Bernoulli **8**(4), 2002, 423–449

Nonlinear kernel density estimation for binned data: convergence in entropy

GORDON BLOWER* and JULIA E. KELSALL**

*Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK. E-mail: *g.blower@lancaster.ac.uk; **julia.kelsall@mayo.edu*

A method is proposed for creating a smooth kernel density estimate from a sample of binned data. Simulations indicate that this method produces an estimate for relatively finely binned data which is close to what one would obtain using the original unbinned data. The kernel density estimate $\hat{f}$ is the stationary distribution of a Markov process resembling the Ornstein–Uhlenbeck process. This $\hat{f}$ may be found by an iteration scheme which converges at a geometric rate in the entropy pseudo-metric, and hence in L^1 and transportation metrics. The proof uses a logarithmic Sobolev inequality comparing relative Shannon entropy and relative Fisher information with respect to $\hat{f}$.

Keywords: binned data; density estimation; kernel estimation; logarithmic Sobolev inequality; transportation

1. Introduction

This paper proposes a variant of kernel density estimation for binned data. Let $x_1, x_2, \ldots, x_N$ be a random sample from a distribution with density $f(x)$. Then the standard kernel density estimate of f is

$$\hat{f}(x) = \frac{1}{N}\sum_{j=1}^{N} K(x - x_j), \tag{1.1}$$

where K is the kernel function which is positive, symmetric about zero and integrates to one. Our choice of kernel is the Gaussian kernel

$$K(x - y) = \frac{1}{(2\pi)^{1/2}h}\exp\left\{-\frac{1}{2h^2}(x - y)^2\right\}, \qquad x, y \in \mathbb{R}. \tag{1.2}$$

We refer to the standard deviation h as the smoothing parameter or *bandwidth*. The appearance of the density estimate depends crucially on the choice of h. The choice of kernel function is not so crucial in terms of asymptotic statistical properties.

For reasons of efficient data storage, large data sets are often binned to some degree. The mildest form of binning may just amount to rounding to some fixed number of decimal places. At the other extreme, data may be counts over relatively large intervals or spatial

1350–7265

regions. In health and social research, for confidentiality reasons, there is often no access to individual level data, and data must be aggregated to a sufficient level.

In this paper we introduce a nonlinear variant of the kernel density estimator for binned data, which has the attractive property of reducing to standard kernel density estimation in the limit as the bin size diminishes. We assume that the data have been pre-binned, and that we have no control over the binning process. We consider binning as a nuisance, and aim for a density estimate which is as close as possible to what we would have obtained if we had the unbinned data. There are two approaches already available in this context. The first is to use standard kernel density estimation, treating all observations as if they were equal to the midpoint of their corresponding bins; this performs well when the binning is very fine, and becomes a much poorer approximation as the binning becomes coarser (Scott and Sheather 1985). Another approach is called 'weighted average of rounded points' (Härdle and Scott 1992). This is an approximation to kernel density estimation which copes better with relatively larger bin-widths than the simple estimator. Literature on these models is focused on their relative efficiencies for computing kernel density estimates from huge sets of data. In this scenario, the unbinned data *are* available yet, due to computational considerations, binning is chosen. The investigator has control of the choice of bin-width *as well as* the bandwidth.

The method that we propose performs well, at least qualitatively, even when the binning is quite coarse. If the bin sizes are large, however, then any information about small-scale fluctuations in the density will be lost irretrievably. The best we can hope for is an estimate which is as close as possible to the true density using the data available. Our approach is based upon similar ideas to Titterington (1983), who considers kernel density estimation for situations in which only a small proportion of the data is binned.

The simplest density estimate from binned data is the histogram or the frequency polygon, which is a trapezium with nodes above the midpoint of the bins. Minnotte (1996) suggests a 'bias-optimized frequency polygon' that preserves the mass proportions within bins. Most further improvements obtain a smooth density estimate from the histogram, employing spline-based methods. Early examples include Boneva *et al*. (1971), and that of Tobler (1979) who produces a smooth version of a *bivariate* histogram. More recently, Minnotte (1998) extended this work by introducing his version of 'histosplines'. These estimate the density as a sum of splines of even order, and reduce to the bias-optimized frequency polygon in the case of linear splines. Like the kernel estimates for binned data described above, most of these methods assume that the investigator has control over the choice of bins. In Minnotte's approach, the estimator always preserves bin mass proportions, which entails that the method will not work well when sample sizes or bin-widths are small, in which case observed bin proportions cannot be relied upon.

Minnotte's method can be considered as an alternative to kernel density estimation, which has the bin-width as its smoothing parameter rather than the bandwidth of the kernel function. In simulations it was found to perform well, compared to kernel approaches, for large sample sizes and large bin-widths (Minnotte 1998). An unpleasant property of this estimator is its tendency to produce estimates which are negative in places. The obvious remedy, of truncating the estimates at zero, abandons the property of preserving mass proportions within bins. Koo and Kooperberg (2000) extend ideas of logspline density

estimation in an approach to be used for binned data (see Stone *et al*. 1997; Kooperberg and Stone 1991). This approach, like ours, assumes that the data are pre-binned, and does not insist upon preserving bin mass proportions exactly. The 'smoothing parameter' for this approach is the number of knots assigned to the logspline. As estimation is on the log scale, density estimates are non-negative.

The theoretical results of Minnotte (1998) and Barron and Sheu (1991), which ensure convergence of the splines, are expressed in terms of the higher-order derivatives of the densities, without an intuitive probabilistic interpretation. Tobler (1979) seeks to smooth binned spatial point data by solving a discrete Dirichlet problem subject to mass-preserving constraints on each bin. There does not seem to be any simple characterization of the solution that minimizes the Dirichlet integral subject to both the mass-preserving constraints and the positivity condition. Further, Tobler's principal example concerned the population of an isolated country, where it is natural to impose the Neumann (zero normal derivative) boundary condition on the minimizer of the Dirichlet integral; but it is not clear that one can assume this in other cases.

Our binned kernel estimator has attractive properties, including the following.

(i) As the ratio of bin-width of the histogram to the bandwidth of the kernel diminishes, our method reduces to standard kernel density estimation.
(ii) In the limit as the smoothing parameter diminishes, we obtain a smooth estimate of the density which preserves mass proportions, with results similar in appearance to those of the cubic histospline estimator of Minnotte (1998).
(iii) The same method works for binned data in $\mathbb{R}^d$, given any finite collection of bounded and abutting bins.
(iv) No boundary conditions are required, and there are no smoothness assumptions on the underlying density.
(v) Estimates are always non-negative.

To introduce our method, we let A_j $(1 \leqslant j \leqslant m)$ be consecutive, abutting and bounded intervals in $\mathbb{R}$, with $\bigcup_{j=1}^{m} A_j = A$; and let $n_j > 0$ be such that $\sum_{j=1}^{m} n_j = N$. Now let $x_{j1}, x_{j2}, \ldots, x_{jn_j}$ denote the true (unknown) locations of the n_j data points in A_j. Then, referring to equation (1.1), the standard kernel density estimate is

$$\hat{f}(x) = \frac{1}{N} \sum_{j=1}^{m} \sum_{k=1}^{n_j} K(x - x_{jk}). \tag{1.3}$$

Since the true locations within the intervals A_j are unknown, we cannot construct this density estimate. Suppose for a moment that we *did* know the underlying density $f(x)$ on $\mathbb{R}$. Given that a particular observation is censored within the interval A_j, its location will be a realization from the density

$$f_j(x) = f(x) \left\{ \int_{A_j} f(y) \mathrm{d}y \right\}^{-1}$$

on A_j. For given j and k, since we do not know the true value of $K(x - x_{jk})$ in (1.3), we could substitute it by its expectation

$$\mathbb{E}\{K(x - X_{jk})\} = \int_{A_j} K(x - y) f_j(y) \mathrm{d}y.$$

Thus we obtain the revised estimator

$$\tilde{f}(x) = \sum_{j=1}^{m} \frac{n_j}{N} \frac{\int_{A_j} K(x - y) f(y) \mathrm{d}y}{\int_{A_j} f(y) \mathrm{d}y}, \qquad x \in \mathbb{R}. \tag{1.4}$$

But we do not know $f(y)$. For any positive and piecewise continuous function g, let us write

$$Tg(x) = \sum_{j=1}^{m} \frac{n_j}{N} \frac{\int_{A_j} K(x - y) g(y) \mathrm{d}y}{\int_{A_j} g(y) \mathrm{d}y}, \qquad x \in \mathbb{R}. \tag{1.5}$$

Equation (1.4) suggests that a smooth estimate $\hat{f}$ can be found as the limit of the iterative scheme

$$\hat{f}_{k+1} = T\hat{f}_k, \qquad k \geqslant 0; \tag{1.6}$$

thus $\hat{f}$ is a fixed point of T. The limit $\hat{f} = \lim_{k \to \infty} \hat{f}_k$ depends upon T; that is, upon h, the A_j and n_j/N. The natural choice of $\hat{f}_0$ is the density of the histogram of observed bin counts

$$\hat{f}_0(x) = \sum_{j=1}^{m} \frac{n_j}{N|A_j|} \mathbb{I}_{A_j}(x), \qquad x \in \mathbb{R},$$

where $\mathbb{I}_{A_j}$ denotes the indicator function of A_j. Empirically, however, convergence is found to be achieved faster by starting with a smooth estimate. In this paper we establish the following.

Theorem 1.1. *The operator T has fixed points $\hat{f} = T\hat{f}$ which are smooth and positive probability density functions of rapid decay at infinity. For suitable n_j/N, A_j and h, the iteration scheme* (1.6) *with initial data $\hat{f}_0$ converges geometrically, so that, for some C, $\eta > 0$, the iterates satisfy*

$$\int_{-\infty}^{\infty} |\hat{f}_k(x) - \hat{f}(x)| \mathrm{d}x \leqslant C \exp(-\eta k), \qquad k \geqslant 0.$$

More precise statements of the technical hypotheses will be given in Theorem 7.1 below. The probabilistic interpretation of the iteration scheme is as follows.

Let X be a random variable with density g, and Y a random variable with bin quotas $\mathbb{P}[Y \in A_j] = n_j/N$, and which has the same conditional distribution on A_j as X, so $Y|_{[Y \in A_j]} \sim X|_{[X \in A_j]}$ for each j. Let Y evolve to $Z = Y + B_\tau$ by addition of an independent

Brownian motion with $B_0 = 0$. Then Tg is the density of Z. The density $g = \hat{f}$ is a fixed point for T when X and Z have the same distribution.

Exploiting special properties of the fixed point, we shall deduce the following.

