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 $\boldsymbol{x}$.*

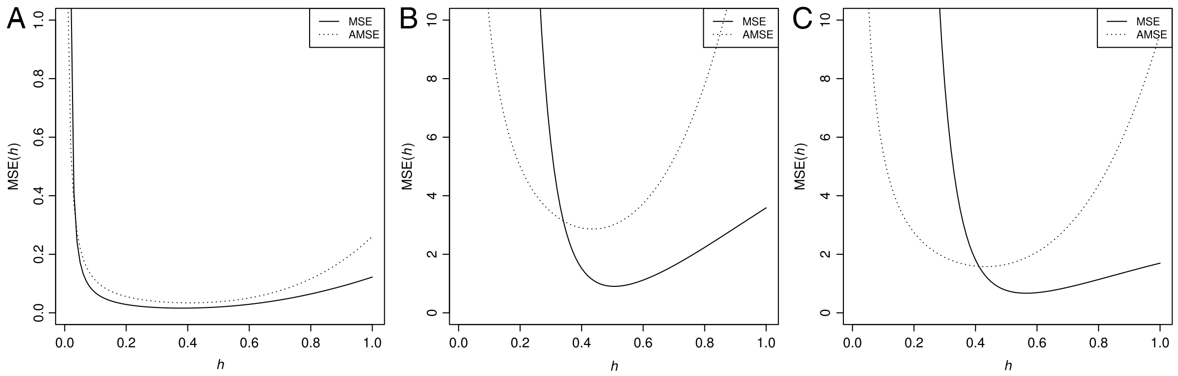

Fig. 1. Comparison of exact and asymptotic versions of the mean squared error of $\hat{\psi}_h(\boldsymbol{x})$ for three problems. For panels (A) and (B) the target density is standard normal, estimated at points $x = 0$ and $x = 3$ respectively. For panel (C) the target density is a bivariate standard normal, estimated at $\boldsymbol{x} = (2, 2)^\mathsf{T}$. In each case the estimate is computed from a sample of size $n = 100$.

(A3) $f(\boldsymbol{x}) > 0$.
(A4) $h = O(n^{-1/(d+4)})$.

Then $|\hat{\psi}_h(\boldsymbol{x})|^M$ *is uniformly integrable for any positive integer* M, *and hence* $\hat{\psi}_h(\boldsymbol{x})$ *has finite moments of all orders for all* $n > 0$.

The proof appears in Appendix A.

Our aim is to select the bandwidth h so as to minimize the mean squared error,

$$\begin{aligned}\mathsf{MSE}\left(\hat{\psi}_h(\boldsymbol{x})\right) &= \mathsf{E}\left[\left(\hat{\psi}_h(\boldsymbol{x}) - \psi(\boldsymbol{x})\right)^2\right] \\ &= \mathsf{Bias}(\hat{\psi}_h(\boldsymbol{x}))^2 + \mathsf{Var}(\hat{\psi}_h(\boldsymbol{x})).\end{aligned}$$

Given sufficient regularity conditions on f and K, a standard asymptotic expansion gives

$$\mathsf{MSE}\left(\hat{\psi}_h(\boldsymbol{x})\right) = \mathsf{AMSE}\left(\hat{\psi}_h(\boldsymbol{x})\right) + o(h^4 + n^{-1}h^{-d})$$

with

$$\mathsf{AMSE}\left(\hat{\psi}_h(\boldsymbol{x})\right) = \frac{h^4}{4}\frac{(\nabla^2 f(\boldsymbol{x}))^2}{f(\boldsymbol{x})^2} + \frac{R(K)}{nh^d f(\boldsymbol{x})} \tag{2}$$

where $R(g) = \int g(\boldsymbol{x})^2\, d\boldsymbol{x}$ for any square integrable function g. This is minimized by the asymptotically optimal bandwidth

$$h_{as} = \left(\frac{dR(K)f(\boldsymbol{x})}{(\nabla^2 f(\boldsymbol{x}))^2}\right)^{1/(d+4)} n^{-1/(d+4)}. \tag{3}$$

It can be shown that $\mathsf{AMSE}(\hat{\psi}_h(\boldsymbol{x})) = \mathsf{AMSE}(\hat{f}(\boldsymbol{x}))/f(\boldsymbol{x})^2$, where $\mathsf{AMSE}(\hat{f}(\boldsymbol{x}))$ is the asymptotic mean squared error for the raw density estimate. This implies that the asymptotically optimal bandwidth is the same whether working with the density on the log or raw scale. Nevertheless, while the asymptotics provide a reliable guide to finite sample behaviour for estimation of f, this is not the case when estimating the log-density, particularly when the estimation point $\boldsymbol{x}$ lies in an area of low density. These remarks are particularly pertinent for applications to approximate likelihood inference, where we may expect observed data to lie far into the tails of the log-density when working with simulations generated at parameter values some distance from the maximum likelihood estimate.

As an illustration, we compare in Fig. 1 plots of $\mathsf{AMSE}(\hat{\psi}_h(\boldsymbol{x}))$ and $\mathsf{MSE}(\hat{\psi}_h(\boldsymbol{x}))$ (computed from 10^5 simulated datasets) for a univariate standard normal model, with log-density estimation at $x = 0$ and $x = 3$ based on a sample of size $n = 100$, and a bivariate standard normal model with log-density estimation at $\boldsymbol{x} = (2, 2)^\mathsf{T}$. Clearly the asymptotic version of the mean squared error provides a serviceable approximation only in the first case. What is more, it is very apparent that $\mathsf{AMSE}(\hat{\psi}_h(\boldsymbol{x}))$ fails to reflect the large errors that are likely to occur when using too small a bandwidth. We can therefore expect a data-driven bandwidth selector based on asymptotic theory to produce very poor results in a significant fraction of cases.

