Bernoulli **1(1/2)**, 1995, 003–016

Kernel estimation of relative risk

JULIA E. KELSALL and PETER J. DIGGLE*

Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK

Estimation of a relative risk function using a ratio of two kernel density estimates is considered, concentrating on the problem of choosing the smoothing parameters. A cross-validation method is proposed, compared with a range of other methods and found to be an improvement when the actual risk is close to constant. In particular, theoretical and empirical comparisons demonstrate the advantage of choosing the smoothing parameters jointly. The methodology was motivated by a class of problems in environmental epidemiology, and an application in this area is described.

Keywords: cross-validation; epidemiology; kernel density estimation; smoothing parameters

1. Introduction

The work in this paper was motivated by the following class of problems arising in environmental epidemiology. A map gives the locations of all cases of a particular disease within a designated geographical region, and the locations of a set of controls chosen as a random sample from the population at risk in the region. We wish to estimate spatial variation in the disease risk (see, for example, Bithell 1990; 1992). A closely related class of problems involves estimation of the spatial variation in the relative risk of two diseases. Similar questions can be asked of disease distributions in time, or space-time.

As an idealization of the one-dimensional version of this class of problems, we consider a set of points x_i, $i = 1, \ldots, n_1$, arising as a partial realization of a Poisson process on an interval I with intensity $\lambda_1(x)$, together with a set of points y_j, $j = 1, \ldots, n_2$, arising as an independent realization of a Poisson process on I with intensity $\lambda_2(x)$. Our objective is to obtain a nonparametric estimate of the ratio $\lambda_1(x)/\lambda_2(x)$. Our particular motivation has been a study of the spatial variation in relative risk for two types of cancer, for which it is important to treat the two sets of points symmetrically. We shall therefore consider estimation of the function

$$\rho(x) = \log\{\lambda_1(x)/\lambda_2(x)\},$$

which has the required symmetry.

Note that if we condition on the values of n_1 and n_2, the data can be treated as a pair of independent random samples from two probability distributions with densities $f(x)$ and $g(x)$ proportional to $\lambda_1(x)$ and $\lambda_2(x)$, respectively. We shall exploit this duality in the theoretical development of our estimator for $\rho(x)$.

* To whom correspondence should be addressed.

1350-4864

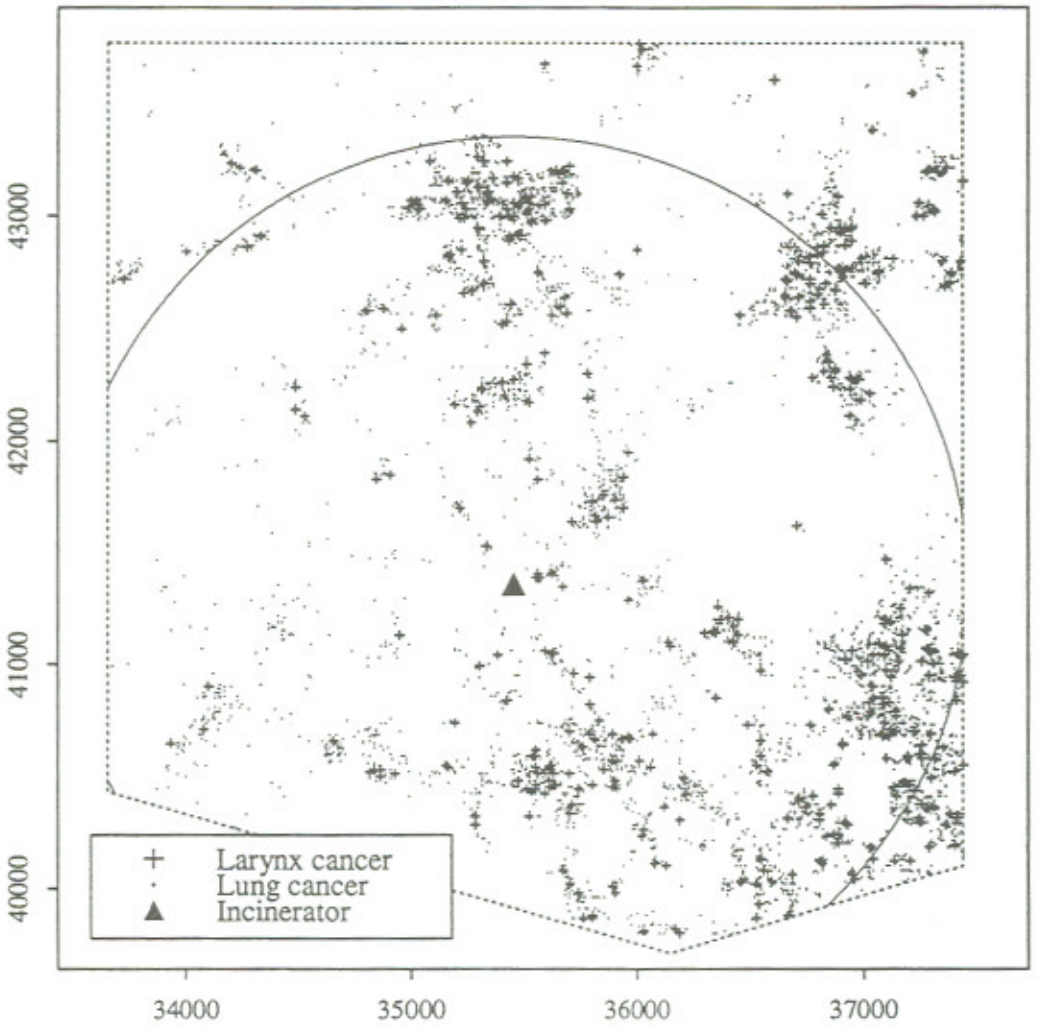

Figure 1. The Chorley-Ribble data

The restriction to one-dimensional processes simplifies notation and various technical details required in the theoretical development. However, our methodology can also be applied to spatial or space-time data.

