J. R. Statist. Soc. B (1983),
45, *No.* 2, *pp.* 258–266

Kernel-based Density Estimates from Incomplete Data

By D. M. TITTERINGTON† and G. M. MILL

University of Glasgow, Scotland

[Received November 1981. Revised August 1982]

SUMMARY

The problem of using non-parametric methods to estimate multivariate density functions from incomplete continuous data does not appear to have been considered before. Methods of producing kernel functions on incomplete observations are suggested involving averaging over the missing variables or substitution of them by simulated values. Consistency of the procedures in terms of integrated mean squared error is investigated and optimal choice of smoothing parameter is discussed. Application in terms of imputation for missing values is discussed.

Keywords: DENSITY ESTIMATION; BIVARIATE DATA; MULTIVARIATE DATA; MISSING DATA; KERNEL; IMPUTATION

1. INTRODUCTION

A fundamental problem in statistics is that of estimating a multivariate density function using a data set of independent observations from the population under investigation. Parametric or non-parametric methods are available to deal with this and most have been adapted to deal with the practical problems generated by missing data. For many parametric problems, versions of the *EM* algorithm (Dempster *et al.*, 1977) exist for the computation of maximum likelihood estimates and Murray (1979) shows how to derive Bayesian predictive densities for estimating multivariate Normal densities. Some suggestions for adapting kernel-based methods of smoothing categorical data are described in Titterington (1977) and in Murray and Titterington (1978). The derivation of non-parametric density estimates based on continuous observations, some of which are incomplete, does not appear to have been considered before. Here we make some suggestions related to kernel-based density estimates for the case where "incompleteness" means that some variables in some observations are missing.

The proposals are to average out the density estimate over the missing values and to impute, by simulation, one or more substitutes for each missing value. The former suggestion, which is computationally heavier, corresponds to simulating infinitely many imputations per missing value. It turns out that both methods produce consistent density estimates, that the imputed values do not distort the data and that the substitution method with only a few imputations per missing value is almost as efficient as the averaging method.

As with other treatments of missing data it will be assumed that the data are missing "at random"; see Rubin (1976).

2. GENERAL APPROACHES

Given a data set $D = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$ of n complete independent observations from a population with *p.d.f.* $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^k$, the kernel method estimates $f(\mathbf{x})$ by

$$\hat{f}(\mathbf{x} \mid D, h) = \frac{1}{n} \sum_{i=1}^{n} K(\mathbf{x} \mid \mathbf{x}^{(i)}, h), \tag{1}$$

† *Present address*: Professor D. M. Titterington, Department of Statistics, University Gardens, Glasgow G12 8QW, Scotland.

 0035–9246/83/45258 $2.00

where, for $i = 1, \ldots, n$, $K(\cdot \mid \mathbf{x}^{(i)}, h)$, the kernel function, is itself a density function, usually with its mode at $\mathbf{x}^{(i)}$, and h is a parameter, possibly vector, to be chosen by the user, which dictates the smoothness of the resulting density estimate. It is generally thought that the choice of the form of K is not important (and this can be exploited to practical advantage), but that the choice of h is; see Wegman (1972), Fryer (1977), Silverman (1978).

If the components of $\mathbf{x}$ are assumed independent then, instead of (1), $f(\cdot)$ can be taken to be the product of univariate density estimates and the incompleteness of the data does not cause difficulty. In general, however, we have to decide what to use for

$$K\{\cdot \mid (\mathbf{y}, \mathbf{z}), h\},$$

where $(\mathbf{y}, \mathbf{z})$ denotes an observation of which $\mathbf{z}$ represents the missing components. Two basic suggestions arise naturally.

(A) Use $K\{\cdot \mid (\mathbf{y}, \hat{\mathbf{z}}), h\}$ where $\hat{\mathbf{z}}$ is an estimate of $\mathbf{z}$.

(B) Use $\int K\{\cdot \mid (\mathbf{y}, \mathbf{z}), h\} \hat{f}(\mathbf{z} \mid \mathbf{y}) dz$, where $\hat{f}(\mathbf{z} \mid \mathbf{y})$ is a *p.d.f.*

Thus the choices of $\hat{\mathbf{z}}$ in (A) and of $\hat{f}(\cdot \mid \mathbf{y})$ in (B) have to be resolved so that suggestion (B) in particular is workable. Whether (A) or (B) is used, (1) will certainly produce a *p.d.f.* as density estimate.

