)}(s)\right], \quad p, q = 0, 1, \ldots, \qquad (8.10)$$

both of which are easily proved from the definitions of equation (8.7).

Using these results and equation (8.8), we can calculate the variance for both smoothed estimates and attempt to compare them with that of the empirical estimates.

Beginning with the first derivative, we have

$$\text{var}\left[\hat{m}_h'(s)\right] = m_K(hs)^2\text{var}[\hat{m}_0'(s)] + h^2m_K'(hs)^2\text{var}[\hat{m}_0(s)] + \frac{2h}{n}m_K(hs)m_K'(hs)\left[m_X'(2s) - m_X(s)m_X'(s)\right]. \qquad (8.11)$$

By using similar arguments to those of the previous section, we can see that the first term is always greater than or equal to the variance of the empirical estimate, and that the second term is always positive. This focuses attention upon the third term.

Note that this final term, which is not obviously always positive, would have be of a sufficent negative value to cancel out both the increase over the variance of the $h = 0$ case from the first term *and* the always positive second term. Thus to demonstrate that smoothing is worthwhile in this case we would have to show that the whole of equation (8.11) is smaller than $\text{var}\,[\hat{m}'_h(s)]$.

We can achieve some insight into this problem by using the results of *Silverman and Young* [75], who develop a theory for the asymptotic MSE of smoothed estimates of linear functionals. They work in the context of empirical and smoothed bootstraps, but we can adapt their methodology to apply in our circumstances, for the case of a single covariate.

Let X be a univariate random variable, K be a symmetric kernel with variance 1 (e.g. the Gaussian kernel), and $A(F)$ a linear functional of some univariate distribution function F with density f. Since A is linear, there exists a function $a(t)$ such that

$$A(F) = \int a(t)f(t)dt.$$

Also, denote the empirical estimate of F by F_0, and the smoothed estimate by F_h. That is

$$\begin{aligned} F_0(x) &= n^{-1}\sum_{i=1}^{n} I(X_i \leq x) = \int \sum_{i=1}^{n} \delta(X_i - x)dx \\ F_h(x) &= n^{-1}\sum_{i=1}^{n} L\left(\frac{x - X_i}{h}\right), \end{aligned}$$

where $L(v)$ is the cumulative distribution function of K. Then the empirical

and smoothed estimates of $A(F)$ will be, respectively,

$$\begin{aligned}\hat{A}_0(F) = A(\hat{F}_0) &= n^{-1}\sum_{i=1}^{n} a(X_i) \\ \hat{A}_h(F) = A(\hat{F}_h) &= \int a(t)\hat{f}_h(t)dt.\end{aligned}$$

Restating Theorem 1 of Silverman and Young, we have

Theorem 1 *Suppose $a(X)$ and $a''(X)$ are negatively correlated. Then*

$$\mathrm{MSE}\left\{\hat{A}_h(F)\right\} \leq \mathrm{MSE}\left\{\hat{A}_0(F)\right\},$$

for a suitably chosen h.

Noting that in our case $\hat{A}(F)$ is also a function of s, we can see that if we take $a_s(t) = e^{st}$, then $a_s''(t) = s^2e^{st}$, $\hat{A}_h(F) = \hat{m}_h(s)$ and

$$\begin{aligned}\mathrm{cov}\left\{a(X), a''(X)\right\} &= s^2\left[m_X(2s) - m_X(s)^2\right] \\ &= s^2\mathrm{var}\left[e^{sX}\right] \geq 0,\end{aligned}$$

confirming the result of Section 8.2 that smoothing can never improve the MSE of the estimate of $m_X(s)$.

