NONPARAMETRIC ESTIMATION OF A MIXING DENSITY VIA THE KERNEL METHOD

Constantinos Goutis

96-30

WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 96–30
Statistics and Econometrics Series 09
May 1996

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid 126
28903 Getafe (Spain)
Fax (341) 624–9849

NONPARAMETRIC ESTIMATION OF A MIXING DENSITY VIA THE KERNEL METHOD

Constantinos Goutis*

Abstract

We present a method to estimate the latent distribution for a mixture model. Our method is motivated by the standard kernel density estimation but instead of using an estimate based on the unobserved latent variables, we take the expectation with respect to their distribution conditional on the data. The resulting estimator is continuous and, hence, is appropriate when there is a strong belief in the continuity of the mixing distribution. We present an asymptotic justification and we discuss the associated computational problems. The method is illustrated by an example of fission track analysis where we estimate the density of the age of crystals.

Key words and phrases : Continuous mixtures, cross-validation, fission track analysis, kernel density estimation.

* Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, tel. (341) 624–9852, e-mail: *costas@est-econ.uc3m.es.* The research was partially carried out during a visit in and supported by CREST-Paris France. I would like to thank Christian Robert for his hospitality and useful discussions and Miguel Delgado and Ismael Sanchez for their help.

1 Introduction

Consider independent random variables x_i, $i = 1, 2, \ldots, n$ having a distribution of the form of a continuous mixture, that is, assume that there exist known distributions $h_i(x_i|y_i)$ and an unknown continuous $f(y)$ such that each x_i has a density

$$m_i(x_i) = \int h_i(x_i|y_i) f(y_i)\, dy_i. \tag{1.1}$$

The unobservable latent variables y_i are independent and identically distributed with density $f(y)$. The goal is to estimate the mixing density $f(y)$ without assuming any particular parametric form.

The nonparametric maximum likelihood of $f(y)$ is discrete with at most n mass points (Laird 1978, Lindsay 1983) and typically the number of mass points is considerably less than n. This is unsatisfactory if we have reasons to believe that $f(y)$ is indeed continuous. The situation is similar to the simpler case of observed y_i's. A nonparametric maximum likelihood estimate of $f(y)$ is a discrete distribution with mass points at y_i, but the vast literature on density estimation indicates that better solutions must be sought.

There are various other approaches to the problem. Perhaps the most extensively studied one uses a deconvolution of a kernel estimator of the observed data (Carroll and Hall 1988, Fan 1991, Liu and Taylor 1989, Stefanski and Carroll 1990, Zhang 1990 among others). However such methods are applicable essentially only in the errors in variables model, that is, when the differences $x_i - y_i$ are independent of each other and of the y_i's and have a common distribution. Other methods are motivated from computational considerations, such as a direct adaptation of the EM algorithm (Vardi and Lee 1993), stopping it before converging (Laird and Louis 1991) or smoothing each step (Eggermont and LaRiccia 1995, Silverman *et al.* 1985).

In this paper we propose a modification of the standard kernel density estimation method in order to estimate the mixing $f(y)$. The idea is simple: instead of using an estimate based on the unobserved y_i, take the expectation with respect to the distribution of y_i given the data x_i. Hence, our method is intuitive and straightforward to use, though it can be computer intensive. Furthermore, the existing techniques in the literature of kernel density estimation allow us to study the statistical properties of our method and, in particular, give some asymptotic justification.

The remainder of the paper is organised as follows: Section 2 motivates and presents the method. In Section 3 we discuss the asymptotic justification and Section 4 examines the problem of choosing the bandwidth of the kernel by cross-validation. Computational issues are tackled in Section 5. The method is illustrated by an example in Section 6 and we conclude with a discussion in Section 7.

2 The method

Our approach borrows heavily from standard kernel density estimation. For a probability density function $K(t)$, to be used as a kernel, if the variables y_i were observed, an estimate

of $f(y)$ would be

$$\check{f}(y) = \frac{1}{n}\sum_{i=1}^{n} K_\lambda(y - y_i), \tag{2.1}$$

where λ is the bandwidth and $K_\lambda(t) = (1/\lambda)K(t/\lambda)$. Of course $\check{f}(y)$ is not an estimator since it depends on unobservables, but, although the variables y_i are not known, we may know or at least estimate their conditional distribution given the data x_i,

$$g(y_i|x_i) = \frac{h(x_i|y_i)f(y_i)}{m(x_i)}\,. \tag{2.2}$$

In the above formula and in the remainder of the paper we omit the subscripts i from the densities for notational convenience. The distribution (2.2) can be used to modify (2.1) by taking the expectation of $\check{f}(y)$ with respect to $g(y_i|x_i)$, that is,

$$\tilde{f}(y) = \frac{1}{n}\sum_{i=1}^{n} \int K_\lambda(y - y_i)g(y_i|x_i)\,dy_i\,. \tag{2.3}$$