3. Approximate smooth cross-validation bandwidth selection

3.1. Method and theory

The previous discussion motivates the use of a bandwidth selector that seeks explicitly to minimize the exact rather than asymptotic version of $\mathsf{MSE}(\hat{\psi}_h(\boldsymbol{x}))$. In fact the asymptotic approximation of the variance term is usually adequate, with the problems arising primarily because of inaccuracies in the representation of the bias. For global density estimation on the raw scale, arguments of this type led Hall et al. (1992) to introduce the idea of smooth cross-validation for bandwidth selection. In the context of local density estimation, the smooth cross-validation estimate of $\mathsf{MSE}(\hat{f}(\boldsymbol{x}))$ is defined by

$$\begin{aligned}\mathsf{SCV}_f(h) &= \left(\mathsf{E}^\dagger[f_h^\dagger(\boldsymbol{x})] - \hat{f}_\lambda(\boldsymbol{x})\right)^2 + \frac{\hat{f}_\lambda(\boldsymbol{x})R(K)}{nh^d} \\ &= \left(\hat{f}_\lambda * K_h(\boldsymbol{x}) - \hat{f}_\lambda(\boldsymbol{x})\right)^2 + \frac{\hat{f}_\lambda(\boldsymbol{x})R(K)}{nh^d}. \qquad (4)\end{aligned}$$

Here $f_h^\dagger(\boldsymbol{x})$ denotes a kernel density estimate constructed using a random sample of size n drawn from pilot density $\hat{f}_\lambda$, and $\mathsf{E}^\dagger$ indicates expectation with respect to this sampling scheme but conditional on the original data. The symbol $*$ denotes a convolution. The squared bias term (first term on the right-hand side) in Eq. (4) is a smoothed bootstrap estimate, but importantly this can be evaluated in closed form, rendering the methodology computationally efficient. The bandwidth selector $\hat{h}_f$ is obtained by minimizing $\mathsf{SCV}_f(h)$, typically using the smallest local minimum in cases where there are multiple turning points.

A direct adaptation of smooth cross-validation to the estimator $\hat{\psi}_h(\boldsymbol{x})$ gives

$$\begin{aligned}\mathsf{SCV}(h) &= \left(\mathsf{E}^\dagger[\hat{\psi}_h^\dagger(\boldsymbol{x})] - \log(\hat{\psi}_\lambda(\boldsymbol{x}))\right)^2 + \frac{R(K)}{\hat{f}_\lambda(\boldsymbol{x})nh^d} \\ &= \left(\mathsf{E}^\dagger[\log((f_h^\dagger(\boldsymbol{x}) + e^{-n})/\hat{f}_\lambda(\boldsymbol{x}))]\right)^2 + \frac{R(K)}{\hat{f}_\lambda(\boldsymbol{x})nh^d} \qquad (5)\end{aligned}$$

where $\hat{\psi}_h^\dagger(\boldsymbol{x}) = \log(f_h^\dagger(\boldsymbol{x}) + e^{-n})$. However, the squared bias term cannot be evaluated in closed form in this case. We therefore propose an approximate smooth cross-validation (ASCV) criterion,

$$\mathsf{ASCV}(h) = \left(\log((\hat{f}_\lambda * K_h(\boldsymbol{x}) + e^{-n})/\hat{f}_\lambda(\boldsymbol{x}))\right)^2 + \frac{R(K)}{\hat{f}_\lambda(\boldsymbol{x})nh^d}. \qquad (6)$$

This is obtained by simply taking the expectation operator $\mathsf{E}^\dagger$ from (5) inside the logarithm. The result is that $\mathsf{ASCV}(h)$ only approximates $\mathsf{SCV}(h)$, but it is straightforward to show that the approximation error is of order $O_p(n^{-1}h^{2-d})$ which is asymptotically negligible. Our idea is that despite the approximation, $\mathsf{ASCV}(h)$ will nonetheless capture the important characteristics of $\mathsf{MSE}(h)$, and in particular will reflect the large errors associated with overly small bandwidths.

The approximate smooth cross-validation bandwidth selector, $\hat{h}$, is the minimizer of $\mathsf{ASCV}(h)$. As is standard with cross-validation procedures, we minimize this function over an interval $[\gamma n^{-1/(d+4)}, \gamma^{-1}n^{-1/(d+4)}]$ where $\gamma > 0$ is some small constant. This ensures that $\hat{h}$ has the optimal asymptotic order.

The properties of $\hat{h}$ depend on the choice of λ, the bandwidth used to construct the pilot (bootstrap) density $\hat{f}_\lambda$. Guidance on this matter is provided by the following theorem, which describes in terms of λ the relative mean squared error of $\hat{h}$ as an estimator of the asymptotically optimal bandwidth h_{as}.

Theorem 2. *Let $\hat{h}$ be the approximate SCV bandwidth selector for the estimator $\hat{\psi}_h(\boldsymbol{x})$, obtained using pilot bandwidth λ. Assume* (A1) *and* (A3) *from* Theorem 1, *and in addition*

(A5) *All partial derivatives of K up to and including order 4 are bounded.*

(A6) *All partial derivatives of f up to and including order 4 are bounded in a neighbourhood of $\boldsymbol{x}$.*

(A7) $\lambda = \alpha n^\beta$ *for* $\alpha > 0$ *and* $-\frac{1}{d+6} < \beta < -\frac{1}{d+10}$.

Then

$$\mathsf{E}\left[\left(\frac{\hat{h} - h_{as}}{h_{as}}\right)^2\right] = \frac{1}{(d+4)^2}\left[\lambda^4\left(\frac{\Theta(\boldsymbol{x})}{\nabla^2 f(\boldsymbol{x})} - \frac{\nabla^2 f(\boldsymbol{x})}{2f(\boldsymbol{x})}\right)^2 + 4\frac{f(\boldsymbol{x})}{(\nabla^2 f(\boldsymbol{x}))^2}\frac{R(\nabla^2 K)}{n\lambda^{d+4}}\right] + o(\lambda^4 n^{-1}\lambda^{-d-4})$$

where

$$\Theta(\boldsymbol{x}) = \sum_{i=1}^{d}\frac{\partial^4}{\partial x_i^4}f(\boldsymbol{x}) + \sum_{i=1}^{d}\sum_{i\neq j\, j=1}^{d}\frac{\partial^2}{\partial x_i^2}\frac{\partial^2}{\partial x_j^2}f(\boldsymbol{x}).$$

The proof appears in Appendix B.

Two immediate corollaries are as follows.

Corollary 1. *Under the assumption that* $2\Theta(\boldsymbol{x})f(\boldsymbol{x}) \neq (\nabla^2 f(\boldsymbol{x}))^2$*, the value of the pilot bandwidth to minimize* $\mathsf{E}[((\hat{h} - h_{as})/h_{as})^2]$ *is*

$$\lambda_0 = \left[\frac{4(d+4)f(\boldsymbol{x})^3 R(\nabla^2 K)}{\left(2\Theta(\boldsymbol{x})f(\boldsymbol{x}) - (\nabla^2 f(\boldsymbol{x}))^2\right)^2} \right]^{1/(d+8)} n^{-1/(d+8)}. \tag{7}$$

Corollary 2. *Using a pilot bandwidth* $\lambda \propto n^{-1/(d+8)}$ *gives the optimal rate of convergence for the bandwidth selector:* $(\hat{h} - h_{as})/h_{as} = O_p(n^{-2/(d+8)})$.

