Nonparametric Estimation of Multivariate Density with Direct and Auxiliary Data and Application

Debasis Sengupta *
Applied Statistics Unit
Indian Statistical Institute
Kolkata 700108
INDIA

Subhadip Bandyopadhyay †
CKDIS, SET Labs
Infosys Technologies Ltd.
Hyderabad 500032
INDIA

Arup Bose ‡
Stat-Math Unit
Indian Statistical Institute
Kolkata 700108
INDIA

May 28, 2009

Abstract

We consider the problem of multivariate density estimation, using samples from the distribution of interest as well as auxiliary samples from a related distribution. We assume that the data from the target distribution and the related distribution may occur individually as well as in pairs. Using nonparametric maximum likelihood estimator of the joint distribution, we derive a kernel density estimator of the marginal density. We show theoretically, in a simple special case, that the implied estimator of the marginal density has smaller integrated mean squared error than that of a similar estimator obtained by ignoring dependence of the paired observations. We establish consistency of the marginal density estimator under suitable conditions. We demonstrate small sample superiority of the proposed estimator over the estimator that ignores dependence of the samples, through a simulation study with dependent and non-normal populations. The application of the density estimator in nonparametric classification is also discussed. It is shown that the misclassification probability of the resulting classifier is asymptotically equivalent to that of the Bayes classifier. We also include a data analytic illustration.

1 Introduction

Multivariate measurements from footprints (pugmarks) of big cats are often used for such purposes as tracking and counting of animals (see Sharma et al., 2005). Usually the mark of one of the two hind pugs is used for such study. However, the mark of the other hind pug carries further information that can be potentially harnessed for the purpose of estimating the density of the requisite measurements. While marks of both the hind pugs would be found in some tracks, one or the other could be obliterated for various reasons in other tracks.

This example leads one to look for a method of multivariate density estimation on the basis of samples from the requisite distribution as well as samples from another distribution. The two samples

*email: sdebasis@isical.ac.in
†email: Subhadip_B@infosys.com

‡email: bosearu@gmail.com. Research supported by J.C. Bose Fellowship, Govt. of India.

can occur together or in isolation. When the two samples occur together, they are dependent. It is this dependence that could make the marginal samples from the second distribution useful for estimation of the first distribution.

Often samples from the auxiliary distribution are easier to obtain. Ong et al. (2005) considers the problem of classification of tissues into normal or malignant type, on the basis of concentrations of some elements in the cell. The concentrations in normal and malignant cell from a single subject forms a pair of observations. However, many more observations for normal cells may be available from subjects who do not have malignant cells. Thus, the data at hand would consist of a set of such paired observations as well as some marginal observations from normal cells. A similar problem arises in other medical examples (see, e.g., ALR, 2009).

If the target and the auxiliary distributions are considered together, the data described above would be regarded as samples from a multivariate distribution with two groups of characters, with data from one or the other group being unavailable in some cases. Thus, the problem is that of density estimation with missing data. Many methods for handling missing data are available in the literature (see Little and Rubin, 1987, Schafer, 1997 and references therein). Cheng and Wei (1986) and Cheng (1994) have considered mean estimation, while Cheng and Chu (1996) have considered estimation of the distribution function and quantiles with missing data. The special nature of the incompleteness mentioned above (i.e., observations of one set of characters or its complement being missing) have also being considered. However, the focus of research in this scenario have been mainly on joint density estimation (Titterington and Mill, 1983) or on conditional density estimation (Cheng and Chu, 1996). The problem of marginal density estimation has received less attention, except in the special case when the target and the auxiliary distribution are both univariate (Hazelton, 1999).

For the special type of incomplete data considered here, another problem of interest is classification into one of two possible classes, where the 'training' samples may occur in pairs or separately for the two populations. The samples that are not paired can be regarded as having a missing part. Bandyopadhyay (1978, 1979), Dasgupta and Bandyopadhyay (1977) and Leung and Srivastava (1983) have considered classification from dependent training data with no missing component. Bandyopadhyay and Bandyopadhyay (2003) have considered the classification problem with incomplete data, and proposed a parametric classifier.

In this article, we derive kernel density estimators for multivariate data with two dependent groups of characters, where observations from one group or the other can go Missing Completely at Random (MCAR). We also propose a kernel density estimator based two-sample classifier that would be appropriate for training data having such incompleteness.