Again, the problem is that $g(y_i|x_i)$ of (2.2) is unknown, since $f(y_i)$ is unkown. Nevertheless, we can use

$$\hat{g}(y_i|x_i) = \frac{h(x_i|y_i)\hat{f}(y_i)}{\hat{m}(x_i)}, \tag{2.4}$$

where

$$\hat{m}(x_i) = \int h(x_i|y_i)\hat{f}(y_i)\,dy_i \tag{2.5}$$

is the estimated marginal distribution of x_i. This suggests the legitimate estimate

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} \int K_\lambda(y - y_i)\hat{g}(y_i|x_i)\,dy_i\,, \tag{2.6}$$

where $\hat{g}(y_i|x_i)$ is given by (2.4)-(2.5). Equality (2.6) is indeed a functional equation since $\hat{f}(y)$ also enters in the right hand side, so the problem is to find a function that satisfies (2.6). Before examining computational issues, we will give some asymptotic justification of the estimate $\hat{f}(y)$.

3 Asymptotics

We take the kernel $K(t)$ to be a symmetric probability function with second moment equal to 1. Furthermore, we assume that the density $f(y)$ to be estimated has an integrable second derivative. The asymptotic results will be obtained under the convergence conditions

$$n \to \infty,\ \lambda \to 0,\ n\lambda \to \infty. \tag{3.1}$$

Standard measures of global discrepancy of a density estimate $\hat{f}(y)$ from the true $f(y)$ are the integrated squared error

$$ISE(\hat{f}(y)) = \int (\hat{f}(y) - f(y))^2\,dy \tag{3.2}$$

or the mean integrated squared error,

$$MISE(\hat{f}(y)) = E\int(\hat{f}(y) - f(y))^2\,dy. \tag{3.3}$$

The first theorem states a result for $\tilde{f}(y)$. Its proof can be found in the Appendix.

Theorem 3.1 *The estimator $\tilde{f}(y)$ is consistent for $f(y)$. In particular, the $MISE(\tilde{f}(y))$ is less than*

$$MISE(\tilde{f}(y)) = \frac{\lambda^4}{4}\int f''(y)\,dy + \frac{1}{n\lambda}\int K(t)^2dt + o(\lambda^4 + \frac{1}{n\lambda}). \tag{3.4}$$

Of course the above theorem is of little use by itself, since $\tilde{f}(y)$ depends on the unknown $g(y_i|x_i)$. However, the next theorem says that even if we use estimates, the resulting estimator of $f(y)$ will be consistent as long as the estimates of $g(y_i|x_i)$ are not terrible. The proof can be found in the Appendix.

Theorem 3.2 *Suppose that we have an estimate $\bar{g}(y|x)$ of $g(y|x)$ so that for every x*

$$E\int[\bar{g}(y|x) - g(y|x)]^2\,dy \tag{3.5}$$

is bounded for a sufficiently large n. Define $\breve{f}(y)$ by

$$\breve{f}(y) = \frac{1}{n}\sum_{i=1}^{n}\int K_\lambda(y - y_i)\bar{g}(y_i|x_i)\,dy_i\,. \tag{3.6}$$

Then the estimate $\breve{f}(y)$ is consistent for $f(y)$ in the $MISE$ sense and the rate of convergence of $MISE(\breve{f}(y))$ is the same as that of $\tilde{f}(y)$.

The expectation in (3.5) is taken over both distributions of the sample and the random variable x. The problem with this Theorem is that typically it is difficult to examine the properties of (3.5) for a given $\bar{g}(y|x)$ since the latter involves the random variable x. It is easy to show that, under some mild conditions, any consistent (in the ISE sense) estimate $\bar{f}(y)$ will yield a $\bar{g}(y|x)$ for which $\int[\bar{g}(y|x) - g(y|x)]^2\,dy$ will tend to zero in probability for any x. We suspect that, under some regularity conditions, a consistent (in the $MISE$ sense) $\bar{f}(y)$ will yield a $\bar{g}(y|x)$ satisfying the assumption of Theorem 3.2. If this is true, it is somewhat at odds with the pessimistic results of Fan (1991) and Zhang (1990). These articles, in a slightly different setup, suggest that it might be too much to expect the same rate of convergence for a general density $f(y)$ as if we had observed y_i. In any case the theorem serves as a motivation for the estimator $\hat{f}(y)$ of (2.6), where we simply take $\bar{g}(y|x) = \hat{g}(y|x)$.