The map in Fig. 1 shows the residential locations of all cases of larynx and lung cancer diagnosed in part of Lancashire, England, during 1974–83. A suspected focus of increased risk of cancer of the larynx was a now disused industrial incinerator located in the centre of the region. An apparent cluster of four cases near the incinerator prompted scrutiny of the data with a view to confirming or refuting the suspicion of association. The locations of lung cancer cases can be viewed as controls which reflect the spatial variation in population intensity, since no association between the

incinerator and lung cancer was suspected. An estimate of relative risk as a function of distance from the incinerator is therefore required. A subset of these data, from a smaller geographical region, was analysed by Diggle (1990). He fitted a parametric model for the variation in relative risk as a function of the squared distance from the incinerator, and found a significant relationship. Elliott *et al.* (1992) analysed data from a number of similar incinerators in the UK, using methodology due to Stone (1988) in which the relative risk is modelled nonparametrically but is assumed to be a monotone non-increasing function of distance from the incinerator. They found that the apparent relationship between relative risk and distance in the Lancashire data was not consistently reproducible at other sites. Our immediate objective is to construct a nonparametric estimate of relative risk without assuming that the relationship with distance from the incinerator is monotone. The circle in Fig. 1 is of radius 20 km, centred on the incinerator.

In Section 2 of the paper, we consider the use of kernel density estimators (Silverman 1986) for the separate densities $f(x)$ and $g(x)$, and show how our interest in the log ratio, $\rho(x)$, influences our choice of smoothing parameters for $\hat{f}(x)$ and $\hat{g}(x)$. Bithell (1990; 1992) suggested the use of kernel estimators in this context. He did not suggest a method for choosing the smoothing parameters, but noted that in one application better results were obtained by using the same value of the smoothing constant for the numerator and denominator densities, despite the fact that the two sample sizes differed by a factor of 3. Our results provide some theoretical support for this under the assumption, often reasonable in the epidemiological setting, that the numerator and denominator densities are approximately equal. In Section 3 we suggest a cross-validatory prescription for the choice of smoothing parameters to estimate an arbitrary, smooth functional of f and g. Estimation of $\rho(\cdot)$ follows as a special case. We also use simulations to compare the performance of the cross-validation prescription with a number of other methods, and in particular demonstrate the advantage of choosing the smoothing parameters jointly when the numerator and denominator densities are approximately equal. Section 4 discusses the construction of pointwise tolerance intervals for $\hat{\rho}(\cdot)$. Section 5 contains the results of applying our method to the data of Fig. 1.

2. Kernel estimation of a log density ratio

2.1. KERNEL ESTIMATION

Suppose that X_i, $i = 1, \ldots, n$, are mutually independent with common, but unknown, density f, defined on $\mathbb{R}$. The kernel density estimate of f is given by

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} h^{-1} K\{h^{-1}(x - X_i)\},$$

where $K(\cdot)$ is the kernel function and h is the smoothing parameter, or *bandwidth*. It is generally accepted that the choice of kernel function is not critical (e.g. Silverman 1986, Chapter 3). We shall use the standard Gaussian density, $K(t) = (2\pi)^{-1/2} \exp(-t^2/2)$. The bandwidth must be chosen more carefully since its value can have a large effect on the resulting estimate of f. It can be shown (see, for example, Silverman 1986, Chapter 3), that the mean and variance of the estimate $\hat{f}_h(x)$ are

given by

$$\mathrm{E}\{\hat{f}_h(x)\} = f(x) + k_2 h^2 f''(x)/2 + o(h^2), \tag{2.1}$$

$$\mathrm{var}\{\hat{f}_h(x)\} = k_1 n^{-1} h^{-1} f(x) + o(n^{-1}h^{-1}), \tag{2.2}$$

where $k_1 = \int K(t)^2\,\mathrm{d}t$ and $k_2 = \int t^2 K(t)\,\mathrm{d}t$. From (2.1) and (2.2) we can derive an asymptotic approximation for the mean integrated square error (MISE), which leads by standard calculus to an 'optimal' smoothing parameter, h_{opt}, given by

$$h_{\mathrm{opt}}^5 = k_1 k_2^{-2} \left\{ \int f''(x)^2\,\mathrm{d}x \right\}^{-1} n^{-1}. \tag{2.3}$$

Unfortunately, this expression depends on the unknown density, f. Silverman (1986) gives a good general discussion of various practical methods of choosing h. Bowman (1985), in a comparative study, concludes that the method of least-squares cross-validation (Rudemo 1982; Bowman 1984) often gives good results.

2.2. APPROXIMATE MISE

Suppose that X_i, $i = 1, \ldots, n_1$, are mutually independent with common density f, and that $Y_j, j = 1, \ldots, n_2$, are also mutually independent (of each other and of the X_i) with common density g. We wish to estimate $\rho(x) = \log\{f(x)/g(x)\}$ on the interval I, using the estimator

$$\hat{\rho}_{h_1,h_2}(x) = \log\{\hat{f}_{h_1}(x)/\hat{g}_{h_2}(x)\}.$$

For our asymptotic approximations to be valid, we must assume that n_1 and n_2 are large and that h_1 and h_2 are small, decreasing as the sample sizes increase. We also assume that f and g are both bounded away from zero and twice differentiable on the interval I.