Let the vectors $X_i = (X_{i1}, \ldots, X_{ip})^T$ and $Y_i = (Y_{i1}, \ldots, Y_{ip})^T$ denote the ith p-variate observation from two distributions. Depending on the possible occurrence of such observations in pairs, we formally define below four types of data.

(i) General data. Suppose we have n pairs of p-variate observations from the two distributions, n_X observations from only the first distribution and n_Y observations from only the second distribution. This leads, without loss of generality, to the following type of data:

$$D_{n,n_X,n_Y} = \left\{ \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}, \begin{pmatrix} X_{n+1}, \ldots, X_{n+n_X} \\ \longleftarrow \text{ not observed } \longrightarrow \end{pmatrix}, \begin{pmatrix} \longleftarrow \text{ not observed } \longrightarrow \\ Y_{n+1}, \ldots, Y_{n+n_Y} \end{pmatrix} \right\}. \quad (1.1)$$

(ii) Completely matched data. A special case of (1.1) is a collection of paired observations on n

sample units:

$$D_{n,0,0} = \left\{ \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \right\}. \tag{1.2}$$

(iii) Completely unmatched data. Another special case of (1.1) consists of observations on two independent sets of sample units from the two distributions:

$$D_{0,n_X,n_Y} = \left\{ \begin{pmatrix} X_1, \ldots, X_{n_X} \\ \leftarrow \text{ not observed} \rightarrow \end{pmatrix}, \begin{pmatrix} \leftarrow \text{ not observed} \rightarrow \\ Y_1, \ldots, Y_{n_Y} \end{pmatrix} \right\}. \tag{1.3}$$

(iv) Single-sample incomplete data. Yet another special case of (1.1) occurs when unmatched observations occur only from one distribution. Thus, the sample is of the form:

$$D_{n,n_X,0} = \left\{ \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}, \begin{pmatrix} X_{n+1}, \ldots, X_{n+n_X} \\ \longleftarrow \text{ not observed} \longrightarrow \end{pmatrix} \right\}. \tag{1.4}$$

or

$$D_{n,0,n_Y} = \left\{ \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}, \begin{pmatrix} \longleftarrow \text{ not observed} \longrightarrow \\ Y_{n+1}, \ldots, Y_{n+n_Y} \end{pmatrix} \right\}. \tag{1.5}$$

Note that the data types $D_{n,0,0}$ and D_{0,n_X,n_Y} are fairly standard, and nonparametric estimation and classification problems for these types of data have been studied extensively.

In Section 2, we propose a nonparametric maximum likelihood estimator (NPMLE) of the joint distribution function of a pair of samples from the two distributions, based on data types D_{n,n_X,n_Y} and $D_{n,n_X,0}$. We use the NPLMLE to derive a multivariate kernel density estimator and estimates of the the corresponding marginal densities for the two distributions. Motivated by the calculation of the mean integrated square error, in the special case of normal parent distribution and standard normal density kernel, we propose a modified estimator of the marginal densities for data type D_{n,n_X,n_Y}. We present large sample results on the behaviour of the NPMLE and the proposed density estimators in Section 3. In Section 4, we propose a likelihood based classification rule using these kernel density estimators and study some large sample properties of the classifier. In Section 5, we report the results of a simulation study. An illustration of the proposed method with a data set on tiger pugmarks is presented in Section 6. We conclude the paper with a discussion on the findings and scope for future work, in Section 7. The proofs of all the results are given in the Appendix.

2 Estimation of distribution function and density

Let $F_{X,Y}$ be the $2p$-variate distribution function of a pair of samples from the two populations. Let $f_{X,Y}$ be the density of this distribution, and f_X and f_Y be the p-variate marginal densities corresponding to populations 1 and 2, respectively. As a precursor to the joint and hence marginal density estimators, we need a nonparametric estimator of $F_{X,Y}$.

2.1 NPMLE of distribution function

When the data are completely matched, the NPMLE of $F_{X,Y}$ is the empirical distribution function (EDF) of the paired observations. When the data are completely unmatched, the NPMLE of $F_{X,Y}$ is the product of the EDF's of the observations from the two populations. The NPMLE has to be worked out in the other two cases.