An obviously sensible $\hat{f}(\cdot \mid \mathbf{y})$ would be some estimate of the true $f(\cdot \mid \mathbf{y})$. Likewise, candidates for the plug-in procedure (A) are as follows.

(AE) Choose $\hat{\mathbf{z}} = \int \mathbf{z} \hat{f}(\mathbf{z} \mid \mathbf{y}) dz$, an estimate of the conditional *expected* value.

(AS1) Choose for $\hat{\mathbf{z}}$, a *simulated* value from $\hat{f}(\cdot \mid \mathbf{y})$.

(ASr) Use $r^{-1} \Sigma_{j=1}^{r} K\{\cdot \mid (\mathbf{y}, \hat{\mathbf{z}}_j), h\}$ where $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_r$ are independently simulated values from $\hat{f}(\cdot \mid \mathbf{y})$.

Method (ASr) is thus an empirical version of method (B). (Although it is not likely to be as efficient, it holds computational advantages, as pointed out later.)

3. BIVARIATE CASE

Consider first the simplest case in which there are missing data on only one component. Suppose a bivariate observation is denoted by $\mathbf{x}^{\mathrm{T}} = (x_1, x_2)$ and that $D = (D_{12}, D_1)$ where $D_{12} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n_0)})$ comprises n_0 (> 0) complete observations and $D_1 = (y_1^{(n_0+1)}, \ldots, y_1^{(n)})$, where $n = n_0 + n_1$, say. From D_{12} we can immediately produce density estimates $\hat{f}(\mathbf{x} \mid D_{12}, h)$, along with

$$\hat{f}_1(x_1 \mid D_{12}, h) = \int \hat{f}(\mathbf{x} \mid D_{12}, h)\, dx_2$$

and

$$\hat{f}(x_2 \mid x_1, D_{12}, h) = \hat{f}(\mathbf{x} \mid D_{12}, h) / \hat{f}_1(x_1 \mid D_{12}, h) \tag{2}$$

for instance.

We shall restrict ourselves to a special product-type kernel function

$$K(\mathbf{x} \mid \mathbf{y}, h) = h^{-2} \prod_{j=1}^{2} K_1\{(x_j - y_j)/h\}, \tag{3}$$

in which $K_1(\cdot)$ is a univariate symmetric unimodal *p.d.f.* with mode at zero and with bounded second moment. In the numerical work we have carried out it greatly aids the workability of method (B) to use the Normal kernel

$$K_1(u) = (2\pi)^{-\frac{1}{2}} \exp(-\tfrac{1}{2} u^2), \quad -\infty < u < \infty.$$

If (3) is used in (2) we obtain

$$\hat{f}(x_2 \mid x_1, D_{12}, h) = \left(h^{-1} \sum_{i=1}^{n_0} \prod_{j=1}^{2} K_1\{(x_j - x_j^{(i)})/h\} \right) \Big/ \left(\sum_{i=1}^{n_0} K_1\{(x_1 - x_1^{(i)})/h\} \right). \quad (4)$$

In particular, for the jth incomplete observation, $y_1^{(j)}$, with $z^{(j)}$ representing the missing $y_2^{(j)}$,

$$\hat{f}(z^{(j)} \mid y_1^{(j)}, D_{12}, h) = \sum_{i=1}^{n_0} w_{ji} h^{-1} K_1\{(z^{(j)} - x_2^{(i)})/h\}, \quad (5)$$

where

$$w_{ji} \propto K_1\{(y_1^{(j)} - x_1^{(i)})/h\}, \quad \sum_{i=1}^{n_0} w_{ji} = 1, \quad j = n_0 + 1, \ldots, n.$$

In method (AE), therefore, the plug-in value for $z^{(j)}$ is

$$\hat{z}^{(j)} = \sum_{i=1}^{n_0} w_{ji} x_2^{(i)},$$

c.f. Watson (1964).

Since the right-hand side of (5) describes a mixture density, with weights $\{w_{ji}: i = 1, \ldots, n_0\}$, simulation of values for $\hat{\mathbf{z}}^{(j)}$ for method (ASr) is easy if random numbers can be generated according to $K_1(\cdot)$.