To use the result from Corollary 1 to choose a pilot bandwidth in practice, we must estimate the density functionals therein. In principle this could involve a number of stages of pilot estimation. Specifically, each functional of f can be estimated using kernel methods. The bandwidth at each stage will depend on further functionals of f, which at some point can be replaced by reference values. Our preference is to employ (normal) reference bandwidths immediately, with any partial derivative of f of order r estimated using bandwidth

$$b_r = \left(\frac{4}{d+2r+2} \right)^{1/(d+2r+4)} s n^{-1/(d+2r+4)}. \tag{8}$$

In Eq. (8), s is a scalar estimate of spread. If the data have been pre-sphered the this can be computed using the standard deviation computed over all variables. Otherwise, s can be set to the geometric mean of the standard deviations for all variables. A robust alternative (used in all the numerical examples presented in this paper) is to replace standard deviations by $\mathsf{IQR}/1.34$, where IQR denotes the interquartile range of the variable in question.

Our justification for using reference bandwidths for functional estimation in (7) is twofold. First, the accuracy of the estimation does not affect the asymptotic order of $(\hat{h} - h_{as})/h_{as}$. Second, experience in other smooth cross-validation local bandwidth selection problems (e.g. Hazelton, 1996) suggests that multiple stages of pilot estimation are often counter-productive. In essence, local bandwidth selection is a challenging problem, more so than global bandwidth selection in which multiple pilot stages are often used. (This is reflected in part by the slower rates of convergence in the local case.) In particular, the bias–variance trade-off for local bandwidth selections appears to place a premium on controlling variability. This matches our experience in some preliminary testing using multiple pilot stages.

3.2. Numerical results

In this section we report on a simulation study to compare the performance of the ASCV bandwidth selector $\hat{h}$ with the selector $\hat{h}_f$ based on minimization of $\mathsf{SCV}_f(h)$ from Eq. (4). As discussed previously, first order asymptotic analysis indicates that $\hat{h}$ and $\hat{h}_f$ should produce log-density estimates $\hat{\psi}_h(\boldsymbol{x})$ with comparable properties for large sample sizes, but the former selector was designed to better capture the form of the $\mathsf{MSE}(h)$ curve for modest values of n.

A total of six test densities are considered. For the univariate case these are standard normal, standard exponential and a t-distribution on 4 degrees of freedom. For the bivariate case we consider a standard normal; a normal distribution with zero mean vector, unit variance in both coordinate directions and correlation coefficient 0.8; and a standard bivariate t-distribution on 4-degrees of freedom. Each of these distributions is estimated at 4 points, one of which is the median (univariate case) or mean (bivariate case), and two of which lie in the tails of the distribution. Three samples sizes are employed: $n = 30, 100, 400$. For each scenario (i.e. combination of test density, estimation point and sample size) a total of 500 datasets are generated. Kernel log-density estimates are then obtained using the two bandwidth selectors, and the squared errors recorded in each case. The results (in terms of squared error on the log-scale) are displayed using boxplots in Figs. 2–7.

There is a very consistent pattern to the results. For log-density estimation at points of (relatively) high density, there is little to choose between the bandwidth selector $\hat{h}$ and $\hat{h}_f$. However, when estimating in the tails, the approximate SCV selector $\hat{h}$ enjoys a substantial advantage in many cases. Moreover, even in situations where the median squared error is quite similar for the two selectors, $\hat{h}$ avoids the excessively large errors that are sometimes encountered when using $\hat{h}_f$. This matches our expectations, in that the approximate SCV approach is more reliably reflecting the likely errors that will arise when the bandwidth is too small.

4. Approximate likelihood inference for a model of migration in Sumba using genetic data

4.1. Overview

In this section we examine an example based on genome-wide data collected from groups of about 30 individuals in several village communities on the island of Sumba in eastern Indonesia. See Guillot et al. (2015). These villages are quite

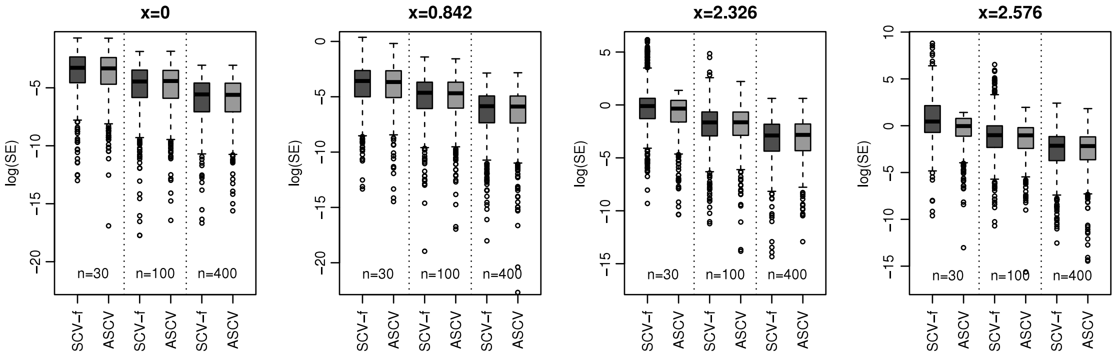

Fig. 2. Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV). The test density is standard normal, and the estimation point indicated in the header to each panel. Results for different sample sizes are separated by vertical lines, and labelled towards the bottom of each plot.

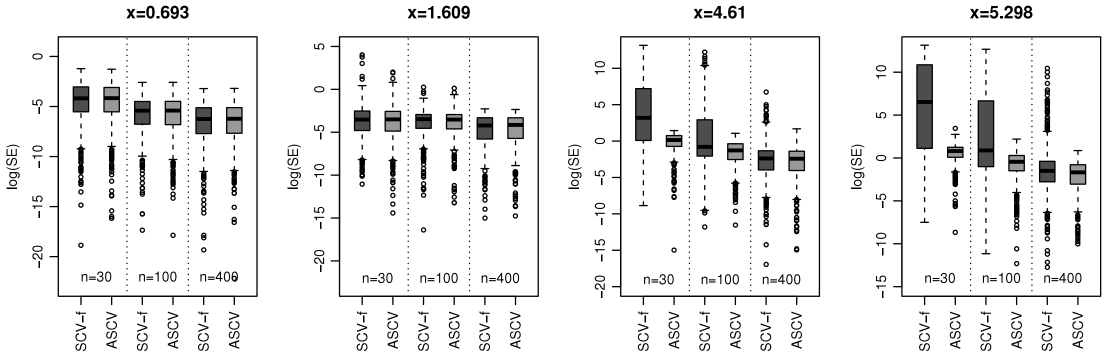

Fig. 3. Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV). The test density is standard exponential, and the estimation point indicated in the header to each panel. Results for different sample sizes are separated by vertical lines, and labelled towards the bottom of each plot.