The likelihood in the case of data type D_{n,n_X,n_Y} may be written as

$$L = \prod_{i=1}^{n} P(X = X_i, Y = Y_i) \prod_{i=n+1}^{n+n_X} P(X = X_i) \prod_{i=n+1}^{n+n_Y} P(Y = Y_i). \tag{2.1}$$

The following result gives the optimum pattern of distribution of the total mass that maximizes the likelihood (2.1). The proof is given in the Appendix.

Theorem 1 *For data type D_{n,n_X,n_Y} with $n_X, n_Y > 0$, let $X_i \neq X_j$ for $1 \leq i \leq n$, $n+1 \leq j \leq n+n_X$ and $Y_i \neq Y_j$ for $1 \leq i \leq n$, $n+1 \leq j \leq n+n_Y$. Then, the likelihood (2.1) is maximized by the distribution having probability mass $m_{i,i}$ at the point (X_i, Y_i), for $1 \leq i \leq n$, and mass $m_{i,j}$ at the point (X_i, Y_j) for $n+1 \leq i \leq n+n_X$, $n+1 \leq j \leq n+n_Y$, where*

$$m_{i,i} = \frac{1}{n+n_X+n_Y}, \quad 1 \leq i \leq n, \tag{2.2}$$

$$m_{i,j} = \frac{n_X+n_Y}{n_X n_Y (n+n_X+n_Y)} + a_{i,j}, \quad n+1 \leq i \leq n+n_X,\ n+1 \leq j \leq n+n_Y, \tag{2.3}$$

and the $a_{i,j}$'s are arbitrary numbers such that $m_{i,j} \geq 0$, $\sum_{i=n+1}^{n+n_X} a_{i,j} = 0$ for $n+1 \leq j \leq n+n_Y$ and $\sum_{j=n+1}^{n+n_Y} a_{i,j} = 0$ for $n+1 \leq i \leq n+n_X$.

Although the NPMLE is not unique, we will use the specific and simple choice $a_{i,j} = 0$ in the sequel. Note that no other choice of $a_{i,j}$ is possible when $n_X = 1$ or $n_Y = 1$. Thus, we have

Corollary 1 *For the situation described in the above theorem, the NPMLE of the distribution function $F_{X,Y}$ has masses only on the points $\{(X_i, Y_i),\ 1 \leq i \leq n\}$ and $\{(X_i, Y_j),\ n+1 \leq i \leq n+n_X,\ n+1 \leq j \leq n+n_Y\}$. A choice of the NPMLE of $F_{X,Y}$ is*

$$\begin{aligned} \hat{F}_{X,Y}(x,y) &= \frac{1}{(n+n_X+n_Y)} \sum_{i=1}^{n} I(X_i \leq x, Y_i \leq y) \\ &\quad + \frac{n_X+n_Y}{n_X n_Y (n+n_X+n_Y)} \sum_{i=n+1}^{n+n_X} \sum_{j=n+1}^{n+n_Y} I(X_i \leq x, Y_j \leq y). \end{aligned} \tag{2.4}$$

The choice is unique if $n_X = 1$ or $n_Y = 1$.

The vector inequalities appearing in (2.4) and in the sequel should be interpreted as a set of simultaneous inequalities between corresponding components of the vectors.

In the case of single-sample incomplete data, the condition of Theorem 1 and Corollary 1 do not hold. In this case, we have

Theorem 2 *For the data type $D_{n,n_X,0}$, let $X_i \neq X_j$ for $1 \leq i \leq n$, $n+1 \leq j \leq n+n_X$. Then the NPMLE of F_{XY} is given by*

$$\hat{F}_{X,Y}(x,y) = \frac{1}{n+n_X} \sum_{i=1}^{n} I(X_i \leq x, Y_i \leq y) + \frac{1}{n+n_X} \sum_{i=n+1}^{n+n_X} I(X_i \leq x) G_i(y) \tag{2.5}$$

where G_i, $n+1 \leq i \leq n+n_X$ are arbitrary distribution functions.

Remark 1 Note that G_i may be interpreted as the conditional distribution of Y given $X = X_i$. Further, the NPMLE is unique only up to the distributions G_i, $n+1 \leq i \leq n+n_X$. A similar non-uniqueness of NPMLE of distribution function was observed in Turnbull (1976) in a truncated and censored data scenario.