If (5) is used for method (B) we obtain, for the new kernel function,

$$h^{-3} K_1\{(x_1 - y_1^{(j)})/h\} \sum_{i=1}^{n_0} w_{ji} \int K_1\{(x_2 - z)/h\} K_1\{(z - x_2^{(i)})/h\}\, dz,$$

which leaves us requiring to evaluate a convolution integral. For the Normal kernel, for instance, the above expression becomes

$$(\sqrt{(2)}h^2)^{-1} K_1\{(x_1 - y_1^{(j)})/h\} \sum_{i=1}^{n_0} w_{ji} K_1\{(x_2 - x_2^{(i)})/\sqrt{(2)}h\}. \quad (6)$$

This can be interpreted as placing a fraction w_{ji} of the jth observation at $(y_1^{(j)}, x_2^{(i)})$, $i = 1, \ldots, n_0$, with an inflation by $\sqrt{2}$ of the smoothing parameter corresponding to the second component.

Method (A), in whichever form, gives, as overall density estimate,

$$\tilde{\hat{f}}(\mathbf{x} \mid D, h) = n^{-1} h^{-2} \left[\sum_{i=1}^{n_0} K(\mathbf{x} \mid \mathbf{x}^{(i)}, h) + r^{-1} \sum_{i=n_0+1}^{n_0+n_1} \sum_{s=1}^{r} K\{\mathbf{x} \mid (y_1^{(i)}, \hat{z}_s^{(i)}), h\} \right], \quad (7)$$

where $\hat{z}_s^{(i)}$, $s = 1, \ldots, r$, are the r plug-in values and $r = 1$ for method (AE).

Method (B) gives

$$\hat{f}(\mathbf{x} \mid D, h) = n^{-1} h^{-2} \left[\sum_{i=1}^{n_0} K(\mathbf{x} \mid \mathbf{x}^{(i)}, h) + \sum_{j=n_0+1}^{n_0+n_1} K_1\{(x_1 - y_1^{(j)})/h\} \times \int K_1\{(x_2 - z)/h\} \hat{f}(z \mid y_1^{(j)}, D_{12}, h)\, dz \right]. \quad (8)$$

From either (7) or (8), new estimates of the marginal and conditional densities can be obtained. The previous results can easily be extended to the data structure

$$D = (D_{12}, D_1, D_2), \tag{9}$$

where D_{12} and D_1 are as in Section 3 and D_2 denotes a set $(y_2^{(n_0+n_1+1)}, \ldots, y_2^{(n)})$ of n_2 independent observations of the second component. Redefine $n = n_0 + n_1 + n_2$. Although, conceivably, we might deal first with (D_{12}, D_1) as in Section 3 and then treat D_2 using (7) or (8) just as D_1 was treated in Section 3 using (2), we state the results for the more symmetric procedure in which D_2 are "completed" using

$$\hat{f}(\cdot \mid x_2) = \hat{f}(\cdot \mid x_2, D_{12}, h).$$

(Asymptotic results such as those given in Section 4 could be obtained for more convoluted procedures but the calculations, though similar, would be even more lengthy.)

Method (A) generates, as an extra term in the square bracket of (7),

$$r^{-1} \sum_{i=n_0+n_1+1}^{n} \sum_{s=1}^{r} K\{\mathbf{x} \mid (\hat{z}_s^{(i)}, y_2^{(i)}), h\}.$$

For method (B), the corresponding extra term in (8) is

$$\sum_{j=n_0+n_1+1}^{n} K_1\{(x_2 - y_2^{(j)})/h\} \int K_1\{(x_1 - z)/h\} \hat{f}(z \mid y_2^{(j)}, D_{12}, h)\, dz, \tag{10}$$

in which each integral is of the form

$$h^{-1} \sum_{i=1}^{n_0} v_{ji} \int K_1\{(x_1 - z)/h\} K_1\{(z - x_1^{(i)})/h\}\, dz,$$

where

$$v_{ji} \propto K_1\{(y_2^{(j)} - x_2^{(i)})/h\}, \quad \sum_{i=1}^{n_0} v_{ji} = 1, \quad \text{for each } j.$$

It is conceivable in some applications that conditional density estimates for use in calculating the kernels for the incomplete observations may be available from independent sources. They may even sometimes be treated as known. Versions of the above methods can be written down and theoretical results obtained on the lines of Section 4. The calculations are easier than the ones we require for the present set-up.