Corollary 1.2. *Let* $\hat{f}$ *and* $\hat{f}_k$ $(k \geqslant 1)$ *be probability density functions generated by the above scheme. Then there exist random variables* X *and* X_k $(k \geqslant 1)$ *on a common probability space such that* X *has density* $\hat{f}$, X_k *has density* $\hat{f}_k$ *and their joint distribution satisfies*

$$\mathbb{E}|X - X_k|^2 \leqslant C \exp(-\eta k), \qquad k \geqslant 1.$$

The remainder of this paper is arranged as follows. The practical performance of the estimator is investigated in Section 2. In Section 3 we establish the existence of fixed points for T. In Section 4 we review the basic properties of entropy and information. The main technical result needed to prove Theorem 1.1 is a logarithmic Sobolev inequality comparing relative entropy and information with respect to a fixed point of T; this is obtained in Section 5. The proof of Theorem 1.1 also requires detailed analysis of the binning operation, which is presented in Sections 6 and 7.

An important feature of certain logarithmic Sobolev inequalities, such as Gross's theorem (Gross 1975), is that they hold in spaces of arbitrarily high dimension, typically with constants which do not depend directly upon dimension. The results of Sections 3, 4 and 5 extend in the obvious way to bounded and abutting rectangular regions in higher dimensions.

2. Implementation

Figure 1 shows a kernel density estimate produced from binned data. In this case there are only ten observations, with a bin size of 0.5, and bandwidth $h = 0.2$. Also shown is the histogram and the kernel density estimate produced from the original data values. The binned kernel estimator produces a sensible estimate, given the extremely limited form of the data.

Changing the bandwidth has much the same effect as in standard kernel density estimation, as can be seen in Figure 2. A sample of size 200 was obtained from an equal mixture of two Gaussian densities (with standard deviation 1.2, and means 2.5 and 7.5), and the data binned. We see that as the bandwidth diminishes, the bin integrals of the density estimate follow more closely those of the histogram. In the limit $h \to 0$, the bin integrals of the histogram are preserved. This binned kernel estimator with '$h = 0$' can be compared with Minnotte's cubic histospline method, as shown in Figure 3. The two estimates are much the same, except that the cubic histospline has the tendency to be negative in some places.

As with standard kernel density estimation, it is vital to consider carefully the issue of choice of the bandwidth. Since the method simplifies to standard kernel density estimation as the bin-width tends to zero, we can use established methods of bandwidth selection for small bin sizes, and versions of them for larger bin-widths. We consider two approaches. The first is the 'plug-in' method of Sheather and Jones (1991), designed by attempting to

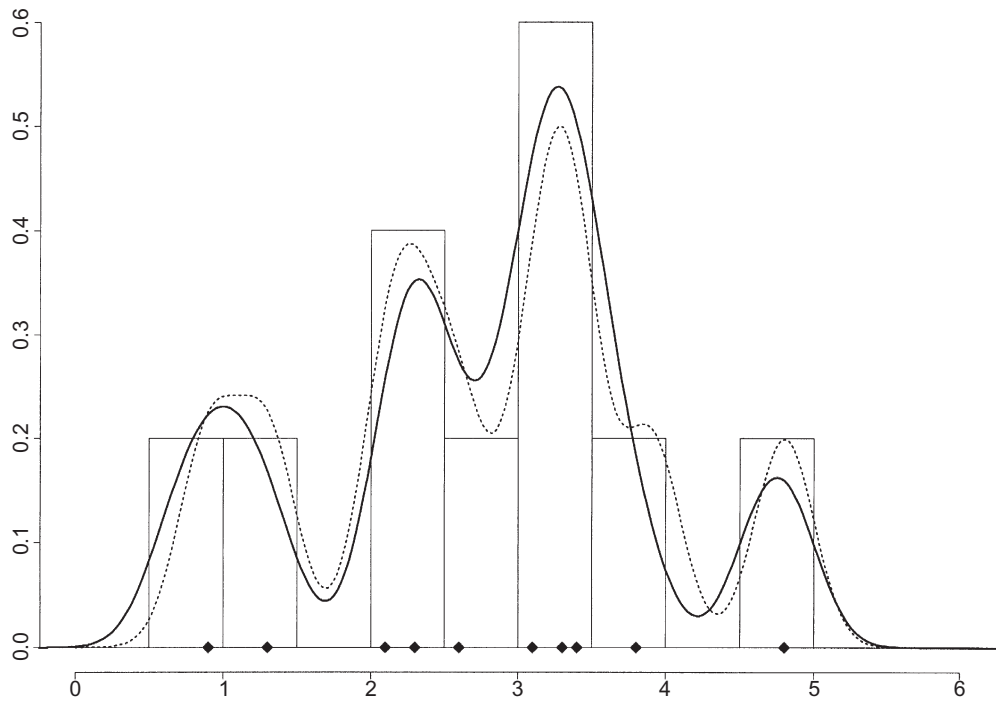

Figure 1. Illustration of kernel density estimates produced with Gaussian kernel and smoothing parameter $h = 0.2$. The dotted line shows the standard kernel density estimate produced from the original data. The solid line shows the estimate created from the histogram of the data with bin-width 0.5 (finer solid line).

minimize an asymptotic form of the mean integrated squared error. (We use the version implemented in the library of S-PLUS functions accompanying Azzalini and Bowman (1997).) The second approach is least-squares cross-validation which attempts to minimize the data-specific integrated squared error (see Bowman 1984). The first of these methods is more stable than the second; yet cross-validation approaches can be generalized more readily. See Wand and Jones (1995) for a fuller discussion of bandwidth choice.

For binned data, a simple approach for choosing the bandwidth is to *simulate* the locations of observations uniformly within bins and use any of the standard methods on these pseudo-data. To minimize the variability that the random simulation may induce, we can repeat this procedure and take the median of the values of smoothing parameter that are obtained. When the density is approximately uniform within bins, we can expect this approach to work. Where the amount of data is very large, this strategy may not be feasible in terms of the computational time involved, and an alternative approach is required.

The least-squares cross-validation method for standard kernel density estimation (as in Bowman 1984) seeks to minimize the integrated squared error (ISE), which amounts to minimizing the following expression when the kernel is Gaussian:

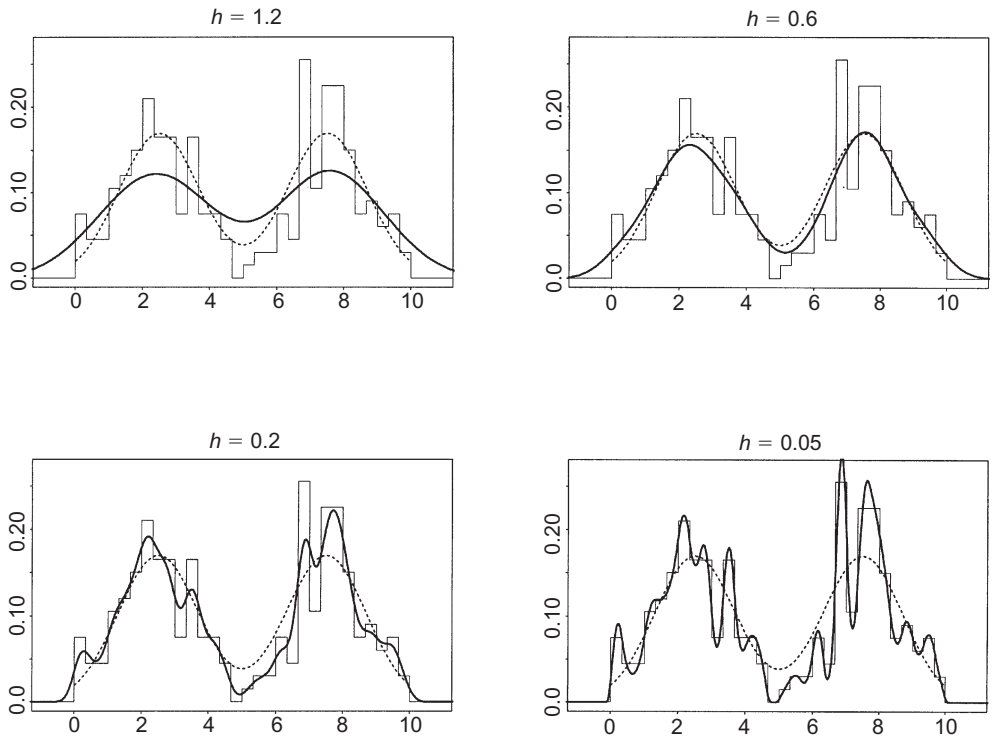

Figure 2. Smooth density estimates constructed from a histogram, using four different values of the smoothing parameter h. The histogram was constructed from a sample of size 200 from a mixture of two Gaussian densities shown on the plots as the dotted line.

$$\mathrm{CV}(h) = \int \hat{f}(x)^2 \, \mathrm{d}x - \frac{2}{N-1} \left\{ \sum_{j=1}^{N} \hat{f}(x_j) - (2\pi)^{-1/2} h^{-1} \right\}.$$

For estimation with binned data, we propose using a similar expression, replacing the summation term with the approximation

$$\sum_{j=1}^{N} \hat{f}(x_j) \approx \sum_{k=1}^{m} \frac{n_k}{|A_k|} \int_{A_k} \hat{f}(x) \mathrm{d}x.$$

For the purposes of this paper, we call this 'binned cross-validation'.

We conduct a simulation study to compare the three methods of choosing the smoothing parameter that we have outlined. In all cases we simulate from a density that is a 1:4 mixture of a normal density with mean 2 and standard deviation 0.17, and a lognormal density with mean 0 and standard deviation 0.5 on the log scale. This density is considered by Koo and Kooperberg (2000) to investigate their logspline approach.