We shall derive an approximation for the MISE,

$$\mathrm{MISE}(\hat{\rho}_{h_1,h_2}) = \int_I \mathrm{E}[\{\hat{\rho}_{h_1,h_2}(x) - \rho(x)\}^2]\,\mathrm{d}x. \tag{2.4}$$

Although this criterion is not directly relevant to the epidemiological applications which motivated us, it is a conventional measure which is widely used in comparisons of automatic procedures for bandwidth choice. We define the *relative error* term for $\hat{f}_{h_1}(x)$ as

$$\epsilon_f(x; h_1) = \{\hat{f}_{h_1}(x) - f(x)\}/f(x), \tag{2.5}$$

for $x \in I$, and similarly $\epsilon_g(x; h_2)$ for $\hat{g}_{h_2}(x)$. According to our assumptions, these error terms will be small. Rearranging (2.5) gives

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

and similarly for $\hat{g}_{h_2}(x)$. This leads, via a first-order Taylor expansion of the log function, to

$$\hat{\rho}_{h_1,h_2}(x) = \rho(x) + \epsilon_f(x; h_1) - \epsilon_g(x; h_2) + O(\epsilon^2), \tag{2.6}$$

where ϵ denotes either $\epsilon_f(x;h_1)$ or $\epsilon_g(x;h_2)$. The following approximations for the mean and variance of $\hat{\rho}_{h_1,h_2}(x)$ on I can now be derived, using equations (2.1), (2.2), (2.5) and (2.6):

$$\mathrm{E}\{\hat{\rho}_{h_1,h_2}(x)\} \simeq \rho(x) + k_2\{h_1^2 f''(x)/f(x) - h_2^2 g''(x)/g(x)\}/2,$$

$$\mathrm{var}\{\hat{\rho}_{h_1,h_2}(x)\} \simeq k_1\{n_1^{-1}h_1^{-1}f(x)^{-1} + n_2^{-1}h_2^{-1}g(x)^{-1}\}.$$

Note in particular that if $f = g$ on I and we fix $h_1 = h_2$, then to this order of magnitude the estimator for $\rho(x)$ is unbiased. To the same degree of approximation, the MISE in (2.4) is

$$\begin{aligned}\mathrm{MISE}(\hat{\rho}_{h_1,h_2}) \simeq\ & k_1(n_1^{-1}h_1^{-1}A_f + n_2^{-1}h_2^{-1}A_g) \\ & + k_2^2(h_1^4 B_{ff} - 2h_1^2h_2^2B_{fg} + h_2^4B_{gg})/4, \end{aligned} \tag{2.7}$$

where $A_f = \int_I f(x)^{-1}\,\mathrm{d}x$, $B_{fg} = \int_I \{f''(x)/f(x)\}\{g''(x)/g(x)\}\,\mathrm{d}x$ and A_g, B_{ff}, B_{gg} are similarly defined. Note that, on our interval I, the closer to zero f and g become, the larger A_f and A_g, so the larger the MISE; the more rapidly fluctuating f and g, especially at low densities, the larger B_{ff}, so the larger the MISE; the more the fluctuations of f and g are out of phase, especially when rapidly fluctuating at low densities, the more negative the B_{fg} term, and so the larger the MISE. These observations are all intuitively reasonable.

Consider now the estimation of $\log f$. Following similar steps to those in the derivation of (2.7) we obtain

$$\mathrm{MISE}(\log \hat{f}_{h_1}) \simeq k_1 n_1^{-1}h_1^{-1}A_f + k_2^2h_1^4B_{ff}/4. \tag{2.8}$$

Thus

$$\mathrm{MISE}(\hat{\rho}_{h_1,h_2}) \simeq \mathrm{MISE}(\log \hat{f}_{h_1}) + \mathrm{MISE}(\log \hat{g}_{h_2}) - k_2^2h_1^2h_2^2B_{fg}/2.$$

This shows the importance of the B_{fg} term. If $B_{fg} \neq 0$, then the jointly optimal bandwidths, h_{J_1} and h_{J_2} for estimating $\rho(x)$, are different from the separately optimal bandwidths, h_{S_1} and h_{S_2} for estimating $\log f(x)$ and $\log g(x)$. From equation (2.8), it is straightforward to show that h_{S_1} and h_{S_2} satisfy

$$\begin{aligned} h_{S_1}^5 &= k_1k_2^{-2}A_fB_{ff}^{-1}n_1^{-1} \\ h_{S_2}^5 &= k_1k_2^{-2}A_gB_{gg}^{-1}n_2^{-1}. \end{aligned} \tag{2.9}$$

Similarly, by straightforward calculus and substitution into (2.9), we can show that the jointly optimal bandwidths, h_{J_1} and h_{J_2}, which minimize the MISE approximation in (2.7), satisfy

$$\begin{aligned} h_{J_1}^5 &= h_{S_1}^5 + B_{fg}B_{ff}^{-1}h_{J_2}^2h_{J_1}^3 \\ h_{J_2}^5 &= h_{S_2}^5 + B_{fg}B_{gg}^{-1}h_{J_1}^2h_{J_2}^3. \end{aligned} \tag{2.10}$$

Although these are not explicit formulae for h_{J_1} and h_{J_2}, they do give some useful insight into their values compared with h_{S_1} and h_{S_2}. From inspection of equations (2.7), (2.8) and (2.10), and the definitions of B_{ff}, B_{fg} and B_{gg}, we obtain the following results:

(i) If $B_{fg} > 0$ then $h_{J_1} > h_{S_1}$ and $h_{J_2} > h_{S_2}$,
(ii) If $B_{fg} = 0$ then $h_{J_1} = h_{S_1}$ and $h_{J_2} = h_{S_2}$,
(iii) If $B_{fg} < 0$ then $h_{J_1} < h_{S_1}$ and $h_{J_2} < h_{S_2}$.

A consequence of (ii) is that if f or g is linear on I, then h_{J_1} and h_{J_2} are the same as h_{S_1} and h_{S_2}, respectively. Another result derived from expressions (2.10) is as follows.