4. ASYMPTOTIC THEORY

In this Section we develop asymptotic theory for the above procedures with data structure given by (9), establishing in particular the convergence in mean integrated squared error of the density estimates to the true densities. The calculations are similar to, if more complicated than, those of Rosenblatt (1956, 1969) and Epanechnikov (1969), among others.

Immediately it can be pointed out that method (AE) will fail on asymptotic grounds. Both (AE) and (ASr) are imputation-based methods. The former can be described as *mean*-imputation and the latter as *random*-imputation. Since, in this example, the data, completed by the imputed values, are then treated as if they were a typical complete data set, the conditionally deterministic mean-imputation method is unsatisfactory. The corresponding trivial univariate version of mean-imputation would impute the same value for all missing values, so that the resulting "completed" data set fails to represent the true random variation.

For the other methods we assume that, given the total size n, $(n_0, n_1, n_2) = \mathbf{n}^T$ is trinomially distributed, with cell probabilities $(\theta_0, \theta_1, \theta_2)$, corresponding to conditions under which the missing data are "missing at random" (Rubin, 1976).

For method (B), under the usual conditions, (see, for instance, Epanechnikov, 1969), it turns out that

$$\mathbb{E}\,\hat{f}(\mathbf{x} \mid D, h) = f(\mathbf{x}) + \tfrac{1}{2} h^2 H(\mathbf{x}) + o(h^2) \tag{11}$$

and

$$\operatorname{var} \hat{f}(\mathbf{x} \mid D, h) = n^{-1} h^{-2}\, \hat{G} f(\mathbf{x}) + o(n^{-1} h^{-2}), \tag{12}$$

where

$$\hat{G} = \theta_0 I_2^2 + 2(\theta_1 + \theta_2) I_2 I_4 + \theta_0^{-1} \{(\theta_1^2 + \theta_2^2) I_2 I_3 + 2\theta_1 \theta_2 I_4^2\}$$

and

$$H(\mathbf{x}) = \{(1 + \theta_1 + \theta_2) \operatorname{tr} \mathbf{D}(\mathbf{x}) - \theta_1\, \gamma_1(\mathbf{x}) - \theta_2 \gamma_2(\mathbf{x})\} I_1 .$$

In these expressions,

$$I_1 = \int u^2 K_1(u)\, du; \quad I_2 = \int K_1^2(u)\, du;$$

$$I_3 = \int \left\{ \int K_1(u)\, K_1(u + \zeta)\, du \right\}^2 d\zeta; \quad I_4 = \int\int K_1(u)\, K_1(v)\, K_1(u + v)\, du\, dv;$$

$$\{\mathbf{D}(\mathbf{x})\}_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}, \quad i, j = 1, 2;$$

$$\gamma_1(\mathbf{x}) = \frac{\partial^2 f_1(x_1)}{\partial x_1^2} \cdot f(x_2 \mid x_1) \quad \text{and} \quad \gamma_2(\mathbf{x}) = \frac{\partial^2 f_2(x_2)}{\partial x_2^2} \cdot f(x_1 \mid x_2),$$

where $f_2(\cdot)$ denotes the marginal *p.d.f.* for x_2.

The " $\mathbb{E}$ " in (11) and "var" in (12) deal with all random variation, due to the data *D and* and the sample sizes $\mathbf{n}$, conditional on n. Although more complicated than in the earlier papers, the proof follows familiar lines based on decomposition of density estimates into

$$\text{true value} + \text{bias} + \text{residual}$$

and subsequent Taylor expansions. An approximation to the estimate of a conditional density was used which is similar to that of Theorem 1 of Rosenblatt (1969). This was required in the consideration of (8) and (10), for instance.

Details of the calculations can be obtained from the authors.