2.2 Nonparametric estimation of density function

Let $\{X_i = (X_{i1}, \ldots, X_{ip})^T,\ \ 1 \le i \le n\}$ be a sample of size n from a p-variate distribution F having density f. A kernel density estimator of f is given by (Scott, 1992)

$$\hat{f}(x) = \frac{1}{nh^p} \sum_{i=1}^{n} \prod_{j=1}^{p} K\left(\frac{x_j - X_{ij}}{h}\right), \quad x = (x_1, \ldots, x_p) \in R^p \tag{2.6}$$

where K is a nonnegative kernel function, satisfying $\int K(u)du = 1$. The estimator can also be written as

$$\hat{f}(x) = \int \prod_{j=1}^{p} \left\{\frac{1}{h} K\left(\frac{x_j - u_j}{h}\right)\right\} d\hat{F}(u_1, \ldots, u_p), \tag{2.7}$$

where $\hat{F}$ is the empirical distribution function (EDF) of $X_1, \ldots, X_n$. The expression given above is the density of $X + h\,Y$, where X is a sample from $\hat{F}$, and Y (independent of X) is a sample from the p-variate product distribution with every one-dimensional marginal density given by K. Using this representation, we will derive kernel density estimators for various data types, while replacing the EDF $\hat{F}$ by the appropriate NPMLE.

(i) For general data (type D_{n,n_X,n_Y}) with $n_X, n_Y > 0$, the NPMLE of the $2p$-variate distribution of a paired sample is given by Corollary 1. The corresponding kernel density estimator evaluated at $x = (x_1, \ldots, x_p)^T \in R^p$ and $y = (y_1, \ldots, y_p)^T \in R^p$ is given by

$$\begin{aligned}
\hat{f}_{X,Y}(x,y) &= \frac{1}{h^{2p}} \int \prod_{j=1}^{p} \left\{K\left(\frac{x_j - s_j}{h}\right) K\left(\frac{y_j - t_j}{h}\right)\right\} d\hat{F}_{X,Y}(s,t) \\
&\qquad \left[\ s = (s_1, \ldots, s_p)^T \in R^p,\ \ t = (t_1, \ldots, t_p)^T \in R^p\ \right] \\
&= \frac{1}{(n + n_X + n_Y)h^{2p}} \sum_{i=1}^{n} \prod_{j=1}^{p} \left\{K\left(\frac{x_j - X_{i,j}}{h}\right) K\left(\frac{y_j - Y_{i,j}}{h}\right)\right\} \\
&\quad + \frac{n_X + n_Y}{n_X n_Y (n + n_X + n_Y)h^{2p}} \\
&\qquad \times \sum_{i=n+1}^{n+n_X} \sum_{l=n+1}^{n+n_Y} \prod_{j=1}^{p} \left\{K\left(\frac{x_j - X_{i,j}}{h}\right) K\left(\frac{y_j - Y_{l,j}}{h}\right)\right\}.
\end{aligned} \tag{2.8}$$

The marginal densities obtained from (2.8) are

$$\begin{aligned}
\hat{f}_X(x) &= \frac{1}{(n + n_X + n_Y)h^p} \sum_{i=1}^{n} \prod_{j=1}^{p} K\left(\frac{x_j - X_{i,j}}{h}\right) \\
&\quad + \frac{n_X + n_Y}{n_X (n + n_X + n_Y)h^p} \sum_{i=n+1}^{n+n_X} \prod_{j=1}^{p} K\left(\frac{x_j - X_{i,j}}{h}\right),
\end{aligned} \tag{2.9}$$

$$\begin{aligned}
\hat{f}_Y(y) &= \frac{1}{(n + n_X + n_Y)h^p} \sum_{i=1}^{n} \prod_{j=1}^{p} K\left(\frac{y_j - Y_{i,j}}{h}\right) \\
&\quad + \frac{n_X + n_Y}{n_Y (n + n_X + n_Y)h^p} \sum_{i=n+1}^{n+n_Y} \prod_{j=1}^{p} K\left(\frac{y_j - Y_{i,j}}{h}\right).
\end{aligned} \tag{2.10}$$

Note that although the NPMLE $\hat{F}_{X,Y}$ given in Corollary 1 is not unique, all possible choices of the NPMLE indicated by Theorem 1 have the same pair of (estimated) marginal distributions corresponding to F_X and F_Y. The density estimators $\hat{f}_X$ and $\hat{f}_Y$ defined above are kernel densities corresponding to these unique marginal distributions.