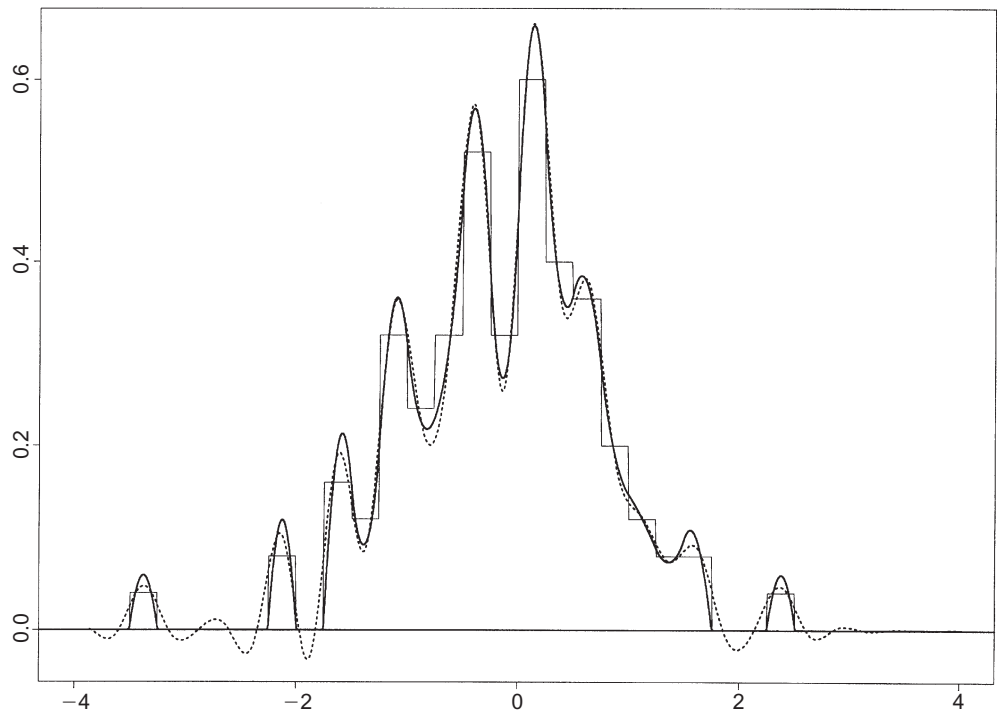

Figure 3. Illustration of the Minnotte (1998) histospline estimator (dotted line) compared with our binned kernel estimator in the limit as $h \to 0$ (solid line). The data were obtained as a sample of size 100 from a standard Gaussian density, and binned with bin-width 0.25.

We consider four different sample sizes, $n = 100$, 300, 1000 and 3000; and compare ISEs attained in 50 simulations for each combination of sample size, possible bin-widths and methods of choosing smoothing parameters. Median ISEs for the binned CV method are shown in Table 1, with the corresponding median smoothing parameters in Table 2; and in each case we show the *gold standard* values obtained by finding smoothing parameters that exactly minimize the ISE.

In Figure 4 are boxplots of the ISEs and smoothing parameter values obtained for the three methods of choosing the smoothing parameter for sample sizes of 100 and 1000. Qualitatively similar results were obtained for sample sizes of 300 and 3000. One can compare these results for medians with the corresponding results for mean ISEs from Table 1 of Koo and Kooperberg (2000) with $\sigma = 0.17$. Although we cannot directly compare means and medians since the distribution of ISEs tend to be positively skewed (see, for example, Figure 4(a)), it is clear that the ISEs obtained using the two approaches tend to be similar in magnitude.

The simulations show that, with optimal choice of smoothing parameter, the binned kernel estimator produces estimates which are hardly any worse, in terms of ISE, than

Table 1. Mean integrated squared errors (×1000) of kernel density estimates constructed from binned data. Smoothing parameters were chosen according to the binned cross-validation method. In parentheses are mean integrated squared errors corresponding to use of the smoothing parameter that minimizes the integrated squared error in each case

	Bin-width			
Sample size	Not binned	0.03	0.1	0.3
100	25.1 (18.1)	23.7 (21.5)	20.7 (17.7)	27.8 (19.6)
300	10.4 (9.1)	10.9 (8.3)	9.6 (8.2)	12.0 (10.4)
1000	3.9 (3.8)	4.2 (3.7)	4.1 (3.6)	5.3 (3.5)
3000	1.8 (1.5)	1.6 (1.6)	1.8 (1.9)	2.7 (2.3)

Table 2. Median values of smoothing parameters (×100) chosen using the binned cross-validation method. In parentheses are the medians of smoothing parameters that minimize the integrated squared error. These smoothing parameters correspond to the results in Table 1

	Bin-width			
Sample size	Not binned	0.03	0.1	0.3
100	14.5 (14.7)	15.9 (14.5)	18.2 (14.2)	21.0 (12.8)
300	11.3 (11.1)	12.1 (11.1)	12.7 (10.6)	14.9 (7.4)
1000	7.7 (8.7)	8.6 (8.5)	9.9 (8.3)	8.8 (3.9)
3000	6.3 (6.8)	6.7 (6.7)	7.7 (5.8)	6.2 (2.4)

kernel density estimation on unbinned data, even for relatively large bin-widths. The Sheather–Jones method for choosing the smoothing parameter works well, so long as the bin size and sample size are relatively small. It does not seem to work well, however, when the sample size is large relative to the bin-width. In this scenario, the fact that the pseudo-data actually follow a piecewise constant version of the true density is likely to be detectable. This will cause 'plug-in' methods, which rely on estimates of the second derivative, to choose values of the smoothing parameter that may be appropriate for estimating the discontinuous density, yet not for the underlying smooth density. Least-squares cross-validation is more variable in its choice of smoothing parameter, and tends to produce larger ISEs than the Sheather–Jones method (a feature well known), yet does not appear to suffer to the same extent the problems associated with having a large sample size relative to the bin-width. The binned cross-validation method performs similarly to cross-validation, yet it tends to choose larger smoothing parameters, with this being most obvious as the sample size increases relative to the bin-width. On the whole, binned cross-validation does not appear appreciably worse than the other methods. It is also computationally much

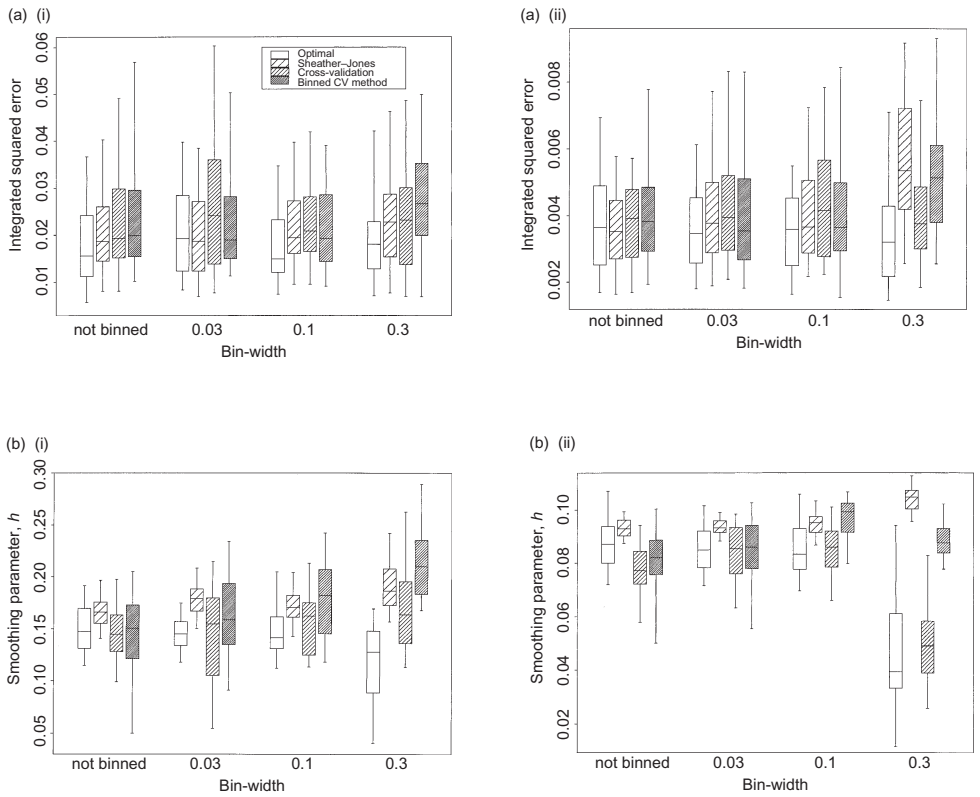

Figure 4. (a) Integrated squared errors of density estimates for three different methods of choosing the smoothing parameter for binned data, for four bin-widths and for sample sizes (i) $n = 100$ and (ii) $n = 1000$. The methods are compared with the optimal possible, where the smoothing parameter is chosen to minimize the integrated squared error for each sample. (b) Corresponding smoothing parameters (i) $n = 100$ and (ii) $n = 1000$.

faster; time depends on the number of bins, rather than the sample size. In conclusion, we recommend the Sheather–Jones method for choosing the bandwidth when the sample size is small (i.e. less than 300), and the binned cross-validation method for larger sample sizes. When the sample size is excessively large relative to the size of the bins such that the density bin proportions are accurately represented, we can effectively fix the smoothing parameter to be the limiting case of '$h = 0$'.

In practice, we find that convergence is slower for smaller smoothing parameter values. It is thus wise to start with a relatively large smoothing parameter and progressively to reduce this, as iterations proceed, to the desired value. This applies especially in the case where we require '$h = 0$'; here we continue until h becomes negligible in value, and convergence of the density estimate is achieved.

The binned kernel density estimation can also be applied to two-dimensional data, and

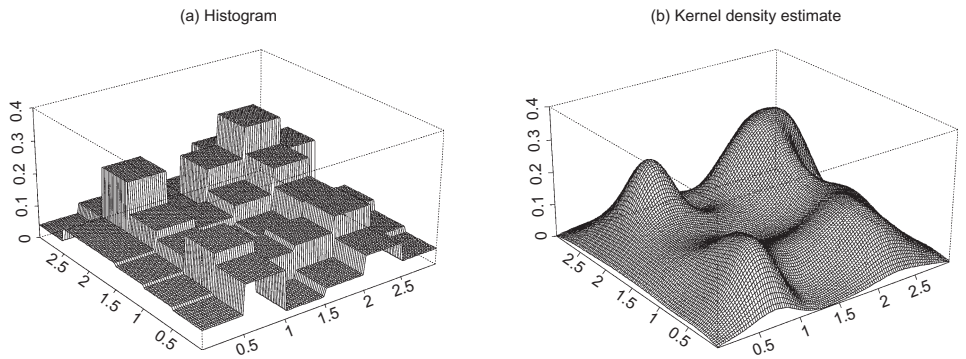

Figure 5. (a) Two-dimensional (scaled) histogram with (b) corresponding smooth density estimate constructed using a smoothing parameter of $h = 0.1$.

Figure 5 demonstrates the implementation of our method for spatial data. There is no need for the bins to be rectangular, which implies that the method has potential applicability in producing density estimates from counts over arbitrarily defined adjoining geographical areas. The binned cross-validation approach is applicable for choosing smoothing parameter(s) for spatial data as it is for one-dimensional data.

3. Existence of the fixed point

In this section we establish the existence of smooth densities that satisfy $Tf = f$ for kernels K such as the Gaussian kernel of (1.2). The kernel function is required to satisfy:

(i) smoothness, so that $K : \mathbb{R} \to \mathbb{R}$ is continuously differentiable;
(ii) positivity, $K(x) > 0$, for all x;
(iii) symmetry, $K(x) = K(-x)$;
(iv) decreasing on $[0, \infty)$; and
(v) normalization, so $\int_{-\infty}^{\infty} K(x)\mathrm{d}x = 1$.

The Gaussian kernel satisfies all these conditions, whereas the Epanechnikov kernel fails (ii) as it is of compact support.