Theorem If $B_{fg} > 0$ and $0 < h_{J_1}, h_{J_2} < \infty$ then

$$h_{J_i} > (1 - B_{ff}^{-1} B_{fg}^2 B_{gg}^{-1})^{-1/5} h_{S_i} \qquad \text{for } i = 1, 2. \tag{2.11}$$

Proof
Rearrange the first expression in (2.10) to give

$$h_{J_1}^5 = h_{S_1}^5 (1 - B_{fg} B_{ff}^{-1} h_{J_2}^2 h_{J_1}^{-2})^{-1}.$$

The expression in parentheses must be strictly positive, so $h_{J_1}^2 h_{J_2}^{-2} > B_{fg} B_{ff}^{-1}$. Similarly, $h_{J_2}^2 h_{J_1}^{-2} > B_{fg} B_{gg}^{-1}$, and this gives

$$1 - B_{fg} B_{ff}^{-1} h_{J_2}^2 h_{J_1}^{-2} < 1 - B_{ff}^{-1} B_{fg}^2 B_{gg}^{-1},$$

which leads to the result. □

The above theory confirms that there is scope to obtain improved estimates of $\rho(x)$ by choosing the two bandwidths jointly. However, as in (2.3) all our formulae involve the unknown densities, f and g, so they are of limited use in practice. In the next section we consider some practical methods for choosing the bandwidths.

3. Choice of bandwidths

3.1. EDGE CORRECTIONS

In Section 2.2, our theory was based on the assumption that we have observations from the distributions f and g both in the interval I and also outside I, but that we are only interested in the ratio of f and g in the interval. In practice, it is usually the case that we only have observations lying in our interval of interest, although the densities are positive outside I. Density estimates, and hence relative risk estimates, constructed from such data tend to be distorted by edge effects. The edge correction we shall use, introduced by Diggle (1985), gives an adjusted estimator

$$\tilde{f}_h(x) = \hat{f}_h(x)/q_h(x),$$

where $q_h(x) = \int_I h^{-1} K\{h^{-1}(x-u)\}\,\mathrm{d}u$. This is equivalent to using a position-dependent kernel K_x, such that $\int_I h^{-1} K_x\{h^{-1}(x-u)\}\,\mathrm{d}u = 1$ for all $x \in I$, which eliminates the bias of order 1. The kernel could be further constrained to remove the bias of order h, but this is not at all straightforward compared with the relatively simple first-order correction given above. Other suggestions for reducing boundary bias are contained in Marron and Ruppert (1994). The edge-corrected density estimate will not strictly integrate to unity on I, although the discrepancy is usually small enough to ignore. When the sample size is small or h is large, the size of the discrepancy can be noticeable. In all

our examples we apply the edge-correction followed by a rescaling of the density estimate so that it integrates to unity.

3.2. A CROSS-VALIDATION PRESCRIPTION

Suppose that f and g are probability density functions defined within an interval $I \subseteq \mathbb{R}$. Also suppose that $\hat{f}_{h_1}$ and $\hat{g}_{h_2}$ are estimates constructed from samples x_i, $i = 1, \ldots, n_1$, and y_j, $j = 1, \ldots, n_2$, using smoothing parameters h_1 and h_2, respectively. To choose h_1 and h_2 for estimation of a smooth functional $\gamma(f, g)$ of f and g on I, a reasonable procedure is to choose the values which minimize the integrated square error (ISE):

$$\mathrm{ISE}\{\gamma(\hat{f}_{h_1}, \hat{g}_{h_2})\} = \int_I [\gamma\{\hat{f}_{h_1}(x), \hat{g}_{h_2}(x)\} - \gamma\{f(x), g(x)\}]^2 \, \mathrm{d}x.$$

This is equivalent to minimizing

$$\begin{aligned} C(h_1, h_2) = & \int_I [\gamma\{\hat{f}_{h_1}(x), \hat{g}_{h_2}(x)\}]^2 \, \mathrm{d}x \\ & - 2 \int_I \gamma\{\hat{f}_{h_1}(x), \hat{g}_{h_2}(x)\} \gamma\{f(x), g(x)\} \, \mathrm{d}x, \end{aligned} \tag{3.1}$$

since the omitted term does not depend on h_1 or on h_2. A first-order Taylor approximation to $\gamma(f, g)$ gives

$$\begin{aligned} \gamma(f, g) \simeq & \gamma(\hat{f}_{h_1}, \hat{g}_{h_2}) + (f - \hat{f}_{h_1}) \gamma^{\alpha}(\hat{f}_{h_1}, \hat{g}_{h_2}) \\ & + (g - \hat{g}_{h_2}) \gamma^{\beta}(\hat{f}_{h_1}, \hat{g}_{h_2}), \end{aligned} \tag{3.2}$$

where $\gamma^{\alpha}(a, b) = \partial\gamma(a, b)/\partial a$ and $\gamma^{\beta}(a, b) = \partial\gamma(a, b)/\partial b$. Substituting equation (3.2) into (3.1), we obtain

$$\begin{aligned} C(h_1, h_2) \simeq & - \int_I [\gamma\{\hat{f}_{h_1}(x), \hat{g}_{h_2}(x)\}]^2 \, \mathrm{d}x + 2 \int_I Q_{\alpha}\{\hat{f}_{h_1}(x), \hat{g}_{h_2}(x)\} \hat{f}_{h_1}(x) \, \mathrm{d}x \\ & + 2 \int_I Q_{\beta}\{\hat{f}_{h_1}(x), \hat{g}_{h_2}(x)\} \hat{g}_{h_2}(x) \, \mathrm{d}x - 2 \int_I Q_{\alpha}\{\hat{f}_{h_1}(x), \hat{g}_{h_2}(x)\} f(x) \, \mathrm{d}x \\ & - 2 \int_I Q_{\beta}\{\hat{f}_{h_1}(x), \hat{g}_{h_2}(x)\} g(x) \, \mathrm{d}x, \end{aligned} \tag{3.3}$$

where $Q_{\alpha}(a, b) = \gamma(a, b)\gamma^{\alpha}(a, b)$ and similarly $Q_{\beta}(a, b) = \gamma(a, b)\gamma^{\beta}(a, b)$.