Similar expansions with appropriate definitions of $a_s(t)$ can be used to investigate the usefulness of otherwise of smoothing for estimates of the first and second derivatives of $m_X(s)$. If we take $a_s(t) = te^{st}$, then $\hat{A}_h(F) = \hat{m}_h'(s)$ and $a_s''(t) = (2s + s^2t)e^{st}$. In this case

$$\begin{aligned}\mathrm{cov}\left\{a(X), a''(X)\right\} &= \mathrm{cov}\left\{s^2Xe^{sX} + 2s\,e^{sX}, Xe^{sX}\right\} \\ &= s^2\mathrm{var}\left[Xe^{sX}\right] + 2s\,\mathrm{cov}\left[Xe^{sX}, e^{sX}\right]. \quad (8.12)\end{aligned}$$

Note that we need only consider the case $s > 0$, since in the context of a risk ratio this corresponds to an increase in Poisson mean with an increase in the covariate, and this can always be achieved through a judicious coding of

the variables. Now for $u > -1$ both e^{su} and ue^{su} are monotonic increasing for all values of s and hence for distributions of X which only take non-negative values, such as the Gamma, Exponential or Poisson distributions, the second covariance term in equation (8.12) will be positive and smoothing will be unnecessary. For non-positive distributions, it would have to be the case that most of the probability density fell below -1 for this term to be negative, suggesting that only in quite extreme cases is there an opportunity for smoothing to improve matters.

A similar argument can be applied to the smoothed estimate of the second derivative to give the condition that smoothing will only reduce the MSE when

$$s^2 \text{var}\left[X^2 e^{sX}\right] + 4s \text{ cov}\left[Xe^{sX}, X^2 e^{sX}\right] + 2 \text{ cov}\left[e^{sX}, X^2 e^{sX}\right] < 0,$$

for some value of $s > 0$. Once again we can argue that e^{su}, ue^{su} and $u^2 e^{su}$ are monotonic increasing (and hence positively correlated) for $u > 0$, and so in the majority of positive distributions and those symmetric distributions centred about zero, smoothing should be unnecessary.

8.4 Estimating $V(s)$

Although we have explored the use of smoothing in the estimation of the MGF and its derivatives, we have yet to consider the crucial function of the previous chapter, namely

$$V(s) = \frac{m_X(s)}{m_X(s)m''_X(s) - m'_X(s)^2}. \tag{8.13}$$

If we substitute either empirical or smoothed estimates into this expression, can we evaluate the mean or the variance? Parallels may be drawn with results from the binary regression estimators of Part I, where ISE-optimal

estimation of the two density components did not automatically translate into better estimation of the probability function.

We begin be expanding $\hat{V}(s)$ about $V(s)$. In all that follows, the dependence upon both X and s have been suppressed. Using the fact that the differences between the estimators and the true functions are 'small' asymptotically, we have

$$\begin{aligned}
\hat{m} &= m\left(1+\frac{\hat{m}-m}{m}\right) \\
\hat{m}'^2 &\simeq m'^2\left(1+\frac{2(\hat{m}'-m')}{m'}\right) \\
\hat{m}\hat{m}'' &\simeq mm''\left(1+\frac{\hat{m}''-m''}{m''}+\frac{(\hat{m}-m)}{m}\right).
\end{aligned}$$

Substituting these values into the formula for $\hat{V}(s)$ derived from equation (8.13) and performing the usual manipulations gives

$$\begin{aligned}
\hat{V}(s) \simeq\ & V(s)+(\hat{m}-m)\frac{V(s)}{m}\left[1-V(s)m''\right]+(\hat{m}'-m')\frac{2m'V(s)^2}{m} \\
& -(\hat{m}''-m'')V(s)^2. \qquad (8.14)
\end{aligned}$$

Thus we can immediately see that $\hat{V}_0(s)$, the empirical estimate of $V(s)$ corresponding to the case $h=0$, is asymptotically unbiased. This suggests that any improvement which smoothing the estimates of the MGF and its derivatives in this case would have to be involve a dramatic reduction in the variance at the expense of a small increase in the bias.