By (i), (ii) and compactness of A, we can introduce a strictly positive constant

$$\eta = \inf\left\{\frac{n_j}{N} K(x - y) \,|\, x,\ y \in A_j;\ 1 \leqslant j \leqslant m\right\}.$$

Let us also introduce the non-empty set of continuous functions

$$\mathbb{S} = \left\{ g \in C(A; \mathbb{R}) \,|\, g(x) \geqslant \eta,\ (1 - \delta)\left(1 - \frac{n_1 + n_m}{2N}\right) \leqslant \int_A g(y)\mathrm{d}y \leqslant 1 \right\},$$

where $1 > \delta > 0$ is selected in (3.1) and depends upon the A_j and the kernel K only. This $\mathbb{S}$ is a closed and convex set of continuous (deficient) probability density functions.

Theorem 3.1. *The operator T of* (1.5) *maps $\mathbb{S}$ to itself, and has some fixed point in $\mathbb{S}$. Such a fixed point is a continuously differentiable function which extends to define a probability density function.*

Proof. Since $g(x) \geqslant \eta$, the denominators are positive and the operator is well defined; clearly $Tg(x)$ is continuous. By the definition of η we have $Tg(x) \geqslant \eta$ for all $x \in A$. We can take $1 > \delta > 0$ with

$$\int_{A_{j-1} \cup A_j \cup A_{j+1}} K(x - y)\mathrm{d}y \geqslant 1 - \delta, \tag{3.1}$$

for all $x \in A_j$ and $2 \leqslant j \leqslant m - 1$; we can make δ as small as we please by shrinking the bandwidth h of K. Hence,

$$\int_A Tg(x)\mathrm{d}x \geqslant \sum_{j=2}^{m-1} \frac{n_j}{N} \frac{\int_{A_j} \int_{A_{j-1} \cup A_j \cup A_{j+1}} K(x - y)\mathrm{d}x\, g(y)\mathrm{d}y}{\int_{A_j} g(y)\mathrm{d}y}.$$

On estimating the inner integrals, we see that this is

$$\geqslant \sum_{j=2}^{m-1} \frac{n_j}{N}(1 - \delta) = \left(1 - \frac{n_1 + n_m}{N}\right)(1 - \delta).$$

By making a slight adjustment to the definition of δ, we can incorporate the terms $j = 1$ and $j = m$, and replace N by $2N$. Hence T maps $\mathbb{S}$ to itself.

The operator T is completely continuous (compact), in the sense that $T(\mathbb{S})$ is a relatively compact set for the standard supremum-norm topology on $\mathbb{S}$, viewed as a subset of $C(A; \mathbb{R})$. By the Arzelà–Ascoli theorem, compact subsets of $C(A; \mathbb{R})$ are characterized by the properties of uniform boundedness and equicontinuity.

To verify the first, we note that $0 \leqslant K(x) \leqslant K(0)$ and hence by convexity $0 \leqslant Tg(x) \leqslant K(0)$ for all $x \in A$ and $g \in \mathbb{S}$. A standard application of the mean value theorem shows that uniform equicontinuity is implied by a uniform bound on the derivative. Here by (i) we have M such that $|K'(x - y)| \leqslant M$ for all $x, y \in A$; whence

$$|(Tg)'(x)| = \left| \sum_{j=1}^{m} \frac{n_j}{N} \frac{\int_{A_j} K'(x - y)g(y)\mathrm{d}y}{\int_{A_j} g(y)\mathrm{d}y} \right| \leqslant M$$

by the triangle inequality, for all $x \in A$ and $f \in \mathbb{S}$.

By the theorem of Schauder, T has a fixed point; that is, $T\hat{f}(x) = \hat{f}(x)$ $(x \in A)$ for some $\hat{f} \in \mathbb{S}$ (see Heuser 1982). It follows from the previous step that $T\hat{f}$, and hence $\hat{f}$, is

continuously differentiable. Indeed, $\hat{f}$ inherits the same degree of smoothness as the kernel function K. Our solution $\hat{f}$ is defined only for $x \in A$; however, we can use the formula $\hat{f}(x) = T\hat{f}(x)$, in which the right-hand side depends only upon $\hat{f}$ restricted to A, to extend $\hat{f}$ to a genuine probability density function on $\mathbb{R}$. Since $\hat{f} \in \mathbb{S}$, most of the probability is concentrated on A. By choosing K to be sharply peaked, so that δ diminishes to zero, we can ensure that less than $(n_1 + n_m)(1 + \varepsilon)/(2N)$ of the probability lies outside the histogram region A. Further, $\hat{f}$, so extended, is continuously differentiable and decreases to zero at infinity. □

The preceding proof does not ensure uniqueness of the fixed point. In applications one has a computationally effective means of estimating bin quotas, so the following partial result is adequate for establishing uniqueness (see also Proposition 6.4 below).

Proposition 3.2. *Let $\hat{f}$ be a fixed point of T, and let $\alpha_j = \int_{A_j} \hat{f}$. If q is any fixed point of T with $\int_{A_j} q = \alpha_j$, then $\hat{f} = q$ on A.*

Proof. Let $w(x) = \sum_{j=1}^{m} n_j/(N\alpha_j)\mathbb{I}_{A_j}(x)$ and let

$$\tilde{K}(x, y) = w(x)^{1/2} K(x - y) w(y)^{1/2}, \qquad x, y \in A,$$

which defines a square-integrable, symmetric and uniformly positive function on $A \times A$. Hence the linear operator $\tilde{T} : L^2(A) \to L^2(A)$ given by $\tilde{T}g = \int_A \tilde{K}(x, y)g(y)\mathrm{d}y$ is compact and self-adjoint, and its norm M coincides with its numerical radius. We deduce that there exists a unique normalized positive eigenfunction f_M corresponding to eigenvalue M. It follows from the fixed-point equation that $w(x)^{1/2}\hat{f}(x)$ and $w(x)^{1/2}q(x)$ are eigenfunctions corresponding to eigenvalue 1. If $M > 1$, then $w(x)^{1/2}\hat{f}(x)$ and f_M would be orthogonal; but this contradicts their positivity. We deduce that $M = 1$, and so $w(x)^{1/2}\hat{f}(x)$ and $w(x)^{1/2}q(x)$ are multiples of $f_M(x)$; and since $w(x) > 0$ on A, we see $\hat{f}$ and q are equal. □

4. Entropy and information

Given probability density functions p and q on the real line, we define the entropy of q relative to p by

$$S(q|p) = \int_{-\infty}^{\infty} q(x)\log q(x)/p(x)\mathrm{d}x.$$

By Jensen's inequality, $S(q|p) \geqslant 0$; so we can always define the relative entropy if we admit infinity as a possible value. The relative entropy $S(q|p)$ is jointly convex in (q, p) since $U(s, t) = s\log s/t$ is a biconvex function of $s, t > 0$. In information theory, $S(q|p)$ is called informational divergence.

We can interpret $S(q|p)$ as a measure of the distance from the probability density q to p, where p is regarded as an equilibrium configuration. Csiszár's inequality (see Barron 1986) shows that

$$\int_{-\infty}^{\infty} |q(x) - p(x)| \mathrm{d}x \leqslant \{2S(q|p)\}^{1/2}. \tag{4.1}$$

Consequently, the sets $G_q(\eta) = \{p|S(p|q) < \eta\}$ $(\eta > 0)$ give a sub-basis for a topology on the densities which is finer than the total variation topology associated with the L^1 metric. In Proposition 4.2 we shall show that, when $q = \hat{f}$, such sub-basic open neighbourhoods are contained in sets which are bounded for the Wasserstein metric.

Let ∇ denote the distributional gradient operator in $L^2(\mathbb{R})$, and $W^{1,2}(\mathbb{R})$ the space of functions $g \in L^2(\mathbb{R})$ for which ∇g also belongs to $L^2(\mathbb{R})$. When p and q are probability density functions with $p^{1/2}$ and $q^{1/2}$ in $W^{1,2}(\mathbb{R})$, we can define the Fisher information of p by

$$I(p) = \int_{-\infty}^{\infty} |\nabla \log p(x)|^2 p(x) \mathrm{d}x = 4 \int_{-\infty}^{\infty} |\nabla \{p(x)^{1/2}\}|^2 \, \mathrm{d}x,$$

and the relative Fisher information of q with respect to p by

$$I(q|p) = \int_{-\infty}^{\infty} |\nabla \log\{q(x)/p(x)\}|^2 q(x) \mathrm{d}x.$$

This defines a convex function of q. A density p has finite information if and only if $p^{1/2} \in W^{1,2}(\mathbb{R})$.

Henceforth we take K to be the Gaussian kernel of (1.2), and write $K = K_\tau$ with $\tau = h^2$. In our iteration scheme we work with densities of the form $p = K_\tau * g$, for some density g; the information of such satisfies by convexity the uniform bound

$$I(p) \leqslant I(K_\tau) = 1/\tau = 1/h^2.$$

Barron (1986) noted a relationship between relative entropy and information which arises under evolution by the heat equation

$$\dot{u}_t(x) = \frac{\partial}{\partial t} u_t(x) = \frac{1}{2} \nabla^2 u_t(x), \qquad x \in \mathbb{R},\ t > 0. \tag{4.2}$$

This will be used in our smoothing scheme in the following way. Let X and Y be random variables with probability density functions q and p respectively, and B_t $(t \geqslant 0)$ be an independent Brownian motion with $B_0 = 0$. Then $X + B_t$ and $Y + B_t$ have probability density functions $u_t = K_t * q$ and $v_t = K_t * p$ respectively, where $u_t(x)$ and $v_t(x)$ satisfy (4.2).