4 Cross validatory choice of λ

Of course, the above method requires a choice of the bandwidth λ and the kernel $K(t)$. Conventional wisdom says that the choice of $K(t)$ is not that important for the statistical properties of a density estimator. Since the normal kernel has computational advantages we will use this. The results, however, depend heavily on λ and a careful choice is necessary. There are various methods for selecting the bandwidth in standard kernel density estimation, surveyed by Marron (1988) and Chiu (1996). Least squares cross-validation (Rudemo 1982, Bowman 1984) is perhaps the most intuitive one and has a strong asymptotic justification (Hall 1983, Hall and Marron 1987, Stone 1984).

The cross-validatory λ for $\check{f}(y)$ minimises

$$\int \check{f}(y)^2(y)\,dy - \frac{2}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{j\neq i}K_\lambda(y_i - y_j), \tag{4.1}$$

where each term of the outer sum is the value at y_i of the estimate of $f(y)$ derived from the remaining data points. The justification is that the above cross-validatory score is an unbiased estimate of the loss $ISE(\check{f}(y)) - \int f^2(y)\,dy$, so choosing a value minimising (4.1) will yield an estimator with small $ISE(\check{f}(y))$.

Clearly in order to use this criterion we must adapt it to the fact that we do not actually observe y_i. In a fashion similar to Section 2 we replace $\check{f}(y)$ by its expectation given x_i and consider the cross-validatory score

$$\int \tilde{f}^2(y)\,dy - \frac{2}{n}\sum_{i=1}^{n}\int \tilde{f}_{-i}(y_i)g(y_i|x_i)\,dy_i \tag{4.2}$$

where

$$\tilde{f}_{-i}(y_i) = \frac{1}{n-1}\sum_{j\neq i}\int K_\lambda(y_i - y_j)g(y_j|x_j)\,dy_j. \tag{4.3}$$

In order to examine if this is a valid cross-validatory score we must see what it is an unbiased estimator of. The following theorem shows that this is indeed the case.

Theorem 4.1 *If $\tilde{f}_{-i}(y_i)$ is given by (4.3) then*

$$E\int \tilde{f}(y)f(y)\,dy = E\frac{1}{n(n-1)}\sum_{i=1}^{n}\tilde{f}_{-i}(y_i). \tag{4.4}$$

Substituting $\tilde{f}^2(y)$ in (4.2) and replacing $n(n-1)$ by n^2, a standard development shows that the cross-validatory score that we want to minimise has the form

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n^2\lambda}\int\int K^*(\frac{y_i - y_j}{\lambda})g(y_i|x_i)g(y_j|x_j)\,dy_i\,dy_j + \frac{2}{n\lambda}K(0) \tag{4.5}$$

where $K^*(t) = K\star K(t) - 2K(t)$ and $\star$ denotes convolution. Clearly (4.5) contains the unknown $g(y_i|x_i)$ and $g(y_j|x_j)$ that have to be replaced by $\hat{g}(y_i|x_i)$ and $\hat{g}(y_j|x_j)$ respectively,

so we minimise

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n^2\lambda}\int\int K^*(\frac{y_i-y_j}{\lambda})\hat{g}(y_i|x_i)\hat{g}(y_j|x_j)\,dy_i\,dy_j+\frac{2}{n\lambda}K(0). \tag{4.6}$$

Since the conditional densities also depend on λ, any minimisation of (4.6) must be done in some iterative way. In the next section we discuss this issue, as well as the computation of $\hat{f}(y)$ itself.

5 Computational issues

We consider first the solution of the defining equation (2.6) and then the efficient computation of the cross-validatory score (4.6). The form of (2.6) suggests an EM type solution, by completing the data by y_i. For the complete data (x_i, y_i), an estimate of the density is $\check{f}(y)$. This, though not a maximisation, gives us the step corresponding to the M step. The E step involves the expectation of $\check{f}(y)$ with respect to the current density estimate. This suggests the iterative equation

$$\hat{f}^{(t+1)}(y)=\frac{1}{n}\sum_{i=1}^{n}\int K_\lambda(y-y_i)\frac{h(x_i|y_i)\hat{f}^{(t)}(y_i)}{\int h(x_i|y_i)\hat{f}^{(t)}(y_i)\,dy_i}\,dy_i\,, \tag{5.1}$$

for $t=0,\,1,\,2,\ldots$ starting from some initial estimate $\hat{f}^{(0)}(y)$.