The last two terms in (3.3) are expectations with respect to the unknown densities f and g. We therefore estimate these two expectations by the usual method of 'leave-one-out' averaging. The method then reduces to choosing h_1 and h_2 to minimize $\hat{C}(h_1, h_2)$ which is the same as the approximation of $C(h_1, h_2)$ in (3.3) except that the last two terms are replaced by

$$-2n_1^{-1} \sum_{i=1}^{n_1} Q_{\alpha}\{\hat{f}_{h_1}^{-i}(x_i), \hat{g}_{h_2}(x_i)\} - 2n_2^{-1} \sum_{j=1}^{n_2} Q_{\beta}\{\hat{f}_{h_1}(y_j), \hat{g}_{h_2}^{-j}(y_j)\},$$

where $\hat{f}_{h_1}^{-i}$ is a density estimate of f constructed from all the data points except x_i, and similarly $\hat{g}_{h_2}^{-j}$ is a density estimate of g constructed from all the data points except y_j.

When $\gamma\{f(x), g(x)\} = \log\{f(x)/g(x)\} = \rho(x)$, this reduces to

$$\hat{C}(h_1, h_2) = -\int_I \{\hat{\rho}_{h_1,h_2}(x)\}^2 \, dx - 2n_1^{-1} \sum_{i=1}^{n_1} \log\{\hat{f}_{h_1}^{-i}(x_i)/\hat{g}_{h_2}(x_i)\}/\hat{f}_{h_1}^{-i}(x_i)$$

$$+ 2n_2^{-1} \sum_{j=1}^{n_2} \log\{\hat{f}_{h_1}(y_j)/\hat{g}_{h_2}^{-j}(y_j)\}/\hat{g}_{h_2}^{-j}(y_j).$$

A special case of this method results if we fix $h_1 = h_2 = h$ a priori. We have seen that this restriction has some theoretical justification when $f = g$, and will show that it works well in practice when $f \simeq g$. Note also that when $\gamma\{f(x), g(x)\} = f(x)$ the method reduces to ordinary least-squares cross-validation as proposed independently by Rudemo (1982) and Bowman (1984).

3.3. OTHER METHODS

In order to assess the performance of the cross-validation prescription introduced in Section 3.2, we need to compare it with other possible methods. One such method is to choose smoothing parameters, h_{S_1} and h_{S_2}, which separately optimize the estimates of $f(x)$ and $g(x)$ on the interval I, using standard least-squares cross-validation. Another is to choose h_{S_1} and h_{S_2} to optimize the estimates of $\log f(x)$ and $\log g(x)$ on I. This can be accomplished by letting $\gamma(f, g) = \log f$, and $\gamma(f, g) = \log g$, to obtain h_{S_1} and h_{S_2}, respectively, using the cross-validation method described in Section 3.2. Adjusted versions of these methods make direct use of the theoretically derived equations (2.10), by estimating the values of B_{ff}, B_{fg} and B_{gg} from the density estimates $\hat{f}_{h_{S_1}}$ and $\hat{g}_{h_{S_2}}$ and substituting these values into (2.10) to obtain adjusted bandwidths h_{J_1} and h_{J_2}. Another method for choosing a bandwidth is likelihood cross-validation (Habbema *et al.* 1974; Duin 1976), which usually performs well except that it tends to be very sensitive to outliers, creating over-smoothed density estimates. In our case, this may be an advantage since we assume that our underlying densities are bounded away from zero on the interval of interest, and we require the density estimates to have the same property in order to give a sensible ratio estimate. As before, the resulting bandwidths h_{S_1} and h_{S_2} can be adjusted via (2.10) to give bandwidths h_{J_1} and h_{J_2}.

Recent research in bandwidth choice for kernel density estimates suggests that cross-validation can be outperformed by various 'plug-in' methods, which essentially seek to estimate the term $\int f''(x)^2 \, dx$ in equation (2.3) (see, for example, Park and Marron 1990; or Sheather and Jones 1991). However, these methods are not obviously adaptable to the situation in which observations are made only on a finite interval, nor do they take account of the edge corrections which are necessary in most epidemiological applications.

3.4. A COMPARATIVE SIMULATION STUDY

We now present the results of a simulation study to compare the performances of different methods for choosing the bandwidths. In the tabulation of the results, the methods are identified as

follows:

1 new cross-validation method;
2 new cross-validation method with constraint $h_1 = h_2$;
3 least-squares cross-validation;
4 least-squares cross-validation with adjustment;
5 likelihood cross-validation;
6 likelihood cross-validation with adjustment;
7 logarithmic cross-validation;
8 logarithmic cross-validation with adjustment.

Note that in the above list, the phrase 'with adjustment' refers to the use of equations (2.10) as described in Section 3.3.