To summarise our results so far, the empirical estimates of the MGF and its derivatives are, in most circumstances, superior in terms of MSE to kernel smoothed estimates. Furthermore, the empirical estimate of $V(s)$ is asymptotically unbiased. Although we could go on to examine the asymptotic variance of $\hat{V}(s)$ using the above expansion, it quickly becomes clear that the numerous functions of $V(s)$, $m_X(s)$ and its derivatives lead to intractable expressions which can only be considered by specifying particular

distributions of X.

With this in mind it would seem reasonable to recommend that, when using this method to calculate a nonparametric estimate of the function $V(s)$ to calculate sample size or power for Poisson regression, the use of the asymptotically unbiased empirical estimate gives both a computationally simple solution and adequate MSE performance.

8.5 Categorical Covariates

Although we have concentrated almost solely on continuous covariates in this chapter, much of the preceding results can be applied to categorical data as well.

Consider the case of a single Bernoulli covariate X, with parameter π. The empirical MGF will be

$$\hat{m}_0(s) = n^{-1}\sum_{i=1}^{n}\exp(sX_i) = n^{-1}\left[\sum_{i=1}^{n}X_i e^s + (n - \sum_{i=1}^{n}X_i)\right] = \hat{p}e^s + (1-\hat{p}),$$

where $\hat{p}$ is the observed proportion of successes($X_i = 1$). Interestingly, in this case, since $\hat{p}$ is also the MLE of π, the parametric and non-parametric estimates coincide. These calculations can be extended to the multinomial distribution with the same intuitive results.

Thus, for categorical covariates, the EMGF is identical to the function obtained by substituting the observed proportions (the maximum likelihood estimates) into the theoretical MGF.

8.6 Conclusions

In this chapter we have explored the practical application of the methods described in Chapter 7. Estimation of the function $V(s)$ can be easily accomplished by using the empirical moment generating function and its derivatives, and although there are situations where kernel smoothing may be advantageous, these are unlikely to occur in practical situations.

Conclusions

The three components of this thesis are linked both by the use of kernel smoothing methods and the focus on practical issues. In Part I the simple binary regression estimate first described by Copas was extended to a two bandwidth case, which incorporated both the original formulation and the concept of treating the problem as two separate density estimations as special cases. Asymptotic analysis and an extensive simulation experiment showed that this latter approach of decomposing the problem did not work, and that, especially for cases where the variabilities of successes and failures were quite different, the use of two bandwidths can lead to dramatic improvements in estimation.

The use of the much-promoted local polynomial semiparametric models was then explored, and the results compared to the fully nonparametric binary regression estimators (which were shown to be a special case of these more general estimators). Although these more complex methods impose a much greater computational burden, the single bandwidth locally linear logistic estimator is as good as the two bandwidth nonparametric one, except in the most extreme of cases. This highlights the use in the simulation experiments of the 'best possible' bandwidth in every case, to decouple the choice of estimator from the choice of bandwidth. Given that a data-dependent bandwidth must eventually be chosen, it is easier to choose one bandwidth

rather than two.

Chapter 3 completes the list of potential binary regression estimators by extending the locally linear logistic approach to the two bandwidth case. Although in this case the improvements achieved in the estimation as measured by the chosen error function are dramatic, closer inspection reveals that this may be due more to the correction of the crude prevalence rates rather than getting closer to the true probability function, suggesting that these gains are unlikely to be reproducible with a data-dependent bandwidth choice, a fact confirmed by later simulation results.

In the final chapter of Part I two different approaches to bandwidth selection for the binary regression estimators were extended to the new estimators and compared. It was demonstrated that, although a cross-validation method can do better than a simpler plug-in rule for some very easy to estimate situations, the combination of the locally linear logistic single bandwidth estimator and a plug-in bandwidth selection rule performs very well except for cases where the variabilities of successes and failures differed markedly. In these cases a nonparametric two bandwidth estimator should be used, and a simple rule of thumb for practical application was proposed.