Lemma 4.1. *Relative information measures the rate of change of relative entropy under the heat flow, so*

$$\frac{\mathrm{d}}{\mathrm{d}t} S(u_t|v_t) = -\frac{1}{2} I(u_t|v_t), \qquad t > 0. \tag{4.3}$$

Proof. The functions $u_t(x)$ and $v_t(x)$ are smooth for $(x, t) \in \mathbb{R} \times (0, \infty)$, and we can differentiate through the integral sign as in the proofs of Davies (1989) to obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{\infty} u_t(x)\log\{u_t(x)/v_t(x)\}\mathrm{d}x = \int_{-\infty}^{\infty} \dot{u}_t \log\{u_t/v_t\}\mathrm{d}x - \int_{-\infty}^{\infty} u_t\dot{v}_t/v_t\,\mathrm{d}x + \int_{-\infty}^{\infty} \dot{u}_t\,\mathrm{d}x.$$

The latest integral vanishes since probability is conserved under the evolution, and from the heat equation (4.2) we have

$$\frac{\mathrm{d}}{\mathrm{d}t}S(u_t|v_t) = \frac{1}{2}\int_{-\infty}^{\infty}\{(\nabla^2 u_t)\log(u_t/v_t) - (\nabla^2 v_t)u_t/v_t\}\mathrm{d}x.$$

Integration by parts leads to the required result

$$\frac{\mathrm{d}}{\mathrm{d}t}S(u_t|v_t) = -\frac{1}{2}\int_{-\infty}^{\infty}\left(\frac{\nabla u_t}{u_t} - \frac{\nabla v_t}{v_t}\right)^2 u_t\,\mathrm{d}x.$$

□

To exploit this to full advantage, we need another relation between entropy and information, namely a logarithmic Sobolev inequality. Let $K_1(x) = (\exp(-x^2/2))/(2\pi)^{1/2}$ be the standard Gaussian density. Gross (1975; 1993) showed $S(q|K_1) \leqslant I(q|K_1)/2$ for all densities q with $q^{1/2} \in W^{1,2}(\mathbb{R})$. His proof began with a logarithmic Sobolev inequality for a product of two-point spaces, and then realized the Gaussian density K_1 as a weak limit via the central limit theorem. In the next section we deduce from Gross's theorem a logarithmic Sobolev inequality for relative information and entropy with respect to a density $\hat{f}$ that satisfies $T\hat{f} = \hat{f}$.

Relative entropy is also related to transportation cost between distributions. See Dudley (1989) for an account of the theory.

Proposition 4.2. *Let $\hat{f}$ be a fixed point of T, and g be a probability density function which has finite entropy relative to $\hat{f}$. Then there exists a probability measure $\mathbb{P}$ on $\mathbb{R}^2$, with marginal densities $\hat{f}$ and g, such that the quadratic transportation cost is bounded in terms of the relative entropy*

$$\iint_{\mathbb{R}^2} |x-y|^2\mathbb{P}(\mathrm{d}x\,\mathrm{d}y) \leqslant 2C_T S(g|\hat{f}); \tag{4.4}$$

here C_T depends only upon the initial data and h.

Proof. Talagrand (1996) showed that a similar result holds with the Gaussian density K_1 replacing $\hat{f}$; and we shall see that (4.4) follows from his result by duality.

The density $\hat{f}$ is bounded and uniformly positive on $A = [a, b]$, and decays rapidly outside of $[a, b]$. Consequently, the increasing real function ψ defined by $\int_{-\infty}^{\psi(x)} \hat{f}(t)\mathrm{d}t = \int_{-\infty}^{x} K_1(t)\mathrm{d}t$ is bi-Lipschitz, so that $|x-y|/c_L \leqslant |\psi(x) - \psi(y)| \leqslant c_L|x-y|$ $(x, y \in \mathbb{R})$ for some constant c_L, with $0 < c_L < \infty$, depending upon $\hat{f}$. Furthermore, for any bounded and continuous real function v, we have

$$\int_{-\infty}^{\infty} v(x)\hat{f}(x)\mathrm{d}x = \int_{-\infty}^{\infty} v(\psi(x))K_1(x)\mathrm{d}x;$$

thus ψ induces $\hat{f}(x)\mathrm{d}x$ from $K_1(x)\mathrm{d}x$.

By the Kantorovich–Rubinstein duality theorem, the transportation cost $\mathrm{Tc}(g, \hat{f})$ of (4.4) for the cost function $|x-y|^2$ and the optimal transportation strategy $\mathbb{P}$ is equal to

$$\sup_{u,v}\left\{\int_{-\infty}^{\infty} u(x)g(x)\mathrm{d}x - \int_{-\infty}^{\infty} v(y)\hat{f}(y)\mathrm{d}y \;\middle|\; u(x)-v(y) \leqslant |x-y|^2;\ x,\ y \in \mathbb{R}\right\}.$$

Let G be the density of the measure induced from $g(x)\mathrm{d}x$ by the inverse function ψ^{-1}, just as K_1 is the density of the measure induced from $\hat{f}(x)\mathrm{d}x$. Since ψ is Lipschitz, we have

$$\mathrm{Tc}(g, \hat{f}) \leqslant c_L^2 \mathrm{Tc}(G, K_1).$$

By Talagrand's theorem, this transportation cost is bounded by a multiple of the relative entropy, and is

$$\leqslant C_1 c_L^2 S(G|K_1).$$

The relative entropy is equal to

$$S(G|K_1) = \sup_V\left\{\int_{-\infty}^{\infty} V(x)G(x)\mathrm{d}x \;\middle|\; \int_{-\infty}^{\infty} \mathrm{e}^{V(x)}K_1(x)\mathrm{d}x \leqslant 1\right\},$$

and it follows that $S(G|K_1) \leqslant S(g|\hat{f})$ since ψ^{-1} induces $G(x)\mathrm{d}x$ and $K_1(x)\mathrm{d}x$ from $g(x)\mathrm{d}x$ and $\hat{f}(x)\mathrm{d}x$, respectively. □

Remark on Corollary 1.2. We proceed to translate Proposition 4.2 into probabilistic language. Let X and Y be random variables on a common probability space with laws $\hat{f}$ and g, respectively. The left-hand side of (4.4) is $\mathbb{E}|X-Y|^2$, which we see is bounded in terms of the relative entropy when the joint distribution $\mathbb{P}$ is suitably chosen. Thus Corollary 1.2 becomes a consequence of convergence in relative entropy as established below.

5. Logarithmic Sobolev inequality

Theorem 5.1. *Let $\hat{f}$ be a fixed point of T. Then there exists a constant $\kappa > 0$, depending only upon $h = \tau^{1/2}$ and the initial data, such that*

$$S(q|\hat{f}) \leqslant \frac{1}{2\kappa} I(q|\hat{f}) \tag{5.1}$$

for all densities q with $q^{1/2} \in W^{1,2}(\mathbb{R})$.

Proof. The fixed point $\hat{f}$ has the form $\hat{f} = K_h * g$, where $g(x) = \sum_{j=1}^m p_j \mathbb{I}_{A_j}(x)$, with $p_j > 0$ $(j = 1, 2, \ldots, m)$, is a probability density function. The measure $g(x)\mathrm{d}x$ is induced from $K_h(x)\mathrm{d}x$ by $\phi : \mathbb{R} \to [a, b]$ that is defined by

$$\int_a^{\phi(x)} g(t)\mathrm{d}t = \int_{-\infty}^{x} K_h(t)\mathrm{d}t, \qquad x \in \mathbb{R};$$

and, moreover, ϕ is Lipschitz continuous with

$$0 < \phi'(x) = \frac{K_h(x)}{g(\phi(x))} \leqslant c_L := (2\pi h)^{-1/2}\Big(\min_j p_j\Big)^{-1}.$$

By Gross's theorem (Gross 1975; 1993), the bivariate Gaussian probability measure $K_h(x)K_h(y)\mathrm{d}x\,\mathrm{d}y$ satisfies the logarithmic Sobolev inequality

$$\iint_{\mathbb{R}^2} F(x,\,y)^2 \log\left(F(x,\,y)^2 \Big/ \iint_{\mathbb{R}^2} F^2 K_h \otimes K_h\right) K_h(x)K_h(y)\mathrm{d}x\,\mathrm{d}y$$

$$\leqslant \frac{1}{2\gamma}\iint_{\mathbb{R}^2} \|\nabla F(x,\,y)\|^2 K_h(x)K_h(y)\mathrm{d}x\,\mathrm{d}y, \tag{5.2}$$

for all $F \in L^2(K_h \otimes K_h)$ with L^2 distributional gradient.

The map $(x,\,y) \mapsto x + \phi(y)$ is Lipschitz continuous $\mathbb{R}^2 \to \mathbb{R}$, and induces $\hat{f}(t)\mathrm{d}t = K_h * g(t)\mathrm{d}t$ from $K_h(x)K_h(y)\mathrm{d}x\,\mathrm{d}y$; so when we substitute $u(x + \phi(y)) = F(x,\,y)$ into the left-hand side of (5.2) we obtain the relative entropy expression

$$\int_{\mathbb{R}} u(t)^2 \log\left(u(t)^2 \Big/ \int_{\mathbb{R}} u^2 \hat{f}\right)\hat{f}(t)\mathrm{d}t. \tag{5.3}$$

With this choice we have $\|\nabla F(x,\,y)\|^2 = |u'(x + \phi(y))|^2(1 + \phi'(y)^2)$, so the right-hand side of (5.2) is

$$\leqslant (2\gamma)^{-1}(1 + c_L^2)\int_{\mathbb{R}} |u'(t)|^2 \hat{f}(t)\mathrm{d}t. \tag{5.4}$$

When $q = u^2\hat{f}$ we obtain (5.1) from (5.2), (5.3) and (5.4) with $\kappa = \gamma/(1 + c_L^2)$. □

Corollary 5.2. *Let g be a probability density function on A and $\hat{f}$ be a fixed point for T. Then smoothing decreases the relative entropy, so*

$$S(K_\tau * g|K_\tau * \hat{f}) \leqslant (1 + \tau\kappa/2)^{-1} S(g|\hat{f}). \tag{5.5}$$

Proof. We consider the evolution of $S(K_t * g|K_t * \hat{f})$ from $t = \tau/2$ to $t = \tau$. By taking $\tau/2 > 0$ as the starting time we can work with smooth functions; while $t = \tau$ gives our usual smoothing kernel. By Lemma 4.1 we have $(\mathrm{d}/\mathrm{d}t)S(K_t * g|K_t * \hat{f}) = -\frac{1}{2}I(K_t * g|K_t * \hat{f})$; integrating this identity, we achieve

$$S(K_\tau * g|K_\tau * \hat{f}) = S(K_{\tau/2} * g|K_{\tau/2} * \hat{f}) - \frac{1}{2}\int_{\tau/2}^{\tau} I(K_t * g|K_t * \hat{f})\mathrm{d}t.$$

From the differential equation (4.3), or joint convexity of the relative entropy, we have $S(K_{\tau/2} * g|K_{\tau/2} * \hat{f}) \leqslant S(g|\hat{f})$. To bound the integral we use the logarithmic Sobolev inequality of Theorem 5.1, which also holds for information and entropy defined with respect to the density $K_t * \hat{f}$ for $\tau/2 \leqslant t \leqslant \tau$. On substituting this into the previous identity, we have

$$S(K_\tau * g|K_\tau * \hat{f}) \leqslant S(g|\hat{f}) - \kappa\int_{\tau/2}^{\tau} S(K_t * g|K_t * \hat{f})\mathrm{d}t$$

$$\leqslant S(g|\hat{f}) - \frac{\kappa\tau}{2} S(K_\tau * g|K_\tau * \hat{f}),$$

which implies (5.5). □

The density $\hat{f}$ is associated with a Dirichlet form and a self-adjoint operator L with core $C_c^\infty(\mathbb{R})$ in $L^2(\hat{f})$ defined by

$$\langle Lg|g\rangle_{L^2(\hat{f})} = \int_{-\infty}^{\infty} |\nabla g(x)|^2 \hat{f}(x)\mathrm{d}x, \qquad g \in C_c^\infty(\mathbb{R}).$$

Let λ_1 be the best possible constant in the Poincaré inequality $\int |g - \int g\hat{f}|^2\hat{f} \leqslant \lambda_1^{-1}\int |\nabla g|^2\hat{f}$ for this Dirichlet form. Let us also introduce the variance of the probability distribution with density $\hat{f}$, by $\sigma^2 = \int (x-\bar{x})^2\hat{f}(x)\mathrm{d}x$, where $\bar{x} = \int x\hat{f}(x)\mathrm{d}x$ is the mean.