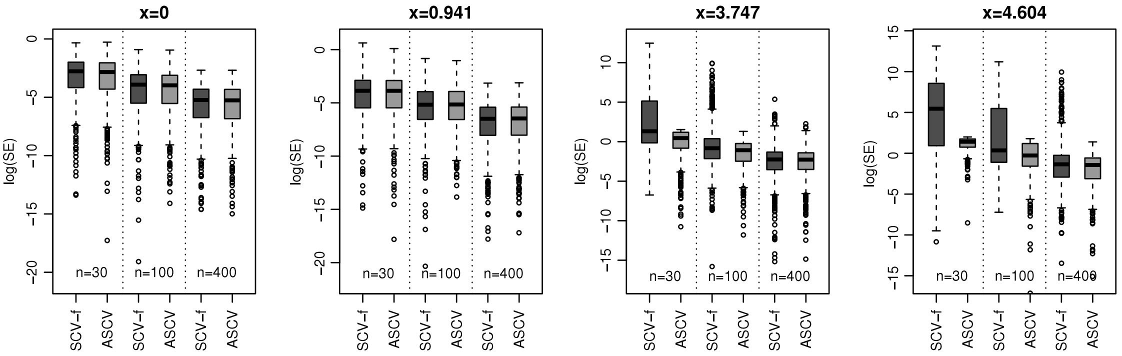

Fig. 4. Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV). The test density is t_4, and the estimation point indicated in the header to each panel. Results for different sample sizes are separated by vertical lines, and labelled towards the bottom of each plot.

isolated, and so the populations therein have evolved reasonably independently. There are nine languages spoken on Sumba, which can be broken down into five language clusters (Lansing et al., 2007). Our focus here is on two pairs of villages and the rates of migration between them. The first pair is Mamboro and Wanokaka, whose inhabitants speak languages in the same language cluster. The second pair comprises villages, Loli and Kodi, from two different language clusters. Apart from language, the two village pairs are very comparable. Our particular interest is in the rates of within-pair migration, taking into account the geographical distances between the villages, and in particular, whether the migration rates are influenced by language differences.

We have available a complex stochastic model, based on a structured coalescent (Chen et al., 2009), that describes the genetic profile of each community as a function of the migration rates between villages in each pair. The migration rate for pair i is described by a parameter θ_i, which represents the proportion of genes that move between the two populations in

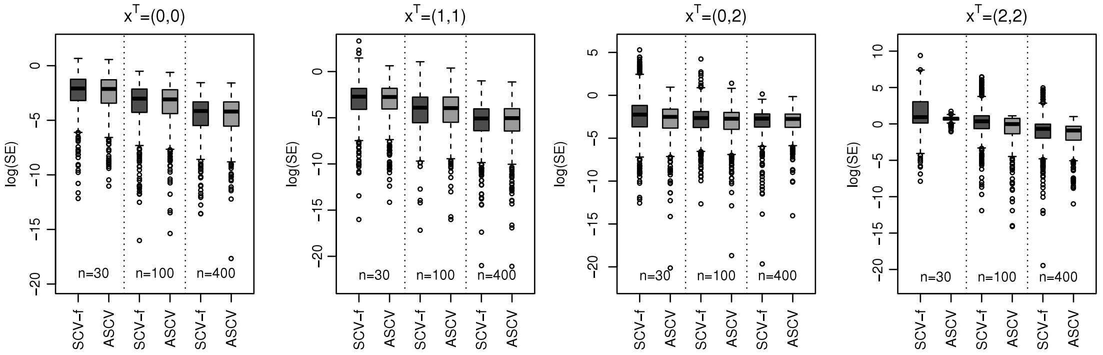

Fig. 5. Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV). The test density is standard bivariate normal, and the estimation point indicated in the header to each panel. Results for different sample sizes are separated by vertical lines, and labelled towards the bottom of each plot.

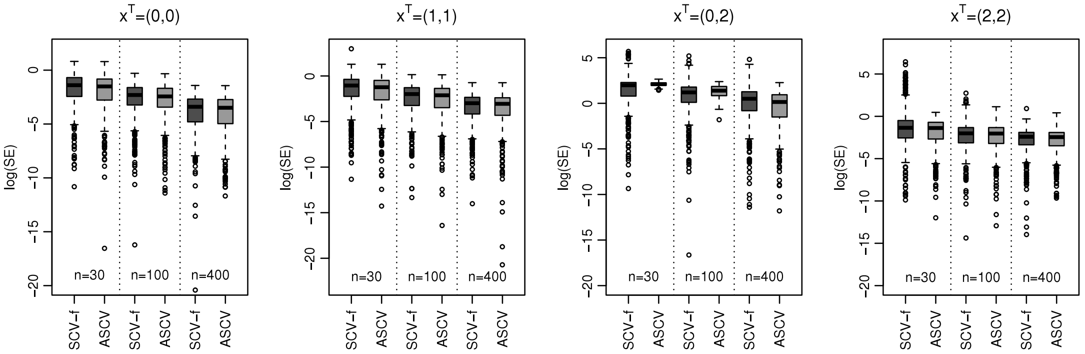

Fig. 6. Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV). The test density is a correlated bivariate normal ($\rho = 0.8$) with zero mean and unit variance for each variable, and the estimation point indicated in the header to each panel. Results for different sample sizes are separated by vertical lines, and labelled towards the bottom of each plot.

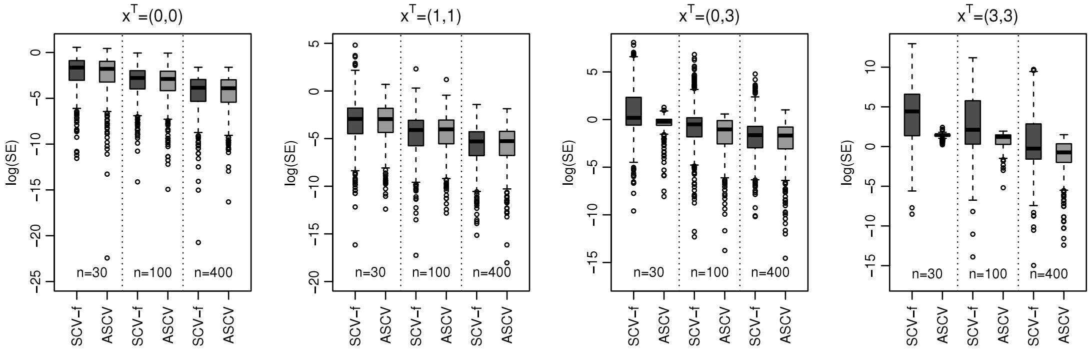

Fig. 7. Log-squared error for log-density estimates using smooth cross-validation for density estimation (SCV-f) and approximate smooth cross-validation for log-density estimation (ASCV). The test density is a bivariate standard t on 4 degrees of freedom, and the estimation point indicated in the header to each panel. Results for different sample sizes are separated by vertical lines, and labelled towards the bottom of each plot.

one generation. The model is mathematically intractable, so that we are unable to obtain an analytical expression for the probability density for the genetic profiles or any (useful) summaries thereof. Nevertheless, we can obtain simulations of the genetic profiles for any given value of the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2)^{\mathsf{T}}$.