From (11) and (12) we obtain

$$\mathbb{E}\,\{\hat{f}(\mathbf{x} \mid D, h) - f(\mathbf{x})\}^2 = n^{-1} h^{-2}\, \hat{G} f(\mathbf{x}) + \tfrac{1}{4} h^4 H^2(x) + o(h^4 + n^{-1} h^{-2})$$

and

$$\int \mathbb{E}\,\{\hat{f}(\mathbf{x} \mid D, h) - f(\mathbf{x})\}^2 d\mathbf{x} = n^{-1} h^{-2} \hat{G} + \tfrac{1}{4} h^4 H^2 + o(h^4 + n^{-1} h^{-2}), \tag{13}$$

where

$$H^2 = \int H^2(\mathbf{x})\, d\mathbf{x}.$$

Thus, provided both $h \to 0$ and $nh^2 \to \infty$, as $n \to \infty$, mean-integrated squared-error (MISE)

consistency holds for the density estimation procedure. The dominant term in (13) is minimized by

$$\hat{h} = \hat{c}n^{\hat{\alpha}},$$

with

$$\hat{\alpha} = -\tfrac{1}{6} \text{ and } \hat{c} = (2\hat{G}/H^2)^{1/6}.$$

In principle this provides an optimal value for h although, as usually happens in this approach, some modification is necessary in practice because H^2 involves the true $f(\cdot)$; see Scott and Factor (1981), for instance.

The corresponding minimum MISE is

$$\hat{S} \propto (\hat{G}Hn^{-1})^{2/3}.$$

For method (ASr) we obtain, similarly, and with the same constant of proportionality,

$$\tilde{S} \propto (\tilde{G}Hn^{-1})^{2/3},$$

where $\tilde{G} = \hat{G} + r^{-1}(\theta_1 + \theta_2) I_2^2$. The additional term corresponds to a contribution to the variance resulting from the simulation variability.

To indicate the behaviour of the density estimators using D as opposed to only D_{12}, consider the particular example of using the Normal kernel when the true density is the circular Normal,

$$f(\mathbf{x}) = (2\pi)^{-1} \exp(-\tfrac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}).$$

In this case $I_1 = 1$, $I_2 = (2\sqrt{\pi})^{-1}$, $I_3 = (2\sqrt{2\pi})^{-1}$ and $I_4 = (\sqrt{6\pi})^{-1}$. Suppose also $\theta_1 = \theta_2 = \theta$, so that $\theta_0 = 1 - 2\theta$. Then, since $f(\mathbf{x})$ represents an independence model,

$$\theta\gamma_1(\mathbf{x}) + \theta\gamma_2(\mathbf{x}) = \theta \operatorname{tr} D(\mathbf{x})$$

and

$$H(\mathbf{x}) = (1 + \theta) \operatorname{tr} D(\mathbf{x}),$$

giving

$$H^2 = (1 + \theta)^2 \int \{\operatorname{tr} D(\mathbf{x})\}^2 \, d\mathbf{x}$$

$$= (1 + \theta)^2 H_0^2, \quad \text{say.}$$

If only the complete observations are used, then the corresponding minimum MISE is, using the results of Epanechnikov (1969), for instance, approximately

$$S_0 \propto \{G_0 H_0 (n\theta_0)^{-1}\}^{2/3},$$

where

$$G_0 = I_2^2.$$

Thus $(\hat{S}/S_0)^{3/2} = (1 + \theta)(1 - 2\theta)\hat{G}/G_0$.

We have, for the right-hand side,

$$(1 + \theta)(1 - 2\theta)(1 - 2\theta + 4\sqrt{2\theta}/\sqrt{3}) + \theta^2(\sqrt{2} + 4/3)\}.$$

Similarly,

$$(\tilde{S}/S_0)^{3/2} = (1 + \theta)\{(1 - 2\theta)(1 - 2\theta + 4\sqrt{2\theta}/\sqrt{3} + 2\theta/r) + \theta^2(\sqrt{2} + 4/3)\}. \quad (14)$$

For $\theta = 0.1\ (0.1)\ 0.4$ and $r = 1, 2, 5, 10$ and ∞ (method (B)) the values of the right-hand side are given in Table 1.

It appears that a method using five simulations of each missing value performs very well and one with r as low as 2 does quite well, compared with method (B).

There is, however, certainly no evidence of improvement in MISE over the value corresponding to D_{12}. To some extent the picture is influenced by the choice of kernel function, $K_1(\cdot)$. For the uniform kernel,

$$K_1(u) = \tfrac{1}{2} \quad |u| \leqslant 1$$
$$= 0 \quad \text{otherwise},$$

for which $I_1 = I_3 = \frac{1}{3}$, $I_2 = \frac{1}{2}$, $I_4 = \frac{3}{8}$ and the right-hand side of (14) is

$$(1+\theta)\{(1-2\theta)(1+\theta+2\theta/r) + 59\theta^2/24\},$$

the values in Table 1 decrease by factors ranging from 3 per cent in the first column up to 9 per cent in the fourth.