(ii) For completely matched data ($n_X = n_Y = 0$), the estimators (2.8) and (2.10) reduce to the following well-known forms:

$$\hat{f}_{X,Y}(x,y) = \frac{1}{nh^{2p}} \sum_{i=1}^{n} \prod_{j=1}^{p} \left\{ K\left(\frac{x_j - X_{i,j}}{h}\right) K\left(\frac{y_j - Y_{i,j}}{h}\right) \right\}, \tag{2.11}$$

$$\hat{f}_X(x) = \frac{1}{nh^{p}} \sum_{i=1}^{n} \prod_{j=1}^{p} K\left(\frac{x_j - X_{i,j}}{h}\right), \tag{2.12}$$

$$\hat{f}_Y(y) = \frac{1}{nh^{p}} \sum_{i=1}^{n} \prod_{j=1}^{p} K\left(\frac{y_j - Y_{i,j}}{h}\right). \tag{2.13}$$

(iii) For completely unmatched data ($n = 0$), the estimators (2.8) and (2.10) simplify to the following well-known forms:

$$\hat{f}_X(x) = \frac{1}{(n+n_X)h^{p}} \sum_{i=1}^{n+n_X} \prod_{j=1}^{p} K\left(\frac{x_j - X_{i,j}}{h}\right), \tag{2.14}$$

$$\hat{f}_Y(y) = \frac{1}{(n+n_Y)h^{p}} \sum_{i=1}^{n+n_Y} \prod_{j=1}^{p} K\left(\frac{y_j - Y_{i,j}}{h}\right), \tag{2.15}$$

$$\hat{f}_{X,Y}(x,y) = \hat{f}_X(x)\hat{f}_Y(y). \tag{2.16}$$

(iv) For single-sample incomplete data (type $D_{n,n_X,0}$), the NPMLE of $F_{X,Y}$ is given by Theorem 2. The corresponding kernel density estimator at $x = (x_1, \ldots, x_p)^T \in R^p$ and $y = (y_1, \ldots, y_p)^T \in R^p$ is

$$\begin{aligned} \tilde{f}_{X,Y}(x,y) = {} & \frac{1}{(n+n_X)h^{2p}} \sum_{i=1}^{n} \prod_{j=1}^{p} \left\{ K\left(\frac{x_j - X_{i,j}}{h}\right) K\left(\frac{y_j - Y_{i,j}}{h}\right) \right\} \\ & + \frac{1}{(n+n_X)h^{2p}} \sum_{i=n+1}^{n+n_X} \left\{ \prod_{j=1}^{p} K\left(\frac{x_j - X_{i,j}}{h}\right) \right\} \int \prod_{j=1}^{p} K\left(\frac{y_j - z_j}{h}\right) dG_i(z), \\ & \left[z = (z_1, \ldots, z_p)^T \in R^p \right]. \end{aligned} \tag{2.17}$$

This expression can be written in terms of the conditional distribution of Y given $X = X_i$ (see Remark 1), as follows:

$$\begin{aligned} \tilde{f}_{X,Y}(x,y) = {} & \frac{1}{(n+n_X)h^{2p}} \sum_{i=1}^{n} \prod_{j=1}^{p} \left\{ K\left(\frac{x_j - X_{i,j}}{h}\right) K\left(\frac{y_j - Y_{i,j}}{h}\right) \right\} \\ & + \frac{1}{(n+n_X)h^{2p}} \sum_{i=n+1}^{n+n_X} \left\{ \prod_{j=1}^{p} K\left(\frac{x_j - X_{i,j}}{h}\right) \right\} \\ & \times \int f_{Y|X=X_i}(z) \left\{ \prod_{j=1}^{p} K\left(\frac{y_j - z_j}{h}\right) \right\} dz_1 \cdots dz_p. \end{aligned} \tag{2.18}$$

