Computational Statistics and Data Analysis 101 (2016) 12–28

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Symmetric adaptive smoothing regimens for estimation of the spatial relative risk function

Tilman M. Davies [a,*], Khair Jones [b], Martin L. Hazelton [b]

[a] Department of Mathematics & Statistics, University of Otago, Dunedin, New Zealand

[b] Institute of Fundamental Sciences—Statistics, Massey University, Palmerston North, New Zealand

ARTICLE INFO

Article history:
Received 13 February 2015
Received in revised form 1 February 2016
Accepted 12 February 2016
Available online 23 February 2016

Keywords:
Bandwidth
Case-control data
Kernel smoothing
Log-relative risk function
Pilot estimation
Simulation study

ABSTRACT

The spatial relative risk function is now regarded as a standard tool for visualising spatially tagged case-control data. This function is usually estimated using the ratio of kernel density estimates. In many applications, spatially adaptive bandwidths are essential to handle the extensive inhomogeneity in the distribution of the data. Earlier methods have employed separate, asymmetrical smoothing regimens for case and control density estimates. However, we show that this can lead to potentially misleading methodological artefacts in the resulting estimates of the log-relative risk function. We develop a symmetric adaptive smoothing scheme that addresses this problem. We study the asymptotic properties of the new log-relative risk estimator, and examine its finite sample performance through an extensive simulation study based on a number of problems adapted from real life applications. The results are encouraging.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The spatial relative risk function was introduced by Bithell (1990) to describe geographical variation in disease risk. It is now regarded as a standard tool for visualisation of spatially tagged case-control data, and also for case–case data when comparison between diseases or disease strains is of interest. Recent applications in human and veterinary epidemiology include Firestone et al. (2012), Zhang et al. (2013), Jaros et al. (2013), Denzin et al. (2013), Mweu et al. (2014) and Lemke et al. (2015). Use of the relative risk function has also spread to other areas that involve comparison of spatial distributions. For example, Howe and Leiserowitz (2013) apply this tool to study the geographical variation in binary responses to a survey on global warming, Davies et al. (2013) examine the spatial dispersion of fast- and slow-twitch muscle fibres in human fascicles, and Smith et al. (2015) analyse patterns in different types of archaeological finds in a particular excavation site.

The spatial relative risk function is defined as follows. Given any point $\boldsymbol{x}$ in some compact geographical region $\mathcal{R} \subset \mathbb{R}^2$ of interest, the relative risk at $\boldsymbol{x}$ is given by

$$r(\boldsymbol{x}) = \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \tag{1}$$

where f and g are respectively the probability density functions for the locations of disease cases and controls over $\mathcal{R}$. Noting that f and g are strictly positive, it is usual to work with the log-relative risk $\rho(\boldsymbol{x}) = \log\{r(\boldsymbol{x})\}$ for data visualisation, in part because of the resultant symmetrisation of the way in which the two densities f and g are handled.

* Corresponding author.
E-mail address: tdavies@maths.otago.ac.nz (T.M. Davies).

http://dx.doi.org/10.1016/j.csda.2016.02.008
0167-9473/© 2016 Elsevier B.V. All rights reserved.

In practice $r(\boldsymbol{x})$ must be estimated from spatially stamped case-control data (we use the case-control terminology henceforth, on the understanding that the methods described may equally well be applied to case–case or similarly structured data). Given the complexity of the geographical distribution of most human and animal populations, it is natural to estimate $r(\boldsymbol{x})$ nonparametrically.

By far the most popular approach has been to replace the unknown densities f and g in (1) by kernel density estimates thereof. See Bithell (1990), Kelsall and Diggle (1995) for example. However, results using standard fixed bandwidth techniques have been disappointing in many applications (Davies and Hazelton, 2010). The problem is that the spatial distributions of many populations of interest are extremely heterogeneous, for example with wide variations in density between rural areas and urban centres. Fixed bandwidth kernel estimators are typically incapable of simultaneously providing sufficiently high levels of smoothing so as to iron out stochastic troughs and valleys in areas of low density (which typically account for the majority of most study regions) while permitting adequate resolution of details in cities.

