Computational Statistics and Data Analysis 103 (2016) 56–67

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bandwidth selection for kernel log-density estimation

Martin L. Hazelton *, Murray P. Cox

Massey University, New Zealand

ARTICLE INFO

Article history:
Received 19 January 2016
Received in revised form 12 April 2016
Accepted 5 May 2016
Available online 11 May 2016

Keywords:
Approximate likelihood inference
Kernel smoothing
Mean squared error
Simulation
Smooth cross-validation

ABSTRACT

Kernel estimation of the logarithm of a probability density function at a given evaluation point is studied. The properties of the kernel log-density estimator are heavily influenced by the unboundedness of the log function at zero. In particular, standard asymptotic expansions can provide a poor guide to finite sample behaviour for this estimator, with consequences for the choice of methodology for bandwidth selection. In response, a new approximate cross-validation bandwidth selector is developed. Its theoretical properties are explored and its finite sample behaviour examined in numerical experiments. The proposed methodology is then applied to estimation of log-likelihoods for a complex genetic model used in determining migration rates between village communities on the Indonesian island of Sumba.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The probability density function is fundamental to a huge swathe of statistical theory and methods. However, in many cases the density occurs naturally on the log-scale. A critical example is likelihood theory, where it is the log-likelihood rather than likelihood that plays a predominant role. Other instances of the use of the log-density include elements of information criteria and entropy functions (e.g. Hall and Morton, 1993) and the log-relative risk function in spatial epidemiology (Bithell, 1990, 1991).

In this paper we consider nonparametric estimation of log-densities. The log-density function is estimated directly in some approaches to density estimation, including the maximum penalized likelihood method (Silverman, 1982), local likelihood density estimation (Loader, 1996) and spline-based methods (e.g. O'Sullivan, 1988). Nonetheless, an attractive alternative is simply to log-transform a standard kernel density estimate, at least in part because of the success of kernel density estimation in practice and the vast theory that exists on this topic. We adopt this approach.

The choice of the bandwidth smoothing parameter is critical in kernel density estimation. It has been the subject of much research effort, with reliable methods now available for both univariate (Sheather and Jones, 1991) and multivariate (Duong and Hazelton, 2003) settings. However, bandwidth selection for log-density estimation poses some interesting problems, stemming primarily from the unboundedness of the logarithmic function at zero. An immediate consequence is that it is impossible to select a globally optimal bandwidth (to minimize mean integrated squared error, for example) for any log-density with infinite support. One alternative would be to consider estimation of the log-density over some compact set. We go further, and focus on bandwidth selection for estimation of the log-density at a single point. Nonetheless, even then the asymmetric shape of the logarithmic curve has major ramifications for bandwidth selection when working with log-densities.

* Correspondence to: Institute of Fundamental Sciences, Massey University, Private Bag 11222, Palmerston North, New Zealand. Tel.: +64 6 356 9099; fax: +64 63557953.

E-mail address: m.hazelton@massey.ac.nz (M.L. Hazelton).

http://dx.doi.org/10.1016/j.csda.2016.05.003
0167-9473/© 2016 Elsevier B.V. All rights reserved.

A naive application of standard asymptotic methods fails to adequately capture the asymmetry in the mean squared error (as a function of bandwidth) for log-density estimation, as we show in Section 2. As a consequence, plug-in type bandwidth selectors, which target the asymptotic form of the error function, are unlikely to perform well. A natural response is to employ bootstrap (e.g. Faraway and Jhun, 1990; Hazelton, 1996) or cross-validation (e.g. Duong and Hazelton, 2005) bandwidth selection techniques. These aim to estimate the exact mean squared error function, which is then minimized to derive a bandwidth. However, these methods become unacceptably computationally expensive unless the estimated bias can be written in closed form. While this is possible for density estimates on the raw scale, it is not the case for log-densities. We therefore introduce a new approximate smooth cross-validation methodology in Section 3, which retains computationally efficiency but nonetheless still captures the crucial features of the mean squared error as a function of the bandwidth.

An interesting application of kernel log-density estimation is approximate likelihood inference (ALI): that is, methods involving approximate log-likelihood functions constructed using simulations from complex models where the likelihood is not available analytically. See Diggle and Gratton (1984). Bandwidth selection for ALI is concerned with minimization of the error in estimating a sum of log-densities evaluated as some specific observed values. Our work does not directly address this general problem, being focussed on estimation at a single point. Nonetheless, it provides insight. Moreover, it is immediately applicable in cases where the observed data comprise a single multivariate observation.