In Part II, attention moved to density estimation, and in particular an attempt was made to assess and compare the more promising of the large number of 'improvements' upon the standard kernel density estimator which have been proposed. All these estimators are, in the asymptotic sense, 'better' than the basic estimator, so the focus was on small-sample performance, and a simulation experiment using the optimal bandwidth in each case was again performed. The results suggest that, at least for sample sizes of 500 or less, theoretical enhancements of the standard KDE are not always carried forward into practice.

The most promising of the improved estimators, however, was the multiplicative bias correction of Jones, Linton and Nielsen, and a simple rule-of-thumb bandwidth selector based upon a sophisticated scale estimate for the standard KDE case was extended to provide a simple data-dependent bandwidth selector for the new methods. The fact that the 'best case' improvements observed in the previous simulations can be carried through to practice even when using a very simple bandwidth selector is reassuring. Furthermore, the results would seem to support the more general conclusion that the gains made by considering more sophisticated estimators are greater than those achieved by more sophisticated bandwidth selection algorithms for the simple estimators.

Finally, Part III considers the calculation of sample size and/or power for Poisson regression, and how this can be achieved using knowledge about the distribution of the covariates, expressed through their moment generating function (MGF). The estimation of this function and its derivatives can be achieved using kernel smoothing, but it was shown that, in most situations, the kernel smoothed estimate with a bandwidth tending to zero, i.e. the empirical MGF, was most appropriate.

To summarise, the use of kernel smoothing has been considered in three practical problems. The use of two bandwidths in binary regression is a novel approach with clear benefits for certain cases. The work on bandwidth selection is the first comparison between any of the suggested approaches to this crucial problem, and the rule-of-thumb derived for practical use can be easily applied. The comparison of a large number of higher-order kernel density estimators is an important unification and a practical attempt to weed out the less promising approaches, and the extension of the simple rule-of-thumb bandwidth selectors to the most promising improved estimator is

also new. Finally, the use of an estimated MGF rather than an educated guess to calculate the sample size or power for a Poisson regression model extends the usefulness of this previously published methodology.

Bibliography

[1] D. Collett. *Modelling Binary Data.* Chapman and Hall, London, UK, 1991.

[2] D.R. Cox and E.J. Snell. *Analysis of Binary Data.* Chapman and Hall, London, UK, second edition, 1989.

[3] P. McCullagh and J.A. Nelder. *Generalized Linear Models.* Chapman and Hall, London, second edition, 1989.

[4] J. Fan, N.E. Heckman, and M.P. Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90:141–150, 1995.

[5] J.B. Copas. Plotting p against x. *Applied Statistics*, 32:25–31, 1983.

[6] E. Fix and J.L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951. Project Number 21-49-004.

[7] E. Fix and J.L. Hodges. Discriminatory analysis - nonparametric discrimination: Consistency properties. *International Statistical Review*, 57:238–247, 1989.

[8] B.W. Silverman and M.C. Jones. E. Fix and J.L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation. *International Statistical Review*, 57:233–238, 1989.

[9] B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, UK, 1996.

[10] E.A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.

[11] G.S. Watson. Smooth regression analysis. *Sankhya, Series A*, 26:101–116, 1964.

[12] M.C. Rodriguez-Campos and R. Cao-Abad. Nonparametric bootstrap confidence intervals for discrete regression functions. *Journal of Econometrics*, 58:207–222, 1993.

[13] R.F. Kappenman. Nonparametric binary regression without replication. *Computational Statistics and Data Analysis*, 6:119–125, 1988.

[14] M.P. Wand and M.C. Jones. *Kernel Smoothing.* Chapman and Hall, London, UK, 1995.

[15] J. Fan. Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87:998–1004, 1992.

[16] J.S. Marron and M.P. Wand. Exact mean integrated squared error. *Annals of Statistics*, 20:712 736, 1992.