Noting that the right hand side of (5.1) is a sum of expectations of $K_\lambda(y-y_i)$, perhaps the best way is to simulate s_0 random variables $y_i^{(s)}$ with distribution proportional to $h(x_i|y_i)\hat{f}^{(t)}(y_i)$ and then take the average over the simulated values of $K_\lambda(y-y_i^{(s)})$, that is, take

$$\hat{f}^{(t+1)}(y)=\frac{1}{ns_0}\sum_{i=1}^{n}\sum_{s=1}^{s_0}\int K_\lambda(y-y_i^{(s)}). \tag{5.2}$$

Acceptance rejection methods are preferable to numerical integration or other simuation methods in this case since they do not require the computation of the normalising constants $\int h(x_i|y_i)\hat{f}^{(t)}(y_i)\,dy_i$. Furthermore, in many cases $h(x_i|y_i)$, considered as a function of y_i, is a density up to a constant. Then an obvious choice is to simulate y_i from the density proportional to $h(x_i|y_i)$, and the simulation algorithm will be efficient since $\hat{f}^{(t)}(y_i)$ will be much more spread than $h(x_i|y_i)$. In practice, the distribution $\hat{f}^{(t+1)}(y_i)$ will have to be computed over a finite grid.

The minimisation of (4.6) can follow a similar pattern. In the case of a normal kernel, we can use the Fourier methods described by Silverman (1982, 1986). The adaptation of the method for our purposes is as follows:

(0) Choose an equally spaced finite grid $t_0,\,t_1,\,t_2,\ldots,t_M$ over the effective support of $f(y)$, for M equal to a power of 2. Let $\delta=t_{k+1}-t_k$ and $u_l=2\pi l(M\delta)^{-1}$ for $-(M/2)\le l\le M/2$.

(1) Evaluate $\hat{f}(y)$ at the points $t_0,\,t_1,\,t_2,\ldots,t_M$ and simulate

$$y_i^{(s)}\sim\hat{g}(y_i|x_i)\propto h(x_i|y_i)\hat{f}(y_i). \tag{5.3}$$

If an acceptance rejection method is used, then, in practice, $y_i^{(s)}$ will be equal to one of the t_k's.

(2) Take

$$\xi_k^{(s)} = \frac{1}{n\delta} \sum_{i=1}^{n} \mathbf{I}\{y_i^{(s)} = t_k\}. \tag{5.4}$$

(3) For $-(M/2) \leq l \leq M/2$ compute

$$Y_l^{(s)} = \frac{1}{M} \sum_{k=0}^{M-1} \xi_k^{(s)} \exp(i2\pi kl/M), \tag{5.5}$$

where $i = \sqrt{-1}$, by fast Fourier transform and calculate $\left|Y_l^{(s)}\right|^2$.

(4) Repeat steps (1)-(3) s_0 times and compute the average

$$\bar{A}_l = \frac{1}{s_0} \sum_{s=1}^{s_0} \left|Y_l^{(s)}\right|^2. \tag{5.6}$$

Then find the λ that minimises

$$M\delta \sum_{l=1}^{M/2} \left\{ \exp\left(-\lambda^2 u_l^2\right) - 2\exp\left(-\frac{1}{2}\lambda^2 u_l^2\right) \right\} \bar{A}_l + \frac{1}{n\lambda\sqrt{2\pi}}. \tag{5.7}$$

(5) Update $\hat{f}(y)$ and repeat steps (1)-(4) till convergence.

In practice, we can fold the two updating schemes in a single one and avoid extra simulations. In other words, the simulated values in step (1) above can be used also to update $\hat{f}^{(t)}(y)$ to $\hat{f}^{(t+1)}(y)$ as descibed by (5.2). If we do so, we always simulate from the first step of (5.2) this will not matter if the values of λ do not change anymore. For values of λ far from the minimising one, we may want to run the updating (5.2) a few times before updating the λ but it is not crucial that we have actually found the fixed point.

6 Example

We now illustrate our method by an example of fission track analysis. The modelling of "mixed" fission track ages can provide estimates of times and temperatures that are of interest in the oil exploration industry and in various geological applications (see Hurford, 1991, for a review). Table 1 (reproduced from Goutis and Galbraith 1995) shows a set of data, which are numbers of spontaneous and induced fission tracks counted in matched areas of crystal and mica for 27 zircon crystals. Spontaneous tracks form over geological time by spontaneous fission of trace ^{238}U. Induced tracks are created artificially by placing the sample in a nuclear reactor and bombarding it with thermal neutrons, a measured proportion of which collide with trace ^{235}U atoms, thereby causing them to fission. This indirectly measures the amount of trace uranium in the crystal.