Proposition 5.3. *The preceding constants satisfy*

$$\kappa \leqslant \lambda_1 \leqslant \sigma^{-2} \leqslant I(\hat{f}).$$

Proof. The left-hand inequality is due to Rothaus (1980), and one can often take equality here; for a general discussion of this point, see Deuschel and Stroock (1990). To achieve the middle inequality, one sets $g(x) = x - \bar{x}$ in Poincaré's inequality. The right-hand inequality is Heisenberg's uncertainty principle applied to $(\hat{f})^{1/2}$. □

Remarks. Otto and Villani (2000) have shown that, under mild technical conditions, the logarithmic Sobolev inequality implies the quadratic transportation inequality for measures on $\mathbb{R}^d$. In particular, the quadratic transportation inequality holds for measures of the form $\mu(\mathrm{d}x) = \mathrm{e}^{-V(x)}\,\mathrm{d}x$, where $V : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable and uniformly convex.

A fixed point $\hat{f}$ of T may be viewed as the stationary distribution for a diffusion process analogous to the Ornstein–Uhlenbeck process, as in Carlen and Soffer (1991). See Kallenberg (1997) for the following result.

Proposition 5.4. *Let Z_t be a Markov diffusion process that satisfies the Langevin equation*

$$\mathrm{d}Z_t = \frac{1}{2}\rho(Z_t)\mathrm{d}t + \mathrm{d}B_t, \qquad t > 0, \tag{5.6}$$

where $\rho(x) = \nabla \log \hat{f}(x)$. Then the transition probabilities for Z_t satisfy

$$\frac{\partial}{\partial t}P_t(x, z) = \frac{1}{2}\nabla^2 P_t(x, z) - \frac{1}{2}\nabla\{\rho(x)P_t(x, z)\}, \qquad x, y \in \mathbb{R};\ t > 0.$$

The formal adjoint of L in $L^2(\mathbb{R})$ is given on smooth functions of compact support by $-L^* g = \nabla^2 g - \nabla(\rho g)$. It is evident one can define a semigroup $\exp(-tL^*)$ by

$$w(x,\, t) = \exp(-tL^*)q(x) = \int_{-\infty}^{\infty} P_t(x,\, z)q(z)\mathrm{d}z$$

such that $\exp(-tL^*)$ preserves the cone of probability density functions on the line. When q is the density of random variable X, one can regard $w_t = \exp(-tL^*)q$ as the density of the random variable obtained by evolving $Z_0 = X$ to Z_t under the Langevin equation (5.6). The operator is so chosen that $L^* \hat{f} = 0$; thus $\hat{f}$ gives the stationary distribution for the evolution under the adjoint semigroup.

Using the same mode of proof as Lemma 4.1, one can show that $(\mathrm{d}/\mathrm{d}t)S(w_t|\hat{f}) = -\frac{1}{2}I(w_t|\hat{f}),\ t > 0$.

6. Convergence of the iteration scheme

Henceforth, let us write $\alpha_j = \int_{A_j} g$. In this section we shall investigate the properties of the α_j, in order to control the binning operation and to estimate the quotas assigned to the respective bins by the fixed point. By a sequence of lemmas, we obtain the following.

Proposition 6.1. *The quotas of a fixed point $\hat{f}$ satisfy the approximate identity $\int_{A_j} \hat{f} \approx \tilde{a}_j$ for small bandwidth, where*

$$\tilde{a}_j = \frac{n_{j-1}h}{N|A_{j-1}|(2\pi)^{1/2}} + \frac{n_j}{N|A_j|}\left\{|A_j| - h\left(\frac{2}{\pi}\right)^{1/2}\right\} + \frac{n_{j+1}h}{N|A_{j+1}(2\pi)^{1/2}}, \qquad 1 \leqslant j \leqslant m. \quad (6.1)$$

The operator T is a composition of the smoothing operator $g \mapsto K * g$ and the (nonlinear) binning operator

$$Bg(x) = \sum_{j=1}^{m} \frac{n_j}{N\alpha_j} g(x)\mathbb{I}_{A_j}(x), \qquad x \in \mathbb{R},$$

both of which preserve the cone of probability density functions. It is also convenient to introduce the conditional expectation with respect to the σ-algebra $\mathbb{A}$ generated by the A_j $(1 \leqslant j \leqslant m)$:

$$\mathbb{E}_{\mathbb{A}} g(x) = \sum_{j=1}^{m} \frac{\alpha_j}{|A_j|}\mathbb{I}_{A_j}(x), \qquad x \in \mathbb{R},$$

and the discrepancy operator

$$\Delta g(x) = Bg(x) - \mathbb{E}_{\mathbb{A}} Bg(x).$$

One may verify that $\mathbb{E}_{\mathbb{A}} Bg = \hat{f}_0$, the histogram density, and so we have

$$Tg = K * Bg = K * \hat{f}_0 + K * \Delta g. \quad (6.2)$$

We recall the error function $\operatorname{erf}(t) = 1 - \Phi(t)$, where Φ is the cumulative distribution function of the standard normal random variable, and we let $\varepsilon_0 = \operatorname{erf}(|A_j|/h)$.

Lemma 6.2. *The* $\tilde{\alpha}_j$ *given by* (6.1) *satisfy*

$$\left| \int_{A_j} K * \hat{f}_0 - \tilde{\alpha}_j \right| \leqslant \varepsilon_0. \tag{6.3}$$

Proof. The error function is rapidly decreasing and satisfies

$$\int_{-\infty}^{0} \int_{0}^{\infty} K(x-y)\mathrm{d}x\,\mathrm{d}y = h \int_{0}^{\infty} \operatorname{erf}(u)\mathrm{d}u;$$

integration by parts reduce this to

$$= \frac{h}{(2\pi)^{1/2}} \int_{0}^{\infty} u \exp\left(\frac{-u^2}{2}\right) \mathrm{d}u = \frac{h}{(2\pi)^{1/2}}. \tag{6.4}$$

We can write the quantity to be approximated by

$$\int_{A_j} K * \hat{f}_0(x)\mathrm{d}x = \sum_{k=1}^{m} \frac{n_k}{N|A_k|} \int_{A_j} \int_{A_k} K(x-y)\mathrm{d}y\,\mathrm{d}x, \tag{6.5}$$

and we observe that, as the Gaussian kernel decays rapidly, the most significant summands come from those k with $|k-j| \leqslant 1$. Indeed, by convexity we have

$$\sum_{k:|k-j|>1} \frac{n_k}{N|A_k|} \int_{A_j} \int_{A_k} K(x-y)\mathrm{d}y\,\mathrm{d}x \leqslant \sup_{x \in A_j} \int_{A_{j\mp 2}} K(x-y)\mathrm{d}y$$

$$\leqslant \operatorname{erf}(|A_{j\pm 1}|/h) = \varepsilon_0.$$

For $k = j+1$ we can use Fubini's theorem to write

$$\frac{n_{j+1}}{N|A_{j+1}|} \int_{A_j} K * \mathbb{I}_{A_{j+1}}(x)\mathrm{d}x = \frac{n_{j+1}}{N|A_{j+1}|} \int_{A_j} \int_{A_{j+1}} K(x-y)\mathrm{d}x\,\mathrm{d}y. \tag{6.6}$$

The main contribution to this double integral comes when $x \in A_j$ and $y \in A_{j+1}$ are close together, so we can extend the ranges of integration symmetrically about the upper endpoint of A_j and use (6.4) to write (6.6) as

$$\frac{n_{j+1}}{N|A_{j+1}|} \int_{-\infty}^{0} \int_{0}^{\infty} K(x-y)\mathrm{d}x\,\mathrm{d}y + O\left(\frac{n_{j+1}\varepsilon}{N}\right) = \frac{n_{j+1}h}{N|A_{j+1}|(2\pi)^{1/2}} + O\left(\frac{n_{j+1}\varepsilon_0}{N}\right). \tag{6.7}$$

A similar identity holds for $j = k-1$.

The principal contribution to the sum (6.5) arises when $k = j$. We use conservation of probability and the definition of the error function to write

$$\frac{n_j}{N} = \int_{\mathbb{R}} K * \frac{n_j \mathbb{I}_{A_j}}{N|A_j|} = \int_{A_{j-1}} + \int_{A_j} + \int_{A_{j-1}} K * \frac{n_j \mathbb{I}_{A_j}}{N|A_j|} + O\left(\frac{n_j \varepsilon}{N}\right). \tag{6.8}$$

Rearranging this, and using (6.7) to deal with the integrals over the neighbouring bins, we obtain

$$\int_{A_j} K * \frac{n_j \mathbb{I}_{A_j}}{N|A_j|} = \frac{n_j}{N|A_j|}\left\{1 - \frac{h}{|A_j|}\left(\frac{2}{\pi}\right)^{1/2}\right\} + O\left(\frac{n_{j\pm1}\varepsilon_0}{N}\right). \qquad \square$$

We now obtain a bound on the norm of the binning operator.

Lemma 6.3. *Let g and g_0 be densities on A; and let $\beta = \max_{1\leqslant j\leqslant m} n_j/(N\alpha_j)$, where as usual $\alpha_j = \int_{A_j} g$. Then*

$$S(Bg|Bg_0) \leqslant \beta S(g|g_0). \tag{6.9}$$

Proof. Writing $\gamma_j = \int_{A_j} g_0$, we have

$$S(Bg|Bg_0) = \sum_{j=1}^{m} \frac{n_j}{N} \int_{A_j} \frac{g(x)}{\alpha_j} \log \frac{g(x)}{\alpha_j} \frac{\gamma_j}{g_0(x)}\, \mathrm{d}x$$

where each summand is non-negative by Jensen's inequality. Hence the sum is

$$\begin{aligned} &\leqslant \beta \sum_{j=1}^{m} \int_{A_j} g(x) \log \frac{g(x)}{g_0(x)} \frac{\gamma_j}{\alpha_j}\, \mathrm{d}x \\ &= \beta(S(g|g_0) - S(\mathbb{E}_{\mathbb{A}} g|\mathbb{E}_{\mathbb{A}} g_0)) \leqslant \beta S(g|g_0), \end{aligned} \tag{6.10}$$

since relative entropy is always non-negative. $\square$

Unfortunately, we cannot expect $-S(\mathbb{E}_{\mathbb{A}} g|\mathbb{E}_{\mathbb{A}} g_0)$ in (6.10) to give us much improvement on the stated bound (6.9); for it may well happen that each $|\alpha_j/\gamma_j - 1|$ is small, even when $\int_{A_j} |\hat{g} - g_0|$ is relatively large.