[17] S.J. Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53:683–690, 1991.

[18] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London, 1996.

[19] M.G. Schimek. Non- and semiparametric alternatives to generalized linear models. *Computational Statistics*, 12:173–191, 1997.

[20] M. Bonneu, M. Delecroix, and E. Malin. Semiparametric versus nonparametric estimation in single index regression model: A computational approach. *Computational Statistics*, 8:207–222, 1993.

[21] R.W. Klein and R.H. Spady. An efficient semiparametric estimator for binary response models. *Econometrica*, 61:387–421, 1993.

[22] P.J. Green. Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55:245–259, 1987.

[23] H. Chen. Asymptotically efficient estimation in semiparametric generalized linear models. *The Annals of Statistics*, 23:1102–1129, 1995.

[24] T.A. Severini and J.G. Staniswallis. Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89:501–511, 1994.

[25] P. Hall and R.D. Murison. Correcting the negativity of high-order kernel density estimators. *Journal of Multivariate Analysis*, 47:103–122, 1993.

[26] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[27] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandanavian Journal of Statistics*, 9:65–78, 1982.

[28] A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71:353–360, 1984.

[29] B.U. Park and J.S. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85:66–72, 1990.

[30] J. Fan, M. Farmen, and I. Gijbels. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society, Series B*, 60:591–608, 1998.

[31] M.C. Jones and D.F. Signorini. A comparison of higher-order bias kernel density estimators. *Journal of the American Statistical Association*, 92:1063–1073, 1997.

[32] M.C. Jones and P.J. Foster. Generalized jackknifing and higher order kernels. *Journal of Nonparametric Statistics*, 3:81–94, 1993.

[33] M.C. Jones, O. Linton, and Neilsen J.P. A simple bias reduction method for density estimation. *Biometrika*, 82:327–338, 1995.

[34] D. Ruppert and D.B.H. Cline. Bias reduction in kernel density estimation by smoothed empirical transforms. *The Annals of Statistics*, 22:185–210, 1994.

[35] N. Victor. Non-parametric allocation rules. In F.T. De Dombal and F. Gremy, editors, *Decision Making and Medical Care: Can Information Science Help?*, pages 515–529. North-Holland, 1976.

[36] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19:135–144, 1977.

[37] I.S. Abramson. On bandwidth variation in kernel estimates - a square root law. *The Annals of Statistics*, 9:168–76, 1982.

[38] M. Samiuddin and G.M. el Sayyad. On nonparametric kernel density estimates. *Biometrika*, 77:865–74, 1990.

[39] N.L. Hjort and I. Glad. Nonparametric density estimattion with a parametric start. *The Annals of Statistics*, 23:882–904, 1995.

[40] M.C. Jones, D.F. Signorini, and N.L. Hjort. On multiplicative bias correction in kernel density estimation. *Sankhya (submitted)*, 1998.

[41] N.L. Hjort and M.C. Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24:1619–47, 1996.

[42] C.R. Loader. Local likelihood density estimation. *The Annals of Statistics*, 24:1602–18, 1996.

[43] J. Fan and J.S. Marron. Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, 3:35–56, 1994.

[44] M.C. Jones, J.S. Marron, and S.J. Sheather. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, 11:337–381, 1996.

[45] S.J. Sheather. The performance of six popular bandwidth selection methods on some real data sets. *Computational Statistics*, 7:225–250, 1992.

[46] B.U. Park and B.A. Turlach. Practical performance of several data driven bandwidth selectors (with discussion). *Computational Statistics*, 7:251–285, 1992.

[47] B.U. Park and J.S. Marron. On the use of pilot estimators in bandwidth selection. *Journal of Nonparametric Statistics*, 1:231–240, 1992.

[48] P. Janssen, J.S. Marron, N. Veraverbeke, and W. Sarle. Scale measures for bandwidth selection. *Journal of Nonparametric Statistics*, 5:359–380, 1995.