Galbraith and Laslett (1993) considered statistical models for such data and give more details on the background. It is supposed that the numbers of spontaneous and induced tracks, R and S, counted over matched areas A for a single crystal, have conditionally

Table 1.
Numbers of spontaneous and induced fission tracks
counted in matched areas for 27 zircon crystals:

crystal	R	S	area	crystal	R	S	area
1	24	459	80	15	2	70	49
2	8	52	30	16	3	94	28
3	136	310	30	17	23	128	60
4	56	257	70	18	153	264	70
5	3	57	70	19	90	143	32
6	6	332	80	20	31	49	16
7	73	98	14	21	38	120	40
8	131	226	50	22	51	46	25
9	9	173	80	23	38	85	12
10	6	28	12	24	127	45	20
11	141	229	70	25	5	24	30
12	11	74	36	26	24	56	20
13	12	61	18	27	10	31	18
14	10	28	40				

independent Poisson distributions with means $A\rho_1$ and $A\rho_2$ respectively. In this context the Poisson model is particularly convincing (Galbraith *et al.* 1990). The spontaneous track mean density ρ_1 depends on the age of the crystal, the amount of trace ^{238}U it contains, and the mean length of spontaneous tracks. The induced track density ρ_2 depends on the amount of trace ^{235}U and on the mean length of induced tracks; ρ_2 also depends on the thermal neutron dose, which is measured independently. To a close approximation, the ratio ρ_1/ρ_2 is given by

$$\frac{\rho_1}{\rho_2} = \frac{2\lambda_f}{\Phi\sigma_f I} T \frac{l_1}{l_2} \tag{6.1}$$

which depends on the crystal's age T, which is of interest in this context, the ratio l_1/l_2 of mean lengths of spontaneous and induced tracks, which reflects the amount of heat the crystal has experienced, the ^{235}U:^{238}U isotopic ratio I, which is usually assumed to be fixed, the thermal neutron dose Φ and constants λ_f and σ_f that are independently calibrated.

Typically the amounts of trace uranium and areas vary substantially between crystals (as they do in Table 1). In a sample of crystals the ratios ρ_1/ρ_2 will vary if the crystals have different ages. They may also vary due to the effect of heat (particularly for the mineral apatite), even if all crystals have the same age, because the spontaneous tracks will shorten, possibly by different amounts for different crystals, so that l_1/l_2 varies.

Goutis and Galbraith (1995) developed a parametric model that allows for variation between crystals of ρ_1, ρ_2 and of ρ_1/ρ_2. Their model was based on a Wishart mixing of the Poisson counts. In this paper, we take a different approach. First we condition on

the total number of counts in each crystal. Conditionally, the number of spontaneous tracks has a binomial distribution with parameters $\rho_1/(\rho_2+\rho_2)$ and $R+S$. Although in the parametric case, in the presense of extra-Poisson variation the statistic $R+S$ is not ancillary and conditioning is not justified, in a nonparametric context concepts such as sufficiency or ancillarity are not defined. Hence it is legitimate to simplify the problem by conditioning and considering a mixing distribution on the log-odds of the binomial distribution. Translating to the notation of (1.1), for the ith crystal, x_i is the number of spontaneous tracks, y_i is $\log(\rho_{1i}/\rho_{2i})$ and $h_i(x_i|y_i)$ is $\mathcal{B}(n_i, [1+\exp(-y_i)]^{-1})$, where n_i is the total number of tracks. The distribution of the ages of the crystals can be readily deduced from the distribution of y_i.

Figure 1 shows the density estimate of the logarithm of the age of the crystals based on the data of Table 1. The ticks on the x-axis indicate the location of the crude age estimates based on the empirical logits $\log[(x_i+0.5)/(n_i-x_i+0.5)]$. The size of the ticks is proportional to $\sqrt{n_i}$. This indicates the precision of the observation and gives a measure of the weight that we would like this observation to have. We computed the density for two values of the smoothing parameter λ. One estimate is for $\lambda \approx 0.5$, the cross-validatory choice, and the other was for $\lambda = 0.3$. For comparative purposes, we include on the figure a crude density estimate, derived by considering the empirical logits as "data" and applying the standard kernel method. The maximum likelihood estimate is also drawn as spikes.