One approach to conducting inference for $\boldsymbol{\theta}$ in this kind of setting is approximate Bayesian computation (ABC) (e.g. Marin et al., 2012; Sunnåker et al., 2013). However, here we choose to conduct a Frequentist analysis. The full genome-wide dataset is unwieldy. Instead, we follow a common path when dealing with genetic data and work with a low dimensional summary statistic. Specifically, the genetic data for community pair i ($i = 1, 2$) are condensed to the fixation index F_{ST} (Plagnol and Wall, 2006), a scalar summary statistic that measures the genetic variation within each village as a proportion relative to the total genetic variation. The observed data is the single bivariate observation $\boldsymbol{y} = (0.01412, 0.01259)$ of F_{ST} statistics.

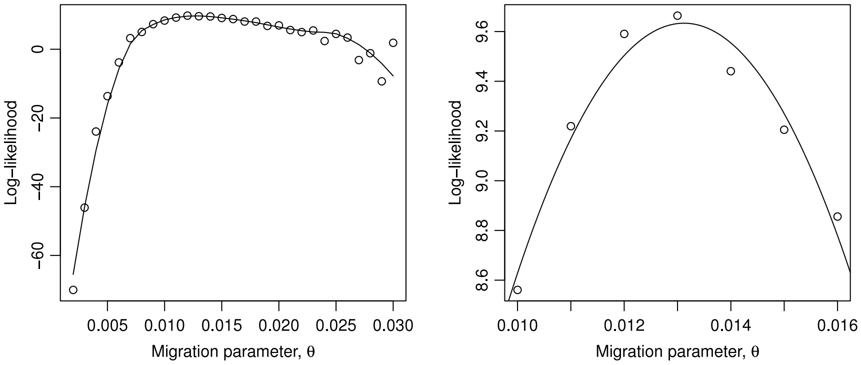

Fig. 8. Estimates of the log-likelihood for the model with common migration rate. The left-hand panel shows estimates based on $n = 500$ simulated datasets for θ on the grid 0.002(0.001)0.03, while the right-hand panel shows results near the maximum based on $n = 2000$ simulated datasets. A loess curve is applied in the left-hand panel, and a quadratic curve fitted in the right-hand panel.

4.2. Approximate likelihood inference methodology

The model log-likelihood is $\ell(\boldsymbol{\theta}) = \log(f(\boldsymbol{y}|\boldsymbol{\theta}))$, where $f(\cdot|\boldsymbol{\theta})$ is the probability density for the fixation indices $\boldsymbol{y}$ at parameter $\boldsymbol{\theta}$. We do not have a mathematically tractable expression for $f(\cdot|\boldsymbol{\theta})$ and hence $\ell(\boldsymbol{\theta})$. Instead we aim to estimate $\ell(\boldsymbol{\theta})$ using simulations $\boldsymbol{x}^{\theta} = (x_1^{\theta}, x_2^{\theta})^{\mathsf{T}}$ of the fixation indices for the two village pairs, following the general methodology proposed by Diggle and Gratton (1984). Rather than using the naive log-density estimator described in Section 2, we estimate the log-likelihood $\ell(\theta)$ by

$$\hat{\ell}(\boldsymbol{\theta}) = \hat{\psi}_h(\boldsymbol{y}|\boldsymbol{\theta}) \tag{9}$$

where $\hat{\psi}_h(\boldsymbol{y}|\boldsymbol{\theta}) = \log(\hat{f}_h(\boldsymbol{y}|\boldsymbol{\theta}) + e^{-n})$ is our modified log-density estimator constructed from the simulations $\boldsymbol{x}^{\theta}$. The bandwidth $h = h_{\boldsymbol{\theta}}$ (to emphasize dependence on $\boldsymbol{\theta}$) is selected using our approximate SCV methodology. We note that the components of $\boldsymbol{y}$ are measured on entirely comparable scales (both being fixation indices), and hence the use of a common bandwidth in both coordinate directions is reasonable.

The entire log-likelihood function (near the maximum) can be obtained by applying a smoother to estimates of $\ell(\boldsymbol{\theta})$ over a grid of parameter values. Calculation of the (approximate) maximum likelihood estimate can proceed by maximizing the fitted smoother.

We can conduct approximate likelihood ratio tests by estimating the statistic $D = -2(\ell(\boldsymbol{\theta}_0) - \ell(\boldsymbol{\theta}))$. We will use the plug-in approximation $\hat{D} = -2(\hat{\ell}(\boldsymbol{\theta}_0) - \hat{\ell}(\boldsymbol{\theta}))$, where $\hat{\ell}(\boldsymbol{\theta}_0)$ is estimated from simulations $\boldsymbol{x}_1^{\boldsymbol{\theta}_0}, \dots, \boldsymbol{x}_n^{\boldsymbol{\theta}_0}$ using bandwidth h_0, and $\hat{\ell}(\boldsymbol{\theta})$ is constructed from simulations $\boldsymbol{x}_1^{\boldsymbol{\theta}}, \dots, \boldsymbol{x}_n^{\boldsymbol{\theta}}$ using bandwidth h. In practice it will be essential to ensure that the simulation induced variability in $\hat{D}$ is controlled sufficiently so that test results are reliable, in the sense of matching those that would have been obtained based on the true likelihood. We can obtain an estimate of the Monte Carlo standard error by making use of the asymptotic approximation

$$\begin{aligned}\mathsf{Var}(\hat{D}|\boldsymbol{y}_1, \dots, \boldsymbol{y}_N) &= \mathsf{Var}(\hat{\psi}_{h_0}(\boldsymbol{y}|\boldsymbol{\theta}_0)) + \mathsf{Var}(\hat{\psi}_h(\boldsymbol{y}|\boldsymbol{\theta})) \\ &\approx \frac{R(K)}{nh_0^2 f(\boldsymbol{y}|\boldsymbol{\theta}_0)} + \frac{R(K)}{nh_0^2 f(\boldsymbol{y}|\boldsymbol{\theta})}.\end{aligned} \tag{10}$$

Specifically, we set the approximate standard error $\hat{\sigma}(\hat{D})$ to be the square root of the previous expression in which the densities in the denominators are replaced by kernel estimates thereof.