Marginal density estimator of f_X obtained from (2.18) coincides with (2.14). The estimator of f_Y is

$$\begin{aligned}\tilde{f}_Y(y) &= \frac{1}{(n+n_X)h^p}\sum_{i=1}^{n}\prod_{j=1}^{p}K\left(\frac{y_j-Y_{i,j}}{h}\right)\\ &\quad+\frac{1}{(n+n_X)h^p}\sum_{i=n+1}^{n+n_X}\int f_{Y|X=X_i}(z)\left\{\prod_{j=1}^{p}K\left(\frac{y_j-z_j}{h}\right)\right\}dz_1\cdots dz_p. \quad (2.19)\end{aligned}$$

The conditional densities appearing in the expressions (2.18) and (2.19) may be replaced by suitable estimators, possibly obtained from the matched part of the data. Estimators for data type $D_{n,0,n_Y}$ are similar to (2.18) and (2.19) and may be obtained by interchanging the samples.

2.3 Utilization of data dependence in marginal density estimation

In order to demonstrate theoretically how dependence of the X and Y samples can lead to improved estimation of marginal density, we present a small sample comparison of performance between the estimator $\tilde{f}_Y$ given by (2.19) and the estimator $\hat{f}_Y$ which is obtained by using only samples from the second population under assumption of normal parent population and standard normal kernel function.

Theorem 3 *Let $f_{X,Y}$ be a $2p$-variate normal density, and f_X and f_Y be its p-variate marginal densities. Suppose $(X_1,Y_1),\ldots,(X_n,Y_n)$ are samples from $f_{X,Y}$, $X_{n+1},\ldots,X_{n+n_X}$ are samples from f_X, and the standard normal density function is used as kernel. Then*

(i) the kernel density estimator $\tilde{f}_Y$ given by (2.19) has smaller integrated mean squared error (IMSE) in comparison with the estimator $\hat{f}_Y$ given by (2.13).

(ii) the kernel density estimator of f_Y given by (2.10) has larger integrated mean squared error (IMSE) in comparison with the estimator given by (2.15).

The first part of the above theorem indicates that a good choice of the conditional density estimator would make $\tilde{f}_Y$ a better estimator than a kernel density estimator based on marginal data alone. However, the estimator given in (2.10) is generally not capable of harnessing information from dependence in the data. The second part of Theorem 3 shows that it may even have larger IMSE than the estimator (2.15) which is based on the marginal data alone.

Since, under the assumption of normal distribution, the estimator (2.10) fares worse than the estimator based on marginal data, an improvement is needed for data type D_{n,n_X,n_Y}. Motivated by the improvement shown by the estimator (2.19) for data type $D_{n,n_X,0}$, we propose the following modified estimators of the marginal densities, for data type D_{n,n_X,n_Y}.

$$\begin{aligned}\tilde{f}_X(x) &= \frac{1}{(n+n_X+n_Y)h^p}\sum_{i=1}^{n+n_X}\prod_{j=1}^{p}K\left(\frac{x_j-X_{i,j}}{h}\right)\\ &\quad+\frac{1}{(n+n_X+n_Y)h^p}\sum_{i=n+1}^{n+n_Y}\int f_{X|Y=Y_i}(z)\left\{\prod_{j=1}^{p}K\left(\frac{x_j-z_j}{h}\right)\right\}dz_1\cdots dz_p, \quad (2.20)\end{aligned}$$

$$\tilde{f}_Y(y) = \frac{1}{(n+n_X+n_Y)h^p}\sum_{i=1}^{n+n_Y}\prod_{j=1}^{p}K\left(\frac{y_j-Y_{i,j}}{h}\right) + \frac{1}{(n+n_X+n_Y)h^p}\sum_{i=n+1}^{n+n_X}\int f_{Y|X=X_i}(z)\left\{\prod_{j=1}^{p}K\left(\frac{y_j-z_j}{h}\right)\right\}dz_1\cdots dz_p. \quad (2.21)$$

Once again, the conditional densities appearing in the expressions (2.20) and (2.21) may be replaced by suitable estimators, possibly obtained from the matched part of the data. The obvious choice might be ratio of estimated joint and marginal densities by kernel method. See Cheng and Chu (1996) for some other choices of the conditional density estimators.

The following theorem, which may be proved along the lines of Theorem 3, indicates that the estimators (2.20) and (2.21) may perform better than (2.14) and (2.15) which are based on marginal data alone.