It is easy to show that $\|Bg\|_{L^\nu} \leqslant \beta \|g\|_{L^\nu}$ for the usual Lebesgue L^ν norm and $1 \leqslant \nu \leqslant \infty$. We have the crude bound

$$1 \leqslant \beta \leqslant (2^{-1} - \varepsilon_0)^{-1} := \beta_0, \tag{6.11}$$

whenever $g = Tg_0$ for a positive continuous density g_0. (In (7.3) below, we improve upon (6.11).) To see this, we simply apply Fubini's theorem to obtain

$$\int_{A_j} Tg_0(x)\mathrm{d}x \geqslant \frac{n_j}{N} \frac{\displaystyle\int_{A_j}\int_{A_j} K(x-y) g_0(y) \mathrm{d}y\, \mathrm{d}x}{\displaystyle\int_{A_j} g_0(y) \mathrm{d}y},$$

and we note that the inner integrand satisfies

$$\int_{A_j} K(x-y)\mathrm{d}x \geqslant 2^{-1} - \mathrm{erf}(|A_j|/h).$$

Proposition 6.4. *Suppose that $\hat{f}$ is a fixed point for T. Then*

$$\left| \int_{A_j} \hat{f} - \tilde{\alpha}_j \right| \leqslant \beta_0^2 h^{-1/2} \{h + O(\varepsilon_0)\} \|\hat{f}\|_{L^2}.$$

Thus the approximate formula of Proposition 6.1 holds for the bin quotas as $h \to 0_+$, provided that the fixed point remains bounded in L^2 norm. The fixed point converges to the histogram in the sense that

$$(\mathbb{E}_{\mathbb{A}}\hat{f} - \hat{f}_0)/\|\hat{f}\|_{L^2} \to 0 \qquad \text{as } h \to 0_+.$$

Proof. In view of (6.3), we need only bound the last term in

$$\int_{A_j} \hat{f} = \int_{A_j} T\hat{f} = \int_{A_j} K * \hat{f}_0 + \int_{A_j} K * \Delta\hat{f}.$$

Following the arguments of Lemma 6.2, we see that the main contribution to $\int_{A_j} K * \Delta\hat{f}$ is

$$\int_{A_j}\int_{A_j} \frac{n_j}{N\alpha_j} K(x-y)\left\{\hat{f}(y) - \frac{\alpha_j}{|A_j|}\right\}\mathrm{d}y\,\mathrm{d}x,$$

which by the definition of conditional expectation and the L^∞–L^1 duality is at most

$$\beta\|\hat{f} - \mathbb{E}_{\mathbb{A}}\hat{f}\|_{L^\infty} \int_{A_j} \left(1 - \int_{A_j} K(x-y)\mathrm{d}y\right)\mathrm{d}x.$$

Using the crude bound (6.11), and (6.4), we see that this is at most

$$4\beta_0\|\hat{f}\|_{L^\infty}\left(\frac{h}{2\pi}\right)^{1/2}.$$

By applying Young's inequality, we see that

$$\begin{aligned}\|\hat{f}\|_{L^\infty} &= \|K * B\hat{f}\|_{L^\infty} \leqslant \|K\|_{L^2}\|B\hat{f}\|_{L^2} \\ &\leqslant h^{-1/2}\beta\|\hat{f}\|_{L^2} \leqslant h^{-1/2}\beta_0\|\hat{f}\|_{L^2}.\end{aligned}$$

□

Theorem 6.5. *Suppose that: $\hat{f}$ is a fixed point for T, there exists a constant ρ with $0 < \rho < 1$ and g is a density for which*

$$\rho\left(1 + \frac{\tau\kappa}{2}\right)\int_{A_j} g \geqslant \frac{n_j}{N}, \qquad 1 \leqslant j \leqslant m. \tag{6.12}$$

Then Tg is closer to $\hat{f}$ than g is to $\hat{f}$, in the sense that

$$S(Tg|\hat{f}) \leqslant \rho S(g|\hat{f}).$$

Proof. By Corollary 5.2 we have

$$S(Tg|\hat{f}) = S(K * Bg|K * B\hat{f}) \leqslant (1 + \tau\kappa/2)^{-1} S(Bg|B\hat{f}),$$

and using (6.12) and our bound from Lemma 6.3 on the binning operator, we deduce that

$$S(Tg|\hat{f}) \leqslant \beta(1 + \tau\kappa/2)^{-1} S(g|\hat{f}) \leqslant \rho S(g|\hat{f}). \qquad \square$$

7. Stability

In order to use Theorem 6.5 in the iteration scheme, one must show that the condition (6.12) is preserved when one replaces g by Tg. In this section we investigate the stability of (6.12). By Proposition 6.4, one should expect a fixed point $\hat{f}$ for T to have $\int_{A_j} \hat{f} \approx \tilde{\alpha}_j$. We note that

$$\frac{n_j}{|A_j|N}\left\{1 - \frac{h}{|A_j|}\left(\frac{2}{\pi}\right)^{1/2}\right\} \leqslant \frac{\tilde{\alpha}_j}{|A_j|} \leqslant \max_k \frac{n_k}{N|A_k|} = \|\hat{f}_0\|_{L^\infty}, \qquad j = 1, 2, \ldots, m. \tag{7.1}$$

For constants c_1 and c_2, possibly depending upon the initial data and bandwidth, we introduce the space of (deficient) probability density functions on A,

$$\mathbb{M} = \left\{ g \in L^2(A) \middle| g \geqslant 0;\ \left|\int_{A_j} g - \tilde{\alpha}_j\right| \leqslant c_1 h^{1/2}\tilde{\alpha}_j;\ \|g - \mathbb{E}_{\mathbb{A}} g\|_{L^2} \leqslant c_2 \right\}. \tag{7.2}$$

The set $\mathbb{M}$ is closed in $L^2(A)$ and convex, and we shall show that it is mapped into itself by T. Evidently T restricted to $\mathbb{M}$ is completely continuous, and so by Schauder's theorem there exists at least one fixed point $\hat{f}$ in $\mathbb{M}$. On this set, the values of $\alpha_j = \int_{A_j} g$ are so close to the $\tilde{\alpha}_j$ that the norm of the binning operator is $\beta = \max_j(n_j/N\alpha_j)$, where

$$\beta \leqslant \bar{\beta} := \max_j \left\{1 - \left(\frac{2}{\pi}\right)^{1/2}\frac{h}{|A_j|} + \min_{\pm}\frac{n_{j+1}}{n_j}\left(\frac{2}{\pi}\right)^{1/2}\frac{h}{|A_j|}\right\}^{-1}\left\{1 + \frac{c_1 h^{1/2}}{(1 - c_1 h^{1/2})^2}\right\}. \tag{7.3}$$

We observe that $\bar{\beta}$ is close to 1 when $h \ll |A_j|$ and $c_1 h^{1/2} \ll 1$. By taking the bandwidth h small in comparison to the bin-width, we can also ensure that

$$\varepsilon_0 = \operatorname{erf}(|A_j|/h) \ll h. \tag{7.4}$$

The smoothing kernel K determines a convolution operator $L^2(\mathbb{R}) \to L^2(\mathbb{R})$ for which the restriction to $\{g \in L^2(A) | \mathbb{E}_{\mathbb{A}} g = 0\}$ has norm

$$\Lambda_1 = \sup\left\{ \left\| \int_A K(x - y)g(y)\mathrm{d}y \right\|_{L^2(\mathbb{R})} \middle| \|g\|_{L^2(A)} \leqslant 1;\ \int_{A_j} g = 0,\ j = 1, 2, \ldots, m \right\}.$$

This quantity is related to $\exp(-\tau\hat{\lambda}_1)$, where $\hat{\lambda}_1$ is the second smallest eigenvalue of $-\nabla^2$ in $L^2(A)$; thus Λ_1 is a spectral gap parameter (there is no weight involved here).

Theorem 7.1. *Suppose that $\hat{f}$ is a fixed point for T in $\mathbb{M}$ and one can choose c_1 and h so that:* (7.4) *holds;* $\Lambda_1\bar{\beta} < 1$*;*

$$(1 + h^2\kappa/2)(1 - c_1 h^{1/2}) > \frac{n_j}{N\tilde{\alpha}_j}; \tag{7.5}$$

$$\frac{3\bar{\beta}^2(1 + c_1 h^{1/2})^{1/2}}{c_1\pi^{3/4}} \leqslant \frac{\tilde{\alpha}_j}{\|\hat{f}_0\|_{L^\infty}^{1/2}}; \tag{7.6}$$

and

$$\frac{3\bar{\beta}^2}{c_1(1 - \bar{\beta}\Lambda_1)\pi^{3/4}} < \frac{\tilde{\alpha}_j}{\|\hat{f}_0\|_{L^\infty}^{1/2}}, \qquad j = 1, 2, \ldots, m. \tag{7.7}$$

Then the iteration scheme (1.6) *converges to $\hat{f}$ in relative entropy so that $S(\hat{f}_k|\hat{f}) \to 0$, geometrically as $k \to \infty$.*

Let us note that the right-hand sides of (7.5), (7.6) and (7.7) depend only upon the observed bin counts and the bandwidth h. By (7.1), $n_j/N\tilde{\alpha}_j \to 1$ as $h \to 0_+$; so (7.5) is a reasonable hypothesis. Further, (7.6) and (7.7) are realistic, in view of (7.1).

We proceed to check the stability of the defining conditions for $\mathbb{M}$ under the operation of T. We begin with some L^2 bounds.

Lemma 7.2. *Each $g \in \mathbb{M}$ satisfies*

$$\|g\|_{L^2} \leqslant (c_2^2 + (1 + c_1 h^{1/2})\|\hat{f}_0\|_{L^\infty})^{1/2}$$

and

$$\|Tg\|_{L^\infty} \leqslant \frac{\bar{\beta}}{(4\pi)^{1/4} h^{1/2}} \|g\|_{L^2}.$$

Proof. Exploiting orthogonality, we obtain

$$\int g^2 = \int |g - \mathbb{E}_{\mathbb{A}} g|^2 + \int |\mathbb{E}_{\mathbb{A}} g|^2 \leqslant c_2^2 + \sum_{j=1}^{m} \alpha_j^2/|A_j| \leqslant c_2^2 + \max_{1 \leqslant j \leqslant m} \alpha_j/|A_j|,$$

where we have used convexity at the last step. Since $g \in \mathbb{M}$, we can take advantage of the bounds on α_j to replace this by

$$c_2^2 + (1 + c_1 h^{1/2}) \max_{1 \leqslant j \leqslant m} \tilde{\alpha}_j/|A_j|;$$

whence the result, by (7.1).