We consider the unit interval, $I = (0, 1)$, three 'denominator' densities, $g_1(x) = 1$, $g_2(x) = 1 + 0.5\sin(2\pi x)$, $g_3(x) = 1 + 0.75\sin(4\pi x)$, and three relative risk functions, $r_1(x) = 1$, $r_2(x) = 1.46\phi(x; 0.5, 0.5)$, $r_3(x) = 0.919 + 0.081\phi(x; 0.5, 0.05)$, where $\phi(x; \mu, \sigma)$ denotes the normal density with mean μ and variance σ^2. Thus r_1 represents uniform risk, r_2 a slow global variation in risk, and r_3 a local 'blip' of increased risk. Combinations of these give nine 'numerator' densities, $f_{11}(x) = r_1(x)\,g_1(x)$, $f_{21}(x) = r_2(x)\,g_1(x)$, etc. For each numerator density f and its associated denominator density g, we generated pseudo-random samples of sizes $n_1 = n_2 = 200$, which we then used to estimate $\rho(x) = \log\{f(x)/g(x)\}$ using each of the eight methods for choosing the bandwidths. In each case, we simulated 20 replicates and calculated the median integrated square errors (ISEs) for each method. The median, rather than the mean, was used because all eight methods occasionally gave rise to an extremely high ISE – in applications, this would be identified as a breakdown of the method and it would then be more sensible to choose the bandwidth subjectively, but this option is not available in a simulation study. We then repeated the whole procedure with sample sizes $n_1 = 50$ and $n_2 = 400$. The results are given in Table 1, including the value of B_{fg} for each case, the importance of which was discussed in Section 2.2. For all cross-validation methods, there is a possibility that an infinite smoothing parameter will be chosen, corresponding to a uniform density on the interval I. This explains why we sometimes obtain zero ISE in the case where $f = g$.

The results suggest that the new cross-validation method is an improvement on the other methods where the denominator density is either g_2 or g_3 but has similar, or worse, performance when it is g_1, the uniform density. This is consistent with the theory given in Section 2.2. When either f or g is uniform, the B_{fg} term in equations (2.10) is zero, implying that jointly chosen bandwidths confer no advantage. In practice it is unlikely that we shall be dealing with near-uniform densities f or g, so this should not be a serious problem. Note also that in these cases, the median ISE is very small for all eight methods of bandwidth choice.

Methods 3, 5 and 7 have similar overall performances in our examples, although we found that the logarithmic method 7 more often gave extremely high ISEs. The adjustment formulae (methods 4, 6 and 8) nearly always improve the estimates, but a greater improvement is usually obtained by using the joint cross-validation procedure (method 1). The best results of all are obtained by using the joint cross-validation method with bandwidths constrained to be equal (method 2). This last remark applies to cases with unequal sample sizes, $n_1 = 50$ and $n_2 = 400$, as well as when $n_1 = n_2 = 200$. However, and as would be expected on theoretical grounds, we have found that

Table 1. Median integrated squared errors of estimates of $\log\{f(x)/g(x)\}$ (20 simulations), for various f, g, n_1 and n_2

Densities	f_{11} and g_1		f_{12} and g_2		f_{13} and g_3	
B_{fg}	0		358.78		35708.38	
n_1	200	50	200	50	200	50
n_2	200	400	200	400	200	400
Method 1	0.00203	0.00447	0.00225	0.00785	0.00459	0.00460
2	0.00038	0.00105	0	0	0	0
3	0.00032	0.00221	0.02403	0.08808	0.08039	0.12473
4	0.00023	0.00221	0.01725	0.08605	0.06459	0.09567
5	0.00031	0.00223	0.02443	0.09070	0.09026	0.18570
6	0.00023	0.00223	0.01619	0.08825	0.07732	0.12312
7	0.00031	0.00217	0.03006	0.07907	0.11467	0.33954
8	0.00023	0.00217	0.01648	0.07636	0.07909	0.28846

Densities	f_{21} and g_1		f_{22} and g_2		f_{23} and g_3	
B_{fg}	0		314.25		35076.71	
n_1	200	50	200	50	200	50
n_2	200	400	200	400	200	400
Method 1	0.02196	0.02532	0.02850	0.03434	0.02472	0.02458
2	0.02236	0.02393	0.02236	0.02236	0.02236	0.02244
3	0.01744	0.02410	0.03679	0.07322	0.08254	0.11708
4	0.01793	0.02265	0.02888	0.06777	0.06983	0.08191
5	0.01804	0.02425	0.03559	0.07730	0.08874	0.14522
6	0.01804	0.02272	0.02684	0.07133	0.07051	0.10954
7	0.01838	0.02309	0.03253	0.07005	0.07430	0.22937
8	0.01838	0.02284	0.02721	0.06167	0.05704	0.20112

Densities	f_{31} and g_1		f_{32} and g_2		f_{33} and g_3	
B_{fg}	0		473.31		40156.17	
n_1	200	50	200	50	200	50
n_2	200	400	200	400	200	400
Method 1	0.02267	0.03520	0.02605	0.02634	0.06013	0.03665
2	0.02260	0.02709	0.02262	0.02267	0.03769	0.02277
3	0.02473	0.04151	0.04129	0.08480	0.09677	0.18956
4	0.02473	0.04151	0.03236	0.07790	0.07717	0.14582
5	0.02314	0.03058	0.03858	0.09514	0.11708	0.18566
6	0.02314	0.03058	0.03358	0.08827	0.10548	0.16073
7	0.02263	0.02748	0.03362	0.10464	0.11809	0.40143
8	0.02263	0.02645	0.02694	0.09369	0.10223	0.38350

method 2 can give poor results when the sample sizes are very different. Note also that the relative benefit of choosing the bandwidths jointly tends to be larger when the sample sizes are different, since all of the available data are then used in choosing both h_1 and h_2. The relative benefits also tend to be large when the densities f and g are rapidly fluctuating, as is the case with g_3, since the value of B_{fg} is then relatively large.

3.5. RECOMMENDATIONS

The simulation results indicate that the new method of cross-validation is the preferred method for bandwidth choice in kernel estimation of $\rho(x)$. If we expect the underlying densities to be nearly equal, as is often the case in epidemiological applications, we also recommend constraining the bandwidths to be equal, so as to reduce bias.