We were somewhat unsatisfied with cross-validation since it seems to oversmooth the data. Experience with simulated data (not shown here) indicated that this happens often. We tried a few values smaller than the cross validatory choice and chose, somewhat arbirtrarily, $\lambda = 0.3$. This gives an aesthetically pleasing picture for the density estimate. The crude estimate based on the empirical logits does not take into account the differences in the distributions $h_i(x|y)$ and misses the differences in precisions. The $\hat{f}(y)$ for $\lambda = 0.3$ shows that there are many high precision observations between 10 Ma and 20 Ma and does not smooth out the high observation. This is ignored by the crude estimate, and as expected, the effect of the high observation is diminished by taking a larger λ. The maximum likelihood estimate suggests a large, somewhat unpalatable, gap in ages between about 1.8 Ma and 6 Ma.

7 Discussion

The method that we propose is an all purpose method for computing an estimate of the mixing density. It is worthwhile noting that it is equally applicable for discrete and for continuous data and does not require exchangeability of the observable random variables. As long as the latent variables have a common distribution, the conditional distributions of the data can have any form. Clearly, it should be used if we believe that the mixing density is continuous, in which case a nonparametric maximum likelihood estimate is not satisfactory. It is often argued (Lindsay 1995) that there is too little information about the mixing distribution to make sensible inferences about its shape. However, any kind of nonparametric estimation makes inferences on the shape, and maximum likelihood infers that the cumulative distribution is a step function. Obviously, if we believe that

a parametric model gives a reasonable approximation, on parsimony grounds one should abandon the nonparametric approach altogether.

Our starting point is the kernel density estimate using the unobserved data. An alternative approach would be to use some kind of penalised likelihood, where one maximises the log-likelihood minus a roughness penalty. This would automatically rule out discontinuous estimates and, by a suitable choice of the weight of the two quantities, loglikelihood and roughness penalty, could give sensible results. We have not pursued this approach, but it would be interesting to do so and compare the two methods.

A more complete asymptotic analysis would also be of interest. Of course, the problem of estimating a density if we observe data contaminated with noise is more difficult than without noise. We suspect that this will reflect on the constants rather than the rate of convergence, which might be the typical $o(n^{-4/5})$, at least in some cases. However, the asymptotic properties of our method seem technically difficult to establish and beyond the scope of this article.

Clearly asymptotics are one aspect of the problem. A more important one is the performance for data sets, which are always finite. Based on our experience on the example of Section 6 and on some limited simulation results, we are optimist. The method gave a believable answer. Some care needs to be taken with the choice of the bandwidth, since we suspect that an automatic least squares cross-validatiotory choice will oversmooth the data and the estimated density will be more spread than it should. Though our results are not conclusive, they suggest that it is worthwhile trying the method to other data sets.

Appendix

Proof of Theorem 3.1. Consider the expectation and the variance of $\tilde{f}(y)$. We have

$$E(\tilde{f}(y)) = \frac{1}{n}\sum_{i=1}^{n}\int\int K_\lambda(y-y_i)g(y_i|x_i)\,dy_i\; m(x_i)\,dx_i \tag{A.1}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int\int K_\lambda(y-y_i)h(x_i|y_i)\,dx_i\; f(y_i)\,dy_i \tag{A.2}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int K_\lambda(y-y_i)f(y_i)\,dy_i \tag{A.3}$$

$$= \int K_\lambda(y-z)f(z)\,dz. \tag{A.4}$$

Similarly

$$n\;var(\tilde{f}(y)) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\int\Big[\int K_\lambda(y-y_i)g(y_i|x_i)\,dy_i\Big]^2 m(x_i)\,dx_i - \Big[\int\int K_\lambda(y-y_i)g(y_i|x_i)\,dy_i\; m(x_i)\,dx_i\Big]^2\Big\} \tag{A.5}$$

$$\le \frac{1}{n}\sum_{i=1}^{n}\Big\{\int\int [K_\lambda(y-y_i)]^2\, g(y_i|x_i)\,dy_i\; m(x_i)\,dx_i - \Big[\int\int K_\lambda(y-y_i)g(y_i|x_i)\,dy_i\; m(x_i)\,dx_i\Big]^2\Big\} \tag{A.6}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big\{\int [K_\lambda(y-y_i)]^2 f(y_i)\,dy_i - \Big[\int K_\lambda(y-y_i)f(y_i)\,dy_i\Big]^2\Big\} \tag{A.7}$$

$$= \int [K_\lambda(y-z)]^2 f(z)\,dz - \Big[\int K_\lambda(y-z)f(z)\,dz\Big]^2 . \tag{A.8}$$