Theorem 4 *Let $f_{X,Y}$ be a $2p$-variate normal density, and f_X and f_Y be its p-variate marginal densities. If $(X_1,Y_1),\ldots,(X_n,Y_n)$ are samples from $f_{X,Y}$, $X_{n+1},\ldots,X_{n+n_X}$ are samples from f_X, $Y_{n+1},\ldots,Y_{n+n_Y}$ are samples from f_Y, and the standard normal density function is used as kernel, then the kernel density estimator of f_Y given by (2.21) has smaller integrated mean squared error (IMSE) in comparison with the estimator given by (2.15).*

3 Large sample results

In this section we study the large sample properties of our estimators and procedures. All proofs are given in the Appendix.

Theorem 5 *Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be samples from $F_{X,Y}$ and $X_{n+1},\ldots,X_{n+n_X}$ be samples from F_X. Suppose $\lim_{n\to\infty,n_X\to\infty} n/(n+n_X)\to k\in(0,1]$.*

(i) Then, $\hat{F}_{X,Y}$ given by (2.5), where G_i is the true conditional distribution function $F_{Y|X=x}$ evaluated at $x=X_i$, converges point-wise to the population cdf $F_{X,Y}$ almost surely.

(ii) If $\hat{F}_{Y|X=x}$ is an estimator of $F_{Y|X=x}$ which satisfies the condition

$$P\left[\lim_{\substack{n\to\infty,\ n_X\to\infty\\ n/(n+n_X)\to k}}\sup_x|\hat{F}_{Y|X=x}(y)-F_{Y|X=x}(y)|=0\right]=1$$

for each y, then the estimator $\hat{F}_{X,Y}$ defined by

$$\hat{F}_{X,Y}(x,y)=\frac{1}{n+n_X}\sum_{i=1}^{n}I(X_i\le x,Y_i\le y)+\frac{1}{n+n_X}\sum_{i=n+1}^{n+n_X}I(X_i\le x)\hat{F}_{Y|X=X_i}(y).$$

converges point-wise to $F_{X,Y}(x,y)$ almost surely.

We now turn to convergence of the marginal density estimators for data type D_{n,n_X,n_Y}. Note that, for the special cases $n=0$ and $n_X=n_Y=0$, i.e., for data types D_{0,n_X,n_Y} and $D_{n,0,0}$, convergence

of the kernel density estimators of the marginal densities follow from standard results (see for example, Silverman, 1986, Chapter 3), under suitable conditions on f_X, f_Y and K. For data type $D_{n,n_X,0}$ also, convergence of the kernel density estimator of f_X follows from standard results. Thus, we only have to consider convergence of (2.19) for data type $D_{n,n_X,0}$ with $n_X > 0$, and of (2.21) for data type D_{n,n_X,n_Y} with $n_X, n_Y > 0$. Convergence of the estimators of f_X for data types $D_{n,0,n_Y}$ and D_{n,n_X,n_Y} need not be considered separately, because of the symmetric nature of the problems.

Theorem 6 *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be samples from $F_{X,Y}$ and $X_{n+1}, \ldots, X_{n+n_X}$ be samples from F_X. Suppose*

(a) $f_{Y|X=x}$ has continuous derivatives up to order four,

(b) $\frac{1}{h^p} \int f_{Y|X=x}(z) \left\{ \prod_{j=1}^{p} K\left(\frac{y_j - z_j}{h}\right) \right\} dz_1 \cdots dz_p$ is uniformly (in x) equicontinuous (in h).

Then,

(i) the estimator $\hat{f}_Y$ given in (2.19), with $f_{Y|X=x}$ representing the true conditional density, converges point-wise to f_Y in probability as $h \to 0$, $n_X \to \infty$, $n \to \infty$, $n/(n+n_X) \to k \in (0,1]$ and $h(n+n_X) \to \infty$.

(ii) the estimator $\hat{f}_Y$ given in (2.21), with $f_{Y|X=x}$ representing the true conditional density, converges point-wise to f_Y in probability as $h \to 0$, $n + n_X + n_Y \to \infty$, $n_Y \to \infty$, $(n+n_Y)/(n+n_X+n_Y) \to k \in (0,1]$ and $h(n+n_X+n_Y) \to \infty$.