Davies and Hazelton (2010) addressed this problem by using adaptive bandwidth kernel estimators. Specifically, they advocated the use of Abramson's spatially adaptive bandwidth selector (Abramson, 1982) to automatically adjust the amount of smoothing employed across the geographical region of interest. Abramson's methodology has much to commend it. The bandwidth varies in a very natural way, taking larger values in areas where the data are most sparse and smaller values in places where the data are numerous. Moreover, density estimates using this type of adaptive bandwidth have improved theoretical properties in comparison to fixed bandwidth competitors, with a reduction in the magnitude of the asymptotic bias.

By allowing different adaptive smoothing schemes for case and control density estimation, Davies and Hazelton (2010) aimed to harness both the intuitive appeal of Abramson's adaptive bandwidths and their theoretical benefits. Their empirical results were very encouraging, with the adaptive log-relative risk estimator having substantially lower integrated squared error than the fixed bandwidth analogue, and also producing more aesthetically pleasing image plots of the log-relative risk on a real data application. This last point is important, because the primary use of estimates of the log-relative risk function has been for data visualisation. In this context, absence of noisy artefacts in the estimated function in areas of low density is important to many users, as is the ability of the methodology to faithfully highlight areas of truly increased risk.

In order to be a robust data visualisation tool which is easily interpretable by users, it is crucial that the estimated relative risk function is not subject to methodological artefacts that could lead to misinterpretation of the data. Unfortunately it transpires that the spatially adaptive log-relative risk estimator proposed in Davies and Hazelton (2010) suffers from this exact problem. As an illustration, consider the plots in Fig. 1. The top left-hand panel displays the control density g; the top right panel the true log-relative risk function ρ, and the bottom left-hand panel an estimate thereof obtained using Davies and Hazelton's (2010) estimator applied to simulated data. The method was implemented using the **sparr** package in R (R Development Core Team, 2015; Davies et al., 2011). An epidemiologist viewing the estimated relative risk function might well postulate that there is a risk factor associated with the outer suburbs of the two population centres on the left, yet the halos of increased risk are merely a subtle artefact of the methodology.

The root cause of artefacts such as those in Fig. 1 is that at most locations in the region of interest we are using a different local bandwidth. This asymmetry is inevitable if we aim for the optimal adaptive smoothing scheme for both case and control densities, unless $f(\boldsymbol{x}) = g(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{R}$. We postpone until Section 2.3 a more detailed explanation of why the adaptive kernel estimator of Davies and Hazelton (2010) is prone to producing halo risk effects. For now we simply note that the methodological artefacts in Fig. 1 have not arisen because the example under study is particularly perverse. Similar problems will be quite common in practical applications.

A simple antidote is to employ identical adaptive smoothing regimens to both case and control density estimates. The bottom right-hand panel demonstrates the resulting elimination of the risk halos. We describe how this methodology can be implemented in Section 2.4. Ostensibly the price to pay for a symmetric smoothing scheme is that the local bandwidths will not be asymptotically optimal, and so the theoretical bias gains mentioned above will not be realised. However, a closer analysis suggests that things are not quite so clear cut on the theoretical front. In many applications we will expect the relative risk function to be constant over much of the region, with only small areas of heightened risk. Symmetric smoothing is asymptotically optimal where r is locally constant, so the theoretical benefits of asymmetrical smoothing will be entirely based upon estimation in the areas of elevated risk. We explore the consequences in Section 3 using a non-standard asymptotic analysis.

Extensive numerical studies on both simulated and real data indicate that a symmetric smoothing regimen returns log-relative risk estimates with smaller integrated squared error than those using asymmetric smoothing in applications with small areas of elevated risk around a few scattered point sources, in line with our theoretical analysis. More unexpectedly, we also see that for sample sizes of $O(1000)$, symmetric smoothing is at least as good as asymmetric smoothing in examples with more extensive geographical variation in risk. These studies are documented in Sections 4 and 5.

2. Adaptive kernel estimation

2.1. Adaptive kernel density estimation

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ denote a random sample from some bivariate density f. The adaptive (or variable bandwidth) kernel density estimator of f is given by

Download English Version:

https://daneshyari.com/en/article/6869007

Download Persian Version:

https://daneshyari.com/article/6869007

Daneshyari.com