4.3. Results

We first consider a model with common migration parameter, $\theta = \theta_1 = \theta_2$. We simulated sets of $n = 500$ pairs of F_{ST} statistics, $\boldsymbol{x}^{\theta} = (x_1^{\theta}, x_2^{\theta})^{\mathsf{T}}$, for θ on the grid 0.002(0.001)0.03. We then applied our log-likelihood estimator from Eq. (9), implemented using ASCV bandwidths. The resulting approximate log-likelihood is displayed in the left-hand panel of Fig. 8, where a loess curve has been applied to give a smooth estimate of the function across the full range of parameter values considered. The right-hand panel of this figure shows estimates based on $n = 2000$ simulated values over a grid 0.010(0.001)0.016 near the maximum. A quadratic curve is fitted to help locate the maximum, giving an approximate maximum likelihood estimate of $\hat{\theta} = 0.0131$.

The second model allows for different migration rates, $\theta_1 \neq \theta_2$, between the first and second pairs of villages. We simulated sets of $n = 500$ pairs of F_{ST} statistics, $\boldsymbol{x}^{\theta} = (x_1^{\theta}, x_2^{\theta})^{\mathsf{T}}$, for $(\theta_1, \theta_2)^{\mathsf{T}}$ on the grid $(0.002(0.001)0.03)^2$. A spline smoothed version of the resultant log-likelihood is displayed in Fig. 9 as a filled contour plot. A further set of $n = 2000$

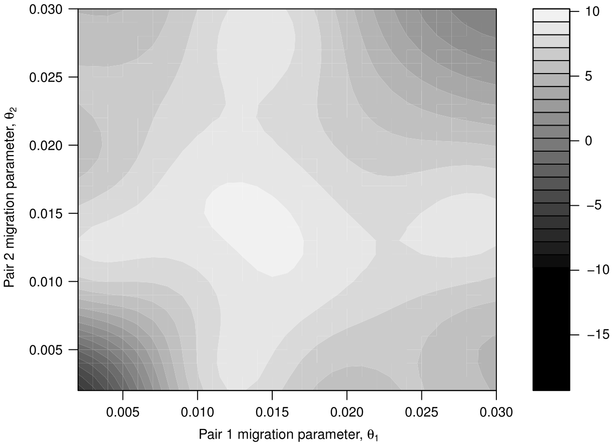

Fig. 9. Smoothed estimate of the log-likelihood surface for the model with differing migration rates. This is based on $n = 500$ simulated datasets for $(\theta_1, \theta_2)^{\mathsf{T}}$ on the grid $(0.002(0.001)0.03)^2$.

simulated datasets were generated over the reduced grid $(0.011(0.001)0.015)^2$. A quadratic response surface was fitted to the results, giving an approximate maximum likelihood estimate of $(\hat{\theta}_1, \hat{\theta}_2)^{\mathsf{T}} = (0.0129, 0.0132)^{\mathsf{T}}$.

Finally, we test the hypothesis $H_0: \theta_1 = \theta_2$ using an approximate likelihood ratio test. This was implemented using $n = 5000$ simulated realizations for both models evaluated at their respective maximum likelihood estimates. The approximate log-likelihood ratio test statistic was $\hat{D} = -0.135$, with an estimated Monte Carlo standard error of $\hat{\sigma}(\hat{D}) = 0.130$. Obviously the true value of D is non-negative, but the negativity of $\hat{D}$ is perfectly explicable in terms of Monte Carlo variation. Equally obviously, we can be very confident that the true test statistic D does not exceed the median (never mind the 95th percentile) of the chi-squared distribution on one degree of freedom. It follows that the data provide no evidence of different migration rates for the two village pairs. The lack of effect of language differences on migration rates is interesting, although not wholly unexpected by some anthropologists (Wilder et al., 2011).

5. Discussion

In this paper we have examined kernel estimation of log-densities at given evaluation points. One common use of kernel log-density estimation has been for evaluating the spatial relative risk function on the log-scale. Despite the significant body of research that now exists on bandwidth selection for that problem (e.g. Kelsall and Diggle, 1995; Hazelton and Davies, 2009; Davies, 2013) this paper is the first to note the problem with standard asymptotic expansions when using the naive log-density estimator.

Approximate likelihood inference (ALI) is another interesting application of kernel log-density estimation. Arguably ALI is a technique that has failed to receive the attention that it deserves. This is particularly unfortunate given the proliferation in recent years of complex models in genetics for which the likelihood function is intractable. One explanation for the rather limited use of approximate likelihood inference is that the technique has been overshadowed by the advent of approximate Bayesian computation (ABC) (e.g. Marin et al., 2012; Sunnåker et al., 2013). Another reason is that the quality of the approximate log-likelihood functions obtained in ALI depend critically on bandwidth selection in the underlying kernel density estimates. Bandwidth selection was handled in a somewhat rudimentary manner in the seminal work of Diggle and Gratton (1984), a perfectly understandable consequence of the immature state of research on bandwidth selection at the time.

A good choice of smoothing parameters is crucial to both approximate Bayesian computation and approximate likelihood inference. Both of these methods work in essence by smoothing over simulated realizations of the model in question. This is explicitly the case in approximate likelihood inference, but is also a valid interpretation of ABC, where a candidate parameter value is accepted only if the observed data (summary statistic) falls within the range of a blurred (i.e. smoothed) simulated value from the model at that parameter value. The degree of smoothing in ABC is controlled by a tolerance parameter describing the width of the blurring. Methods for choosing the value of the tolerance have been examined by a number of researchers (e.g. Beaumont et al., 2002; Ratmann et al., 2009; Beaumont et al., 2009), but this remains an area where further work is needed. In contrast, the problem of selecting the bandwidth to implement approximate likelihood inference can build upon a substantial body of theory on kernel density estimation, including the contributions of this paper.

Acknowledgements

The authors thank three anonymous referees and an Associate Editor for their suggestions, which led to a variety of improvements to this paper. The second author acknowledges funding by way of a Royal Society of New Zealand Marsden grant (11-MAU-007) and a Rutherford Fellowship (RDF-10-MAU-001).