The rest is standard development (see e.g. Silverman 1986). □

Proof of Theorem 3.2. We have

$$(\check{f}(y)-\tilde{f}(y))^2 = \Big[\frac{1}{n}\sum_{i=1}^{n}\int K_\lambda(y-y_i)\,(\bar{g}(y_i|x_i)-g(y_i|x_i))\;dy_i\Big]^2 \tag{A.9}$$

$$\le \frac{2}{n^2}\sum_{i=1}^{n}\Big[\int K_\lambda(y-y_i)\,(\bar{g}(y_i|x_i)-g(y_i|x_i))\;dy_i\Big]^2 \tag{A.10}$$

$$\le \frac{2}{n^2}\sum_{i=1}^{n}\int K_\lambda(y-y_i)\,[\bar{g}(y_i|x_i)-g(y_i|x_i)]^2\;dy_i \tag{A.11}$$

where the last step follows by noting that, considered as a function of y_i, $K_\lambda(y-y_i)$ is a density. Integrating over y, and using the fact that $K_\lambda(y-y_i)$ is also a density in y, we obtain

$$\int(\check{f}(y)-\tilde{f}(y))^2dy \le \frac{2}{n^2}\sum_{i=1}^{n}\int [\bar{g}(y_i|x_i)-g(y_i|x_i)]^2\;dy_i\,, \tag{A.12}$$

Now we take expectations over both sides of (A.12). The asymptotic properties of $\bar{g}(y_i|x_i)$ are the same as the ones of $\bar{g}(y_i|x_i)$ for a given x_i, so we may as well consider x_i to be a random variable independent of the sample. If the expectation of each term of the sum is bounded, the Cesaro averages will also be bounded and the expectation of the left hand side (A.12) will be $O(n^{-1})$. Noting that

$$(\check{f}(y) - f(y))^2 \leq 2\left[(\check{f}(y) - \tilde{f}(y))^2 + (\tilde{f}(y) - f(y))^2\right] \tag{A.13}$$

establishes the result. □

Proof of Theorem 4.1. Consider the expectation of $\tilde{f}_{-i}(y_i)$. This is a function of the random variables y_i and x_j, $j \neq i$. Conditioning first on all $x_1,\ x_2, \ldots, x_n$, we have

$$E\left[\tilde{f}_{-i}(y_i)|x_1,\ x_2, \ldots, x_n\right] = \sum_{j\neq i} \int\int K_\lambda(y_i - y_j) g(y_i|x_i) g(y_j|x_j)\, dy_i\, dy_j. \tag{A.14}$$

To find the expectation of the above we consider each term of the sum separately. Since the jth term is a function of x_i and x_j we multiply by $m(x_i)m(x_j)$ and integrate with respect to x_i and x_j. Following the proof of Theorem 3.1, we interchange the order and integrate $dx_i\, dx_j$. After summing over $j \neq i$ we obtain the expectation of (A.14) and after summing over i we have

$$E\frac{1}{n(n-1)}\sum_{i=1}^{n}\tilde{f}_{-i}(y_i) = \sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{n(n-1)}\int\int K_\lambda(y_i - y_j) f(y_i) f(y_j)\, dy_i\, dy_j \tag{A.15}$$

$$= \int\int K_\lambda(y - z) f(y) f(z)\, dy\, dz. \tag{A.16}$$

On the other hand, from (A.4) we obtain

$$E\int \tilde{f}(y) f(y)\, dy = \int\int K_\lambda(y - z) f(y) f(z)\, dy\, dz \tag{A.17}$$

which completes the proof. □

References

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

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83, 1184-1186.

Chiu, S.-T. (1996). A comparative review of bandwidth selection for kernel density estimation. *Statistica Sinica*, 6, 129-145.

Eggermont, P. P. B. and LaRiccia, V. N. (1995). Maximum smoothed likelihood density estimation for inverse problems. *Ann. Statist.*, 23, 199-220.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19, 1257-1272.

Galbraith, R. F. and Laslett, G. M. (1993). Statistical models for mixed fission track ages. *Nuclear Tracks and Radiation Measurements,* 21, 459-470.

Galbraith, R. F., Laslett, G. M., Green, P. F. and Duddy, I. R. (1990). Apatite fission track analysis: geological thermal history analysis based on a three dimensional random process of linear radiation damage. *Phil. Trans. R. Soc. Lond.* A, 332, 419-438.

Goutis, C. and Galbraith, R. F. (1995). A parametric model for heterogeneity in paired Poisson counts. Working Paper 95-61, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Spain.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.*, 11, 1156-1174.

Hall, P. and Marron, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Prob. Theory Related Fields*, 74, 567-581.