4. Tolerance intervals

Suppose that we have obtained an estimate, $\hat{\rho}_0(x)$ say, using data $x_1, \ldots, x_{n_1}$ from a density f and data $x_{n_1+1}, \ldots, x_{n_1+n_2}$ from a density g, both on an interval I. A pointwise tolerance interval can be constructed to indicate, for each x, the range of values of $\hat{\rho}(x)$ which are consistent with a hypothesis H: $\rho(x) = \rho_H(x)$, say. Under H, the data can be viewed as a single sample from a density $(n_1 f + n_2 g)/(n_1 + n_2)$ followed by independent, random allocations of points to the 'f group' and 'g group' with position-dependent probabilities $p(x)$ and $1 - p(x)$, respectively, where $p(x) = n_1 \exp\{\rho_H(x)\}/[n_1 \exp\{\rho_H(x)\} + n_2]$. A log density ratio estimate which is consistent with H can therefore be created by combining the data, randomly allocating them with these probabilities into two groups of approximate sizes n_1 and n_2 and calculating an estimate, $\hat{\rho}_1(x)$ say, using the same smoothing parameters as for the original estimate $\hat{\rho}_0(x)$. For a 95% tolerance interval, we repeat the random allocation procedure, say s times, at each point x in I, calculate the 2.5 and 97.5 percentiles of the estimates $\hat{\rho}_1(x), \ldots, \hat{\rho}_s(x)$, and plot these on the same graph as the original estimate $\hat{\rho}_0(x)$.

Two hypotheses about $\rho(x)$ are of particular interest. The first is H_0: $\rho(x) = 0$, i.e. a null hypothesis of constant risk. A formal Monte Carlo test of H_0 (Barnard 1963) can be performed using the statistics

$$t_j = \int_I \hat{\rho}_j(x)^2 \, \mathrm{d}x \qquad j = 0, \ldots, s.$$

The p value of the test is $p = (k + 1)/(s + 1)$, where k is the number of $t_j > t_0$.

The second hypothesis of particular interest is H: $\rho(x) = \hat{\rho}_0(x)$, since the resulting interval estimates the sampling variability in $\hat{\rho}_0(x)$.

5. Application

To analyse the data from Fig. 1, we convert the locations to squared distances (up to a distance of

20 km) from the incinerator, and use the resulting samples of size 387 (larynx cancers) and 7672 (lung cancers) to obtain an estimate of $\rho(\cdot)$. Use of squared distances avoids the problem of trying to estimate $\rho(x)$ near $x = 0$ when both $f(x) \to 0$ and $g(x) \to 0$ as $x \to 0$. For ease of interpretation, we then present the relative risk estimate as a function of distance from the incinerator, rather than squared distance. If there is no association between risk and distance we would expect the cases and controls to behave as a random sample from a common underlying distribution, and the relative risk estimate to be approximately constant. Note that although the circle of radius 20 km centred on the incinerator does not lie entirely within the study region, this does not bias the estimation provided that the relative risk is solely attributable to distance from the incinerator. For bandwidth choice, we use the cross-validation procedure constrained to have equal bandwidths. The resulting estimate $\hat{\rho}(x)$ is shown in Fig. 2, with pointwise 95% tolerance intervals for $\hat{\rho}(x)$ both when $\rho(x) = 0$ (Fig. 2a) and when $\rho(x) = \hat{\rho}(x)$ (Fig. 2b), in each case based on 1000 random reallocations. A Monte Carlo test of constant risk gave a p value of 0.075. Note that although there is an apparent local increase in relative risk close to the incinerator, this is not significant, whereas there does appear to be some genuine and unexpected fluctuation in risk between 13 km and 18 km away.

Once we accept the idea that the spatial variation in relative risk cannot be attributed exclusively to the incinerator, it no longer makes sense to analyse the data solely in terms of distances from the incinerator. We are developing analogous methodology for estimating two-dimensional spatial variation in relative risk, and will report the results in due course.

6. Discussion

A natural extension of this approach is to the estimation of spatial variation in relative risk when there is no prior hypothesis of association with a specific location. The spatial method is a direct analogue of the one-dimensional method described in the present paper. The kernel estimator uses a radially symmetric kernel, and the cross-validation calculations involve double integrals over a target region A. In principle, A may be any spatial region, but the computations are eased if A is rectangular.

A completely different approach to the problem of risk estimation is to use nonparametric binary regression. As our starting point, we again consider a set of points x_i, $i = 1, \ldots, n_1$, arising as a partial realization of a Poisson process on an interval I, with intensity $\lambda_1(x)$, and a second set of points y_j, $j = 1, \ldots, n_2$, arising as an independent realization of a Poisson process on I with intensity $\lambda_2(x)$. We wish to estimate the relative risk, $r(x) = \lambda_1(x)/\lambda_2(x)$, with a view to detecting and describing departures from constant risk. If we now think of the data as a single set of locations, x_i, $i = 1, \ldots, n$, where $n = n_1 + n_2$, together with a binary label for each point, then conditional on the points x_i, the labels z_i, $i = 1, \ldots, n$, are mutually independent Bernoulli random variables with $P(Z_i = 1) = p(x_i)$, where

$$p(x) = \lambda_1(x)/\{\lambda_1(x) + \lambda_2(x)\} = r(x)/\{1 + r(x)\}.$$

Then, $p(x)$ and hence $r(x)$ can be estimated by a binary regression of the z_i on the x_i (Hastie and Tibshirani 1990, Section 4.5). In effect, this is equivalent to transforming the functional parameter space from $\lambda_1(x)$ and $\lambda_2(x)$ to $r(x)$ and $s(x) = \lambda_1(x) + \lambda_2(x)$, and conditioning on the x_i which are