Appendix A. Proof of Theorem 1

Let $\mathbf{1}_A$ denote the indicator of the event A. We have

$$\begin{aligned}
&\lim_{z\to\infty}\sup_n \mathsf{E}[|\log(\hat{f}_h(\boldsymbol{x})+e^{-n})|^M \mathbf{1}_{\{|\log(\hat{f}_h(\boldsymbol{x})+e^{-n})|>z\}}]\\
&\qquad\le \lim_{z\to\infty}\sup_n \max\{\log(h^{-d}K(\mathbf{0})+e^{-n}), n^M\}\mathbb{P}(|\log(\hat{f}_h(\boldsymbol{x})+e^{-n})|>z)\\
&\qquad\le \lim_{z\to\infty}\sup_n n^M\mathbb{P}(|\log(\hat{f}_h(\boldsymbol{x})+e^{-n})|>z).
\end{aligned}$$

Now, $n^{2/(4+d)}\hat{f}_h(\boldsymbol{x}) \xrightarrow{d} N(a,b)$ for some finite constants a and b (see e.g. Prasaka Rao, 1983). We can therefore find n_0 such that for all $n \ge n_0$,

$$\mathbb{P}(|\log(\hat{f}_h(\boldsymbol{x})+e^{-n})|>z) \le \mathbb{P}(|\log(\hat{f}_h(\boldsymbol{x}))|>z) \le e^{-\epsilon(f(\boldsymbol{x})-e^{-z})n^{2/(4+d)}}$$

for some $\epsilon > 0$ by applying a very crude bound on tail probabilities for the normal distribution. Moreover,

$$\lim_{z\to\infty}\sup_{n<n_0} n^M\mathbb{P}(|\log(\hat{f}_h(\boldsymbol{x})+e^{-n})|>z) \le \lim_{z\to\infty} n_0^M H(n_0-z) = 0$$

where H is the Heaviside step function. It follows that

$$\lim_{z\to\infty}\sup_n \mathsf{E}[|\log(\hat{f}_h(\boldsymbol{x})+e^{-n})|^M \mathbf{1}_{\{|\log(\hat{f}_h(\boldsymbol{x})+e^{-n})|>z\}}] \le \lim_{z\to\infty}\sup_n n^M e^{-\epsilon(f(\boldsymbol{x})-e^{-z})n^{2/(4+d)}} = 0,$$

completing the proof.

Appendix B. Proof of Theorem 2

A straightforward asymptotic expansion gives

$$\mathsf{ASCV}(h) = \frac{h^4}{4}\frac{(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))^2}{\hat{f}_\lambda(\boldsymbol{x})^2} + \frac{R(K)}{nh^d f_\lambda(\boldsymbol{x})} + o_p(h^4+n^{-1}h^{-d}).$$

Hence

$$\frac{\partial}{\partial h}\mathsf{ASCV}(h) = \frac{h^3(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))^2}{\hat{f}_\lambda(\boldsymbol{x})^2} - \frac{dR(K)}{nh^{d+1}f_\lambda(\boldsymbol{x})} + o_p(h^3+n^{-1}h^{-d-1}). \tag{B.1}$$

By standard arguments,

$$\frac{\hat{h}-h_{as}}{h_{as}} = -\frac{1}{h_{as}}\left(\frac{\partial/\partial h\mathsf{ASCV}(h_{as})}{\partial^2/\partial h^2\mathsf{AMSE}(h_{as})}\right)(1+o_p(1)). \tag{B.2}$$

Now

$$\frac{\partial}{\partial h}\mathsf{AMSE}(h_{as}) = \frac{3h_{as}^2(\nabla^2 f(\boldsymbol{x}))^2}{f(\boldsymbol{x})^2} + \frac{d(d+1)R(K)}{nh_{as}^{d+2}f(\boldsymbol{x})} + o_p(h_{as}^2+n^{-1}h_{as}^{-d-2}). \tag{B.3}$$

Substituting Eqs. (3), (B.1) and (B.2) into (B.3) gives

$$\begin{aligned}
\frac{\hat{h}-h_{as}}{h_{as}} &= -\frac{f(\boldsymbol{x})^2}{\hat{f}_\lambda(\boldsymbol{x})^2 h_{as}}\left(\frac{h^3(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))^2 - d\hat{f}_\lambda(\boldsymbol{x})R(K)n^{-1}h_{as}^{-d-1}}{3h_{as}^2(\nabla^2 f(\boldsymbol{x}))^2} + d(d+1)f(\boldsymbol{x})R(K)n^{-1}h_{as}^{-d-2}\right)(1+o_p(1))\\
&= -\left(\frac{dR(K)f(\boldsymbol{x})(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))^2(\nabla^2 f(\boldsymbol{x}))^{-2} - d\hat{f}_\lambda(\boldsymbol{x})R(K)}{3dR(K)f(\boldsymbol{x}) + d(d+1)f(\boldsymbol{x})R(K)}\right)(1+o_p(1))\\
&= -\frac{1}{d+4}\left(\frac{(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))^2}{(\nabla^2 f(\boldsymbol{x}))^2} - \frac{\hat{f}_\lambda(\boldsymbol{x})}{f(\boldsymbol{x})}\right)(1+o_p(1)).
\end{aligned} \tag{B.4}$$

Lemma 1. *Under the conditions of Theorem 2,*

$$\mathsf{E}[(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))^2] = (\nabla^2 f(\boldsymbol{x}))^2 + \lambda^2\Theta(\boldsymbol{x})\nabla^2 f(\boldsymbol{x}) + \frac{f(\boldsymbol{x})}{n\lambda^{d+4}}R(\nabla^2 K) + o(\lambda^4+n^{-1}\lambda^{-d-4}).$$

Proof.

$$\mathsf{E}[(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))^2] = \mathsf{E}[(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))]^2 + \mathsf{Var}((\nabla^2\hat{f}_\lambda(\boldsymbol{x}))).$$

By a standard asymptotic expansion,

$$\mathsf{E}[(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))] = \nabla^2 f(\boldsymbol{x}) + \frac{\lambda^2}{2}\Theta(\boldsymbol{x}) + O(\lambda^4)$$

and hence

$$(\mathsf{E}[\nabla^2\hat{f}_\lambda(\boldsymbol{x})])^2 = (\nabla^2 f(\boldsymbol{x}))^2 + \lambda^2\Theta(\boldsymbol{x})\nabla^2(\boldsymbol{x}) + O(\lambda^4).$$

Also,

$$\mathsf{Var}(\nabla^2\hat{f}_\lambda(\boldsymbol{x})) = \frac{f(\boldsymbol{x})R(\nabla^2 K)}{n\lambda^{d+4}} + o(n^{-1}\lambda^{-d-4})$$

when the result of the lemma follows. □

Using the result of Lemma 1 and a standard expansion of $\mathsf{E}[\hat{f}_\lambda(\boldsymbol{x})]$, we obtain from (B.4)

$$\begin{aligned}\mathsf{E}\left[\frac{\hat{h}-h_{as}}{h_{as}}\right] &= -\frac{1}{d+4}\left(\frac{\mathsf{E}[(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))^2]}{(\nabla^2 f(\boldsymbol{x}))^2} - \frac{\mathsf{E}[\hat{f}_\lambda(\boldsymbol{x})]}{f(\boldsymbol{x})}\right) + o(\lambda^2 + n^{-1}\lambda^{-d-4})\\ &= -\frac{1}{d+4}\lambda^2\left(\frac{\Theta(\boldsymbol{x})}{\nabla^2 f(\boldsymbol{x})} - \frac{\nabla^2 f(\boldsymbol{x})}{2f(\boldsymbol{x})}\right) - \frac{1}{d+4}\frac{R(\nabla^2 K)}{n\lambda^{d+4}}\frac{f(\boldsymbol{x})}{(\nabla^2 f(\boldsymbol{x}))^2} + o(\lambda^2 + n^{-1}\lambda^{-d-4})\\ &= -\frac{1}{d+4}\lambda^2\left(\frac{\Theta(\boldsymbol{x})}{\nabla^2 f(\boldsymbol{x})} - \frac{\nabla^2 f(\boldsymbol{x})}{2f(\boldsymbol{x})}\right) + o(\lambda^2)\end{aligned} \tag{B.5}$$

since the last term is asymptotically negligible.