We provide an example of this type in Section 4, where a complex model is applied to genetic data from settlements on the island of Sumba, eastern Indonesia, to investigate possible associations between migration rates and language clusters. The raw observed dataset is huge, comprising genome wide information on dozens of individuals. However, much of the information on migration between any given pair of communities is thought to be captured by the fixation index, a scalar measure of the relative genetic similarity between them. Our observed data are hence condensed to a single bivariate observation comprising the fixation indices between two pairs of settlements. For any given value of the migration rates we can employ a complex stochastic model to generate simulations of the fixation indices, from which the log-likelihood can be estimated. We find that the observed data are sufficient to permit useful estimates of migration rates, and to construct an approximate likelihood ratio test statistic to examine whether migration rates are affected by language differences between settlements.

2. Kernel log-density estimation

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a d-dimensional random sample, each drawn from a density function f. Our kernel estimate of f is defined by

$$\hat{f}_h(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i) \tag{1}$$

where $K_h(\boldsymbol{x}) = h^{-d}K(\boldsymbol{x}/h)$ is a scaled kernel, the unscaled kernel function K is a radially symmetric density function with $\int K(\boldsymbol{u})\|\boldsymbol{u}\|^2\, d\boldsymbol{u} = d$, and h is the bandwidth. Eq. (1) employs isotropic smoothing, and hence requires that the variables either be on a comparable scale to begin with, or are pre-scaled. More general smoothing regimens are possible, including use of a full bandwidth matrix to specify both scaling and orientation of the kernel functions (Wand and Jones, 1993, 1994). However, we are interested in estimating f at some prespecified point $\boldsymbol{x}$, so that the bandwidth can (and should) be a function of $\boldsymbol{x}$. It is well known that bandwidth selection for local density estimation is more challenging than global bandwidth selection (e.g. Hazelton, 1996), and that selection of full global bandwidth matrices is very challenging for dimensions $d \geq 4$ without very large sample sizes (Duong and Hazelton, 2005). These comments explain the authors' experience when conducting some preliminary numerical tests, where attempts to select full local bandwidth matrices produced highly variable and unreliable results. We therefore prefer to work with the kernel density estimator in Eq. (1), where we need select only a local scalar bandwidth $h = h(\boldsymbol{x})$.

Estimation of the log-density $\psi(\boldsymbol{x}) = \log(f(\boldsymbol{x}))$ can be done using $\log(\hat{f}_h(\boldsymbol{x}))$. However, an immediate problem with this naive estimator is that it may not have finite moments. Indeed, if K is on compact support, then for any finite bandwidth h, $\mathsf{E}[\log(\hat{f}_h(\boldsymbol{x}))]$ is undefined. To see this, simply note that $\mathbb{P}(\hat{f}_h(\boldsymbol{x}) = 0) > 0$ under those conditions. This problem can be addressed by making a small adjustment to the naive estimator. Specifically, we consider henceforth the modified log-density estimator $\hat{\psi}_h(\boldsymbol{x}) = \log(\hat{f}_h(\boldsymbol{x}) + e^{-n})$. This has finite moments of all orders, a result that we formalize in Theorem 1. What is more, in most practical cases involving a suitable bandwidth, the effect of the adjustment will be negligible for all but tiny sample sizes. It follows that the modification is primarily for theoretical purposes. In practical situations where n is sufficiently small for the adjustment to have a potentially tangible effect, the data should be prescaled (for example, to have unit standard deviation in each coordinate direction), or the estimator revised to $\hat{\psi}_h(\boldsymbol{x}) = \log(\hat{f}_h(\boldsymbol{x}) + e^{-n}/s)$ for some scalar measure of spread s, in order to ensure scale invariance. Such a change has no effect on any of the theoretical results that we present.

Theorem 1. *Assume that*

(A1) *K is a radially symmetric probability density function with $\int K(\boldsymbol{u})\|\boldsymbol{u}\|^2\, d\boldsymbol{u} = d$ and $\int K(\boldsymbol{u})\|\boldsymbol{u}\|^4\, d\boldsymbol{u} < \infty$.*
(A2) *All partial derivatives of f up to and including order 2 are continuous in a neighbourhood of $\boldsymbol{x}$.*

Download English Version:

https://daneshyari.com/en/article/415265

Download Persian Version:

https://daneshyari.com/article/415265

Daneshyari.com