TABLE 1
Values of right-hand side of (14)

r	0.1	0.2	0.3	0.4
1	1.20	1.32	1.35	1.26
2	1.11	1.18	1.19	1.15
5	1.06	1.09	1.10	1.08
10	1.04	1.06	1.07	1.06
∞	1.02	1.03	1.04	1.04

Although this is more encouraging it seems that, as far as estimating the *bivariate* distribution is concerned, performance in terms of MISE is determined by the amount of complete data. Given that we are being non-parametric, this fact seems perfectly reasonable. If a parametric model is assumed, the incomplete data more plausibly contribute extra information about the joint density, via the medium of the parametric form. Even then, however, the relative amount of information provided by the incomplete data is small (Murray, 1979).

What is brought out by the Tables is the usefulness of methods (AS) and (B) in terms of non-parametric imputation for the missing values. One objective of an imputation procedure is to produce a completed data-set that is "representative" of the underlying joint distribution. As mentioned in Section 4, method (AE) fails on these grounds. The "ideal" imputation for a missing value is its conditional distribution given all other data. Method (B) provides an estimate of this. In sample survey practice it is much more common to impute one value for each missing value and method (AS1) represents a "fair" such method. Furthermore, method (ASr) corresponds to the multiple imputation approximation to method (B) favoured by Rubin (1978). Clearly, on the basis of MISE, the methods do not greatly distort the picture, relative to that obtained from the complete data. Also, method (ASr), with quite small r, gets very close to the ideal imputation procedure (B).

This is particularly helpful in practice because of the need, with method (B), of storing the large numbers of weights $\{w_{ji}\}$ and $\{v_{ji}\}$. Ways to deal directly with these difficulties are to modify the procedure of Specht (1970) of replacing a kernel-based density estimate by a polynomial approximation or to reduce the number of weights by neglecting all those which are sufficiently small, as in Sorenson and Alspach (1971); see also Box and Tiao (1968). Another possibility is to use the uniform kernel, or any with compact support, for then many of the weights will be zero.

5. MULTIVARIATE GENERALIZATIONS

Generalization to the multivariate case is in principle straightforward if there is a set of complete data that can be used to derive the incomplete-data kernels as in earlier sections. The calculations parallel those for the bivariate case and here we simply state the results.

Suppose the complete data are k-dimensional, that there n_0 complete observations and that the incomplete data arise in m incompleteness patterns, $\{P_j\}$, with sample sizes $\{n_j\}$ and corres-

ponding multinomial probabilities $\{\theta_j\}$, with $\theta_0 = 1 - \Sigma_{j=1}^{m} \theta_j$. Let $f_{p_j}(\mathbf{x}_{p_j})$ denote the marginal density appropriate to the d_j variables observed in P_j and let $D_{p_j}(\mathbf{x})$ denote the $(d_j \times d_j)$ Hessian matrix for that density, evaluated at $\mathbf{x}$. Then the bias for the analogue of either $\hat{f}$ or $\tilde{f}$, at $\mathbf{x}$, is

$$\tfrac{1}{2} h^2 H(\mathbf{x}) + o(h^2),$$

where

$$H(\mathbf{x}) = \left\{ (1 + \sum_{j=1}^{m} \theta_j) \operatorname{tr} D(\mathbf{x}) - \sum_{j=1}^{m} \theta_j \gamma_j(\mathbf{x}) \right\} I_1,$$

with

$$\gamma_j(\mathbf{x}) = \{ \operatorname{tr} D_{p_j}(\mathbf{x}) \} f(\mathbf{x}) / f_{p_j}(\mathbf{x}_{p_j}).$$