[49] J.M. Stevenson and D.R. Olson. Methods for analysing county-level mortality rates. *Statistics in Medicine*, 12:393–401, 1993.

[50] R.A. Parker. Analysis of surveillance data with Poisson regression : A case study. *Statistics in Medicine*, 8:285–294, 1989.

[51] D.F. Signorini. Sample size for Poisson regression. *Biometrika*, 78:446–50, 1991.

[52] A.S. Whittemore. Sample size for logistic regression with small probability. *Journal of the American Statistical Association*, 76:27–32, 1981.

[53] R. Van de Ven and N.C. Weber. Log-linear models for mean and dispersion in mixed Poisson regression models. *Australian Journal of Statistics*, 37:205–216, 1995.

[54] N.D. Yanez III and J.R. Wilson. Comparison of quasi-likelihood models for overdispersion. *Australian Journal of Statistics*, 37:217–231, 1995.

[55] O Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* Wiley, Chichester, 1978.

[56] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Academic Press, London, 1979.

[57] D.G. Kleinbaum, L.L. Kupper, and K.E. Muller. *Applied Regression Analysis and Other Multivariable Methods.* PWS-KENT, Boston, second edition, 1988.

[58] Ovid Technologies Inc. Medline index of medical abstracts, 1996.

[59] R.P. McMahon, M. Prochan, N. Geller, P.H. Stone, and G. Sopko. Sample size calculations for clinical trials in which entry criteria and outcomes are counts of events. *Statistics in Medicine*, 13:859–870, 1994.

[60] J.L. Fleiss. *Statistical Methods for Rates and Proportions.* Wiley, New York, second edition, 1981.

[61] J. Schwartz. Air pollution and hospital admissions for the elderly in Birmingham, Alabama. *American Journal of Epidemiology*, 139:589–598, 1994.

[62] D.R. Cox and D.V. Hinkley. *Theoretical Statistics.* Chapman and Hall, London, 1974.

[63] S.G. Self and R.H. Mauritsen. Power/sample size calculations for generalized linear models. *Biometrics*, 44:79–86, 1988.

[64] S.G. Self, R.H. Mauritsen, and J. Ohara. Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 48:31–39, 1992.

[65] M. Sankaran. Approximations to the non-central chi-square distribution. *Biometrika*, 50:199–204, 1963.

[66] T.K. Chandra and S.N. Joshi. Comparison of the likelihood ratio, Rao's and Wald's tests and a conjecture of C.R.Rao. *Sankhya, Series A*, 45:226–246, 1983.

[67] T.K. Chandra and R. Mukerjee. Comparison of the likelihood ratio, Wald's and Rao's tests. *Sankhya, Series A*, 47:271–284, 1985.

[68] R.E. Maiboroda. An estimator of the moment generating function of a random variable from the results of observations. *Theory of Probability and Mathematical Statistics*, 32:141–151, 1985.

[69] R.E. Maiboroda. Estimation of a moment generating function from observations with error. *Theory of Probability and Mathematical Statistics*, 39:105–109, 1989.

[70] R.E. Maiboroda. The central limit theorem for empirical moment generating functions. *Theory of Probability and its Applications*, 34:332–335, 1989.

[71] T.W. Epps, K.J. Singleton, and L.B. Pulley. A test of separate families of distributions based on the empirical moment generating function. *Biometrika*, 69:391–399, 1982.

[72] E.E. Gbur and R.A. Collins. Estimation of the moment generating function. *Communications in Statistics, Part B*, 18:1113–1134, 1989.

[73] G. Grimmett and D. Welsh. *Probability: An Introduction.* Clarendon Press, Oxford, 1986.

[74] L. Devroye. *A Course in Density Estimation.* Birkhauser, Boston, 1987.

[75] B.W. Silverman and G.A. Young. The bootstrap: To smooth or not to smooth? *Biometrika*, 74:469–479, 1987.