Turning to the variance, it is straightforward to show that

$$\mathsf{Var}\left(\frac{\hat{h}-h_{as}}{h_{as}}\right) = \frac{1}{(d+4)^2}\frac{\mathsf{Var}\left[(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))^2\right]}{(\nabla^2 f(\boldsymbol{x}))^2}(1+o(1)).$$

Now

$$\begin{aligned}\mathsf{Var}[(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))^2] &= 4(\mathsf{E}[\nabla^2\hat{f}_\lambda(\boldsymbol{x})])^2\mathsf{Var}(\nabla^2\hat{f}_\lambda(\boldsymbol{x}))(1+o(1))\\ &= 4(\nabla^2 f(\boldsymbol{x}))^2 f(\boldsymbol{x})\frac{R(\nabla^2 K)}{n\lambda^{d+4}} + o(n^{-1}\lambda^{-d-4})\end{aligned}$$

so that

$$\mathsf{Var}\left(\frac{\hat{h}-h_{as}}{h_{as}}\right) = \frac{4}{(d+4)^2}\frac{f(\boldsymbol{x})}{(\nabla^2 f(\boldsymbol{x}))^2}\frac{R(\nabla^2 K)}{n\lambda^{d+4}} + o(n^{-1}\lambda^{-d-4}). \tag{B.6}$$

The theorem then follows by combining Eqs. (B.5) and (B.6).

References

Beaumont, M., Cornuet, J.M., Marin, J.M., Robert, C., 2009. Adaptive approximate Bayesian computation. Biometrika 96, 983–990.

Beaumont, M., Zhang, W., Balding, D., 2002. Approximate Bayesian computation in population genetics. Genetics 162, 2025–2035.

Bithell, J.F., 1990. An application of density estimation to geographical epidemiology. Stat. Med. 9, 691–701.

Bithell, J.F., 1991. Estimation of relative risk functions. Stat. Med. 10, 1745–1751.

Chen, G., Marjoram, P., Wall, J., 2009. Fast and flexible simulation of DNA sequence data. Genome Res. 19, 136–142.

Davies, T.M., 2013. Jointly optimal bandwidth selection for the planar kernel-smoothed density-ratio. Spat. Spat.-Temporal Epidemiol. 5, 51–65.

Diggle, P.J., Gratton, R.J., 1984. Monte Carlo methods of inference for implicit statistical models (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 46, 193–227.

Duong, T., Hazelton, M.L., 2003. Plug-in bandwidth matrices for bivariate kernel density estimation. J. Nonparametr. Stat. 15, 17–30.

Duong, T., Hazelton, M.L., 2005. Cross-validation bandwidth matrices for multivariate kernel density estimation. Scand. J. Statist. 32, 485–506.

Faraway, J., Jhun, M., 1990. Bootstrap choice of bandwidth for density estimation. J. Amer. Statist. Assoc. 85, 1119–1122.

Guillot, E., Hazelton, M., Karafet, T., Lansing, J., Sudoyo, H., Cox, M., 2015. Relaxed observance of traditional marriage rules allows social connectivity without loss of genetic diversity. Mol. Biol. Evol. 32, 2254–2262.

Hall, P., Marron, J., Park, B., 1992. Smoothed cross-validation. Probab. Theory Related Fields 92, 1–20.

Hall, P., Morton, S.C., 1993. On the estimation of entropy. Ann. Inst. Statist. Math. 45, 69–88.

Hazelton, M., 1996. Bandwidth selection for local density estimators. Scand. J. Statist. 23, 221–232.

Hazelton, M.L., Davies, T.M., 2009. Inference based on kernel estimates of the relative risk function in geographical epidemiology. Biom. J. 51, 98–109.

Kelsall, J.E., Diggle, P.J., 1995. Kernel estimation of relative risk. Bernoulli 1, 3–16.

Lansing, J.S., Cox, M.P., Downey, S.S., Gabler, B.M., Hallmark, B., Karafet, T.M., Norquest, P., Schoenfelder, J., Sudoyo, H., Watkins, J., Hammer, M., 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. Proc. Natl. Acad. Sci. 104, 16022–16026.

Loader, C., 1996. Local likelihood density estimation. Ann. Statist. 24, 1602–1618.

Marin, J.M., Pudlo, P., Robert, C., Ryder, R., 2012. Approximate Bayesian computational methods. Stat. Comput. 22, 1167–1180.

O'Sullivan, F., 1988. Fast computation of fully automated log-density and log-hazard estimators. SIAM J. Sci. Stat. Comput. 9, 363–379.

Plagnol, V., Wall, J., 2006. Possible ancestral structure in human populations. PLos Genet. 2, e105.

Prasaka Rao, B., 1983. Nonparametric Functional Estimation. Academic Press, London.

Ratmann, O., Andrieu, C., Wiuf, C., Richardson, S., 2009. Model criticism based on likelihood-free inference, with an application to protein network evolution. Proc. Natl. Acad. Sci. 106, 10576–10581.

Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B Stat. Methodol. 53, 683–690.

Silverman, B.W., 1982. On the estimation of a probability density function by the maximum penalized likelihood method. Ann. Statist. 795–810.

Sunnåker, M., Busetto, A., Numminen, E., Corander, J., Foll, M., Dessimoz, C., 2013. Approximate Bayesian computation. PLoS Comput. Biol. 9, e1002803.

Wand, M.P., Jones, M.C., 1993. Comparison of smoothing parameterizations in bivariate kernel density estimation. J. Amer. Statist. Assoc. 88, 520–528.

Wand, M.P., Jones, M.C., 1994. Multivariate plug-in bandwidth selection. Comput. Statist. 9, 97–116.

Wilder, J., Cox, M., Paquette, A., Alford, R., Satyagraha, A., Harahap, A., Sudoyo, H., 2011. Genetic continuity across a deeply divergent linguistic contact zone in North Maluku, Indonesia. BMC Genet. 12, 100.