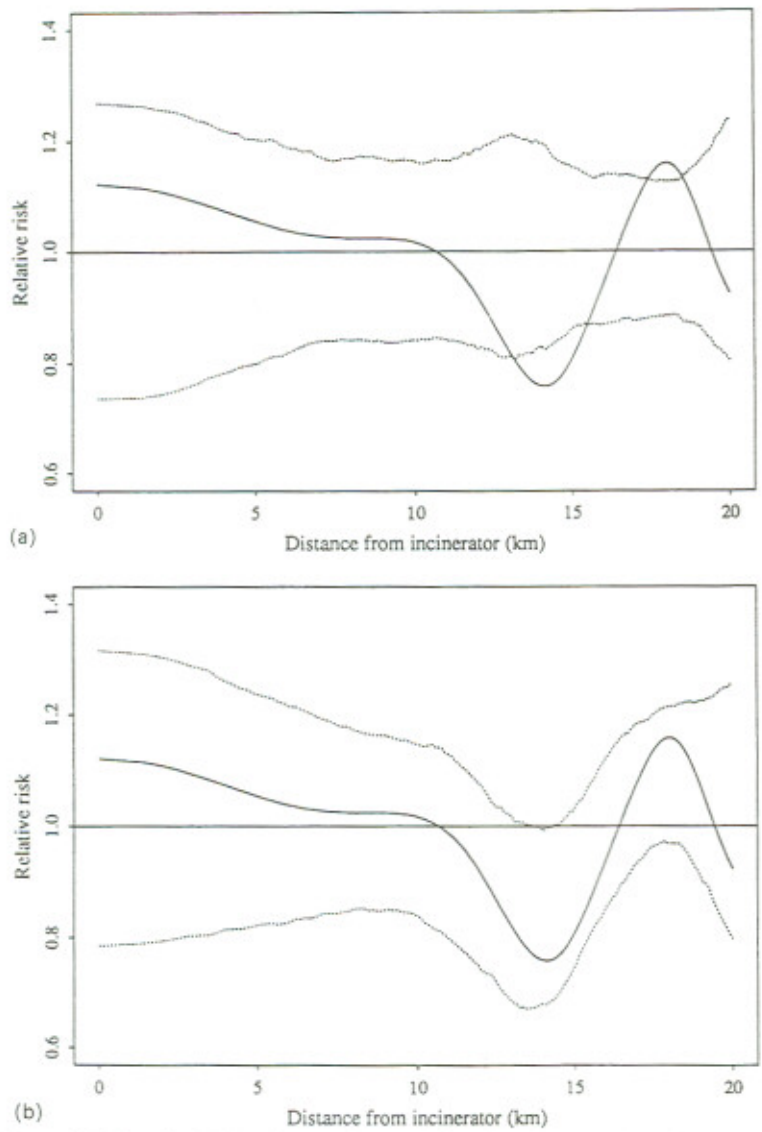

Figure 2. Pointwise 95% tolerance intervals for the estimate of relative risk computed from the Chorley-Ribble data, as a function of distance from the incinerator. (a) Tolerance interval when $\rho(x) = 0$; (b) tolerance interval when $\rho(x) = \hat{\rho}(x)$

sufficient statistics for the nuisance parameter $s(x)$. The same idea was used in a parametric setting in Diggle and Rowlingson (1994).

Results for the spatial extension, and for the binary regression formulation, will be reported in due course.

Acknowledgements

JEK was financially supported by an EPSRC studentship. The cancer data were provided by Tony Gatrell (Department of Geography, Lancaster University).

References

Barnard, G.A. (1963) Contribution to the discussion of Professor Bartlett's paper. *J. Roy. Statist. Soc. Ser. A*, **25**, 294.

Bithell, J.F. (1990) An application of density estimation to geographical epidemiology. *Statistics in Medicine*, **9**, 691–701.

Bithell, J.F. (1992) Statistical methods for analysing point-source exposures. In P. Elliott, J. Cuzick, D. English and R. Stern (eds), *Geographical and Environmental Epidemiology: Methods for Small Area Studies*, pp. 221–230. Oxford: Oxford University Press.

Bowman, A.W. (1984) An alternative method of cross-validation for the smoothing of density estimators. *Biometrika*, **71**, 353–360.

Bowman, A.W. (1985) A comparative study of some kernel-based nonparametric density estimates. *J. Statist. Comput. Simulation*, **21**, 313–327.

Diggle, P.J. (1985) A kernel method for smoothing point process data. *Appl. Statist.*, **34**, 138–147.

Diggle, P.J. (1990) A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. *J. Roy. Statist. Soc. Ser. A*, **153**, 349–362.

Diggle, P.J. and Rowlingson, B.S. (1994) A conditional approach to point process modelling of raised incidence. *J. Roy. Statist. Soc. Ser. A.*, **157**, 433–440.

Duin, R.P.W. (1976) On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Comput.*, **C-25**, 1175–1179.

Elliot, P., Hills, M., Beresford, J., Kleinschmidt, I., Jolley, D., Pattenden, S., Rodrigues, L., Westlake, A. and Rose, G. (1992) Incidence of cancer of the larynx and lung near incinerators of waste solvents and oils in Great Britain. *Lancet*, **339**, 854–858.

Habbema, J.D.F., Hermans, J. and van der Broek, K. (1974) A stepwise discrimination program using density estimation. In G. Bruckman (ed.), *Compstat 1974*, pp. 100–110. Vienna: Physica Verlag.

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*. New York: Chapman & Hall.

Marron, J.S. and Ruppert, D. (1994) Transformations to reduce boundary bias in kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, **56**, 653–671.

Park, B.-U. and Marron, J.S. (1990) Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.*, **85**, 66–72.

Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, **9**, 65–78.

Sheather, S.J. and Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, **53**, 683–690.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Stone, R.A. (1988) Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in Medicine*, **7**, 649–660.

Received May 1994 and revised September 1994