We can convert this uniform L^2-bound on the elements of $\mathbb{M}$ into a uniform L^∞ bound on the elements of $T(\mathbb{M})$ by Young's inequality for convolution:

$$\|Tg\|_{L^\infty} = \|K * Bg\|_{L^\infty} \leqslant \|K\|_{L^2}\|Bg\|_{L^2} \leqslant \frac{\bar{\beta}}{(4\pi)^{1/4}h^{1/2}}\|g\|_{L^2}. \qquad \square$$

Proof of Theorem 7.1. We now check stability of the α_j under the operation of T. By (6.2) we have

$$\int_{A_j} Tg = \int_{A_j} K * B\hat{f}_0 + \int_{A_j} K * \Delta g. \tag{7.8}$$

By Lemma 7.2, the first term is close to $\tilde{\alpha}_j$, up to a negligible error of order ε_0. The last term in (7.8) may be bounded, as in the proof of Proposition 6.4, by

$$\left|\int_{A_j} K * \Delta g\right| \leqslant \bar{\beta}\|g - \mathbb{E}_{\mathbb{A}}g\|_{L^\infty}\left\{\int_{A_j}\left(1 - \int_{A_j} K(x-y)\mathrm{d}y\right)\mathrm{d}x \right.$$

$$\left. + \int_{A_{j\pm1}}\int_{A_j} K(x-y)\mathrm{d}y\,\mathrm{d}x + \sum_{k:|k-j|\geqslant 2}\int_{A_k}\int_{A_j} K(x-y)\mathrm{d}y\,\mathrm{d}x\right\}.$$

Using the argument that led to (6.8), we see that the first few summands give the main contributions and we can bound this expression by

$$\bar{\beta}\|g - \mathbb{E}_{\mathbb{A}}g\|_{L^\infty}\left\{\frac{3h}{(2\pi)^{1/2}} + O(\varepsilon_0)\right\}.$$

When $g = Tg_0$ for some $g_0 \in \mathbb{M}$, we can use Lemma 7.2 to bound

$$\|g - \mathbb{E}_{\mathbb{A}}g\|_{L^\infty} \leqslant \|Tg\|_{L^\infty} + \|\mathbb{E}_{\mathbb{A}}g\|_{L^\infty}$$

$$\leqslant \frac{\bar{\beta}}{(4\pi)^{1/4}h^{1/2}}\{c_2^2 + (1 + c_1h^{1/4})\|\hat{f}_0\|_{L^\infty}\}^{1/2} + (1 + c_1h^{1/2})\|\hat{f}_0\|_{L^\infty}.$$

To ensure stability of the condition in (7.2) on the α_j we need

$$\frac{\bar{\beta}^2}{(4\pi)^{1/4}h^{1/2}}\{c_2^2 + (1 + c_1h^{1/2})\|\hat{f}_0\|_{L^\infty}\}^{1/2}\frac{3h}{(2\pi)^{1/2}} + O(h) + O(\varepsilon_0) \leqslant c_1h^{1/2}\tilde{\alpha}_j,$$

which, in the presence of (7.4), is implied by both the conditions

$$3\bar{\beta}^2c_2 \leqslant \pi^{3/4}c_1\tilde{\alpha}_j, \qquad j = 1, 2, \ldots, m, \tag{7.9}$$

and

$$3\bar{\beta}^2(1 + c_1h^{1/2})\|\hat{f}_0\|_{L^\infty}^{1/2} \leqslant \pi^{3/4}c_1\tilde{\alpha}_j. \tag{7.10}$$

Evidently (7.10) is implied by (7.6). We shall now show that we can select c_1 so that the family of inequalities (7.9) can also be satisfied, together with stability of the bound on $\|g - \mathbb{E}_{\mathbb{A}}g\|_{L^2}$ in (7.2). We start by writing

$$Tg - \mathbb{E}_{\mathbb{A}}Tg = (K * \hat{f}_0 - \mathbb{E}_{\mathbb{A}}K * \hat{f}_0) + K * \Delta g - \mathbb{E}_{\mathbb{A}}K * \Delta g. \tag{7.11}$$

The first term on the right-hand side is independent of g and satisfies

$$\|K * \hat{f}_0 - \mathbb{E}_{\mathbb{A}} K * \hat{f}_0\|_{L^2} \leqslant \|K * \hat{f}_0\|_{L^2} \leqslant \|\hat{f}_0\|_{L^2} \leqslant \|\hat{f}_0\|_{L^\infty}^{1/2}. \tag{7.12}$$

To control the other terms, we use the fact that $\int_{A_j} \Delta g = 0$, from which it follows by definition of Λ_1 that

$$\|K * \Delta g\|_{L^2} \leqslant \Lambda_1 \|\Delta g\|_{L^2} \leqslant \bar{\beta}\Lambda_1 \|g - \mathbb{E}_{\mathbb{A}} g\|_{L^2} \leqslant \bar{\beta}\Lambda_1 c_2. \tag{7.13}$$

This deals with the final two terms in (7.11), since $\mathbb{E}_{\mathbb{A}}$ is an orthogonal projection making $\|K * \Delta g - \mathbb{E}_{\mathbb{A}} K * \Delta g\|_{L^2} \leqslant \|K * \Delta g\|_{L^2}$.

From (7.12) and (7.13) we see that stability of the L^2 bound in (7.2) is implied by

$$\|\hat{f}_0\|_{L^\infty}^{1/2} + \bar{\beta}\Lambda_1 c_2 \leqslant c_2. \tag{7.14}$$

By the hypothesis (7.7), we can select c_1 so that (7.14) may be satisfied together with (7.9) for some c_2.

Hence $K * \hat{f}_0 = T\hat{f}_0$ gives an element of $\mathbb{M}$, and furthermore all the iterates $\hat{f}_k = T^k \hat{f}_0$ ($k \geqslant 1$) belong to $\mathbb{M}$. Further, (7.5) implies that the $\hat{f}_k$ satisfy the hypothesis (6.12), and so by repeated application of Theorem 6.5 we have geometric convergence of the iterates $\hat{f}_k$ to $\hat{f}$ in the relative entropy metric, so $S(\hat{f}_k|\hat{f}) \to 0$. By Csiszár's inequality (4.1), this implies convergence in L^1 norm, so $\int |\hat{f}_k - \hat{f}| \to 0$ at a geometric rate as $k \to \infty$.

□

Thus, under the hypotheses of Theorem 7.1, Theorem 1.1 follows directly and Corollary 1.2 becomes a consequence of Proposition 4.2.

Acknowledgements

We should like to thank L. Gross, M. Ledoux, M. Titterington and our colleagues for helpful discussions. We also thank the editors and referee for improving the exposition. The work of JEK was supported by a Nuffield Foundation grant (NUF – NAL 99).

References

Azzalini, A. and Bowman, A.W. (1997) *Applied Smoothing Techniques for Data Analysis*. New York: Oxford University Press.

Barron, A.R. (1986) Entropy and the central limit theorem. *Ann. Probab.*, **14**, 336–342.

Barron, A.R. and Sheu, C.H. (1991) Approximation of density functions by sequences of exponential families. *Ann. Statist.*, **19**, 1347–1369.

Boneva, L.I., Kendall, D.G. and Stefanov, I. (1971) Spline transformations: Three new diagnostic aids for the statistical data-analyst (with discussion). *J. Roy. Statist. Soc. Ser. B*, **33**, 1–70.

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

Carlen, E.A. and Soffer, A. (1991) Entropy production by block variable summation and central limit theorems. *Comm. Math. Phys.*, **140**, 339–371.

Davies, E.B. (1989) *Heat Kernels and Spectral Theory.* Cambridge, Cambridge University Press.

Deuschel, J.-D. and Stroock, D.W. (1990) Hypercontractivity and spectral gap of symmetric diffusions with applications to the stochastic Ising models. *J. Funct. Anal.*, **92**, 30–48.

Dudley, R.M. (1989) *Real Analysis and Probability.* Pacific Grove, CA: Wadsworth and Brooks/Cole.

Gross, L. (1975) Logarithmic Sobolev inequalities. *Amer. J. Math.*, **97**, 1061–1083.

Gross, L. (1993) Logarithmic Sobolev inequalities and contractivity properties of semigroups. In G. Dell'Antonio and U. Mosco (eds), *Dirichlet Forms* (*Varenna, 1992*), Lecture Notes in Math. 1563, pp. 54–88. Berlin: Springer-Verlag.

Härdle, W.K. and Scott, D.W. (1992) Smoothing by weighted averaging of rounded points. *Comput. Statist.*, **7**, 97–128.

Heuser, H.G. (1982) *Functional Analysis*. New York: Wiley.

Kallenberg, O. (1997) *Foundations of Modern Probability.* New York: Springer-Verlag.

Koo, J.Y. and Kooperberg, C. (2000) Logspline density estimation for binned data. *Statist. Probab. Lett.*, **46**, 133–147.

Kooperberg, C. and Stone, C.J. (1991) A study of logspline density estimation. *Comput. Statist. Data Anal.*, **12**, 327–347.

Minnotte, M.C. (1996) The bias-optimized frequency polygon. *Comput. Statist.*, **11**, 35–48.

Minnotte, M.C. (1998) Achieving high-order convergence rates for density estimation with binned data. *J. Amer. Statist. Assoc.*, **93**, 663–672.

Otto, F. and Villani, C. (2000) Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, **173**, 361–400.

Rothaus, O.S. (1980) Logarithmic Sobolev inequalities and the spectrum of Sturm–Liouville operators. *J. Funct. Anal.*, **39**, 42–56.

Scott, D.W. and Sheather, S.J. (1985) Kernel density estimation with binned data. *Comm. Statist. Theory Methods*, **14**, 1353–1359.

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.

Stone, C.J., Hansen, M.H., Kooperberg, C. and Truong, Y.K. (1997) Polynomial splines and their tensor products in extended linear modelling. *Ann. Statist.*, **25**, 1454–1470.

Talagrand, M. (1996) Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.*, **6**, 587–600.

Titterington, D.M. (1983) Kernel-based density estimation using censored, truncated or grouped data. *Comm. Statist. Theory Methods*, **12**, 2151–2167.

Tobler, W.R. (1979) Smooth pycnophylactic interpolation for geographical regions. *J. Amer. Statist. Assoc.*, **74**, 519–536.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman & Hall.

Received July 1999, and revised August 2001 and February 2002