Hurford, A. J. (1991). Uplift and cooling pathways derived from fission track analysis and mica dating: a review. *Geologisches Rundschau,* 80, 349-368.

aird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.*, 73, 805-811.

ird, N. M. and Louis, T. A. (1991). Smoothing the non-parametric estimate of a prior distribution by roughening: a computational study. *Comput. Statist. Data Anal.*, 12, 27-37.

dsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, 11, 86-94.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5. Hayward, CA: IMS.

Liu, M. C. and Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canad. J. Statist.*, 17, 427-438.

Marron, J. S. (1988). Automatic smoothing parameter selection: a survey. *Empirical Economics*, 13, 187-208.

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

Silverman, B. W. (1982). Kernel density estimation using the fast Fourier transform. Statistical Algorithm AS 176. *Appl. Statist.*, 31, 93-97.

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

Silverman, B. W., Jones, M. C., Wilson, J. D. and Nychka, D. W. (1990). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). *J. R. Statist. Soc. Ser.* B, 52, 271-324.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 2, 169-184.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12, 1285-1297.

Vardi, Y. and Lee, D. (1993). From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems (with discussion). *J. R. Statist. Soc. Ser.* B, 55, 569-612.

Zhang, C. H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.*, 18, 806-830.

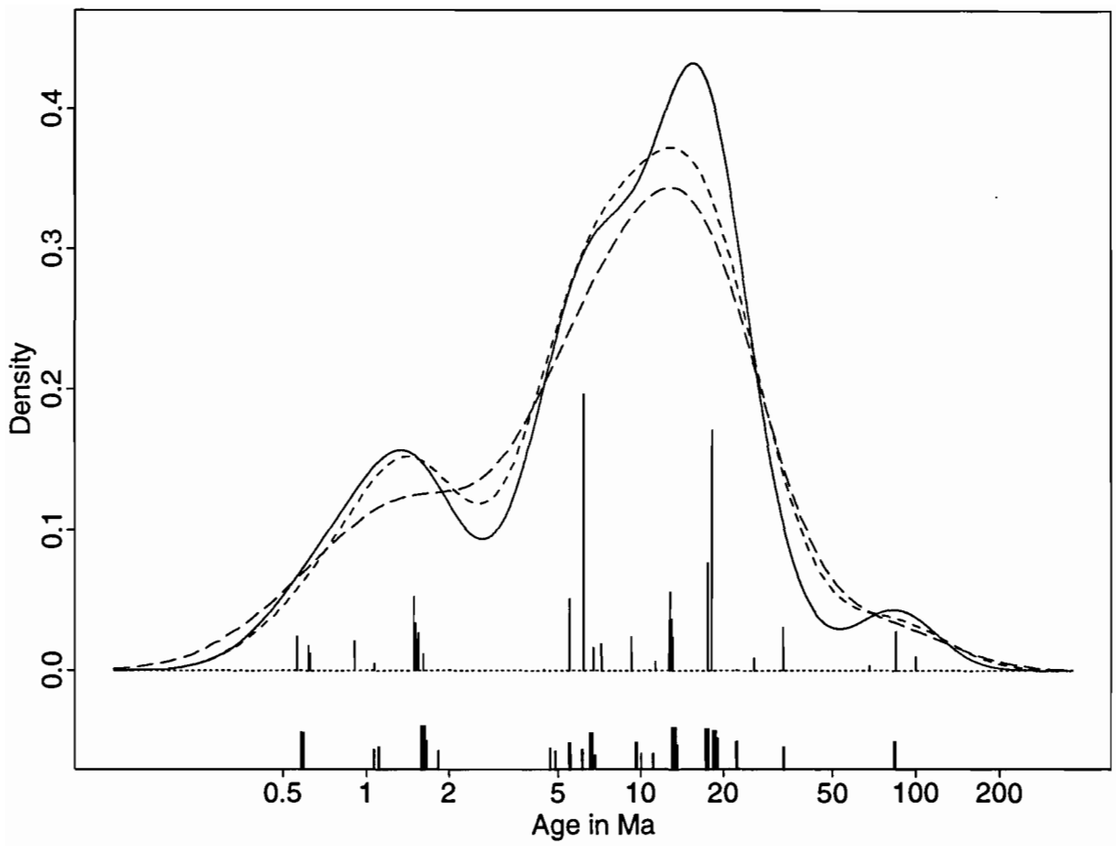

Figure 1: Density estimate of the age of the crystals data. The solid line is based on a bandwidth $\lambda = 0.3$, and the long-dashed line on $\lambda \approx 0.5$. The short-dashed line is the kernel density estimate based on the empirical logits and the vertical spikes indicate the maximum likelihood estimate.