The variances are

$$\operatorname{var} \hat{f}(\mathbf{x} \mid D, h) = n^{-1} h^{-k} \hat{G} f(\mathbf{x}) + o(n^{-1} h^{-k})$$

and

$$\operatorname{var} \tilde{f}(\mathbf{x} \mid D, h) = n^{-1} h^{-k} \tilde{G} f(\mathbf{x}) + o(n^{-1} h^{-k}),$$

where

$$\hat{G} = \theta_0 I_2^k + \theta_0^{-1} \sum_{j=1}^{m} \theta_j^2 I_2^{k-d_j} I_3^{d_j} + 2 \sum_{j=1}^{m} \theta_j I_2^{k-d_j} I_4^{d_j}$$

$$+ \theta_0^{-1} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ i \neq j}}^{m} \theta_i \theta_j I_2^{k-d_i-d_j+d_{ij}} I_4^{d_i+d_j-2d_{ij}} I_5^{d_{ij}},$$

where d_{ij} is the number of variables observed in both P_i and P_j and

$$I_5 = \int\int\int K_1(u) K_1(v) K_1(w) K_1(u+v+w) \, du \, dv \, dw.$$

For the Normal kernel, $I_5 = (2\sqrt{2\pi})^{-1}$.
Correspondingly,

$$\tilde{G} = \hat{G} + r^{-1} I_2^k \sum_{j=1}^{m} \theta_j.$$

Optimal smoothing parameters are of the form $cn^{-1/\alpha}$, where $\alpha = k + 4$ and $c = \hat{c}$ or $\tilde{c}$ as in Section 4 where, for instance, $\hat{c}^{k+4} = k\hat{G}/H^2$.

Again the simulation approach, with small r, will be easier to use than method (B), with little loss of efficiency.

If there is no complete data, practical procedures can be suggested. Suppose, for instance, for $k = 3$, there are data sets on all pairs of variables, D_{12}, D_{13} and D_{23}, in an obvious notation. The missing values on variable 3 in D_{12} can be dealt with using D_{13} and/or D_{23} and similarly for the other cases. Attempts to use the data to best advantage are not easy to formulate, in spite of valiant efforts to this end in as yet unpublished work by Dr J. Hilden. Unfortunately, without complete data there is no guarantee of consistency, simply because knowledge of all marginal distributions does not necessarily imply knowledge of the overall joint distribution.

ACKNOWLEDGEMENTS

G. M. M. was partially supported by a grant from the Science Research Council during the course of this work. Communication with Dr J. Hilden of the Institute of Medical Genetics, University of Copenhagen, was much appreciated.

REFERENCES

Box, G. E. P. and Tiao, G. C. (1968) A Bayesian approach to some outlier problems. *Biometrika*, **55**, 119–129.

Dempster, A. P., Laird, N. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the *EM* algorithm. *J. R. Statist. Soc.* B, **39**, 1–38.

Epanechnikov, V. (1969) Nonparametric estimation of a multidimensional probability density. *Theory of Prob. and Applics.*, **14**, 153–158.

Fryer, M. J. (1977) A review of some non-parametric methods of density estimation. *J. Inst. Maths. Applics.*, **20**, 335–354.

Murray, G. D. (1979) The estimation of multivariate normal density functions using incomplete data. *Biometrika*, **66**, 375–380.

Murray, G. D. and Titterington, D. M. (1978) Estimation problems with data from a mixture. *Appl. Statist.*, **27**, 325–334.

Rosenblatt, M. (1956) Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837.

——— (1969) Conditional probability density and regression estimators. In *Multivariate Analysis II* (P. R. Krishnaiah, ed.), pp. 25–31. New York: Academic Press.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

——— (1978) Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proc. Section on Surv. Res. Meth., Amer. Statist. Assoc.*, 20–34.

Scott, D. W. and Factor, L. E. (1981) Monte Carlo study of three data-based nonparametric probability density estimators. *J. Amer. Statist. Ass.*, **76**, 9–15.

Silverman, B. W. (1978) Choosing the window width when estimating a density. *Biometrika*, **65**, 1–12.

Sorenson, H. W. and Alspach, D. L. (1971) Recursive Bayesian estimation using Gaussian sums. *Automatica*, **7**, 465–479.

Specht, D. F. (1970) Series estimation of a probability density function. *Technometrics*, **13**, 409–424.

Titterington, D. M. (1977) Analysis of incomplete multivariate binary data by the kernel method. *Biometrika*, **64**, 455–460.

Watson, G. S. (1964) Smooth regression analysis. *Sankhyā* A, **26**, 359–372.

Wegman, E. J. (1972) Nonparametric probability density estimation, II. *J. Statist. Comput. Simul.*, **1**, 225–245.