4 Application in classification: Kernel classifier

The standard likelihood-based approach of classifying a new observation is to compare the ratio of the joint likelihoods under the alternative scenarios. Thus, a new p-variate observation Z is classified into population 1 if

$$L = \frac{\left[\left\{ \prod_{i=1}^{n} \tilde{f}_{X,Y}(X_i, Y_i) \right\} \left\{ \prod_{i=n+1}^{n+n_X} \tilde{f}_X(X_i) \right\} \left\{ \prod_{i=n+1}^{n+n_Y} \tilde{f}_Y(Y_i) \right\} \tilde{f}_X(Z) \right]}{\left[\left\{ \prod_{i=1}^{n} \tilde{f}_{X,Y}(X_i, Y_i) \right\} \left\{ \prod_{i=n+1}^{n+n_X} \tilde{f}_X(X_i) \right\} \left\{ \prod_{i=n+1}^{n+n_Y} \tilde{f}_Y(Y_i) \right\} \tilde{f}_Y(Z) \right]} > c, \tag{4.1}$$

where c is a cutoff value depending on the prior probabilities and cost of misclassification. The estimators in the numerator and denominator are obtained by treating the new observation alternatively as having come from the two populations. One major problem while working with (4.1) is computational complexity. A simpler approach is to estimate densities using training data only. That is, to use

$$L = \frac{\tilde{f}_X(Z)}{\hat{f}_Y(Z)} > c. \tag{4.2}$$

If $n_X, n_Y > 0$, then the marginal density estimates in the numerator and denominator are obtained from (2.20) and (2.21) respectively. If $n_X > n_Y = 0$, then the estimators in the numerator and

denominator have to be replaced by (2.14) and (2.19), respectively. A similar adjustment is needed when $n_Y > n_X = 0$. If $n_X = n_Y = 0$ or $n = 0$, the estimators (2.14) and (2.15) are to be used in the numerator and denominator, respectively.

The proposed classifier differs from a regular kernel classifier only when $n_X > 0$ and/or $n_Y > 0$. Theorems 3 and 4 suggest that the kernel density estimators, which utilize pairing of a part of the data, may be superior as density estimator to those that do not utilize it. However, these results do not directly imply superiority of a classifier based on the 'better' estimators. Some evidence of improved classification is given through simulation results reported in Section 5.

Finally, we deal with average misclassification probability for rule (4.2).

Theorem 7 *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be samples from $F_{X,Y}$ and $X_{n+1}, \ldots, X_{n+n_X}$ be samples from F_X, and the conditions of Theorem 6 hold.*

(i) Then, the average misclassification probability of the classification rule (4.2), with $\tilde{f}_X$ and $\tilde{f}_Y$ given by (2.14) and (2.19), respectively, converges to the average Bayes misclassification probability as $h \to 0$, $n_X \to \infty$, $n \to \infty$, $n/(n + n_X) \to k \in (0, 1]$ and $h(n + n_X) \to \infty$.

(ii) Further, suppose that $Y_{n+1}, \ldots, Y_{n+n_Y}$ is a sample from F_Y and conditions (i) and (ii) of Theorem 6 also holds for the conditional density of X given Y. Then the average misclassification probability of the classification rule (4.2), with $\tilde{f}_X$ and $\tilde{f}_Y$ given by (2.20) and (2.21), respectively, converges to the average Bayes misclassification probability as $h \to 0$, $n_X/(n + n_X + n_Y) \to k_x \in (0, 1]$, $n_Y/(n + n_X + n_Y) \to k_y \in (0, 1]$ and $h(n + n_X + n_Y) \to \infty$.

5 Simulation study

A mixture of two four-variate normal distributions was used for a Monte carlo simulation study of the performance of the proposed estimator. The mean vectors of the two normal distributions are $(1\ 1\ -1\ -1)^T$ and $(1\ -1\ -1\ 1)^T$, respectively. All the correlations are zero, all the characters have variance σ^2 and the mixing proportion is even. Three values of σ^2 were used: 0.5, 1 and 2. The last two of the four characters were treated as the vector of interest (Y) and the other two, as the vector of auxiliary variables (X). We generated one-sample incomplete data of the type $D_{n,n_X,0}$ for two sets of sample sizes: $n = 15$, $n_X = 35$ and $n = 15$, $n_X = 85$. The estimator $\tilde{f}_Y(\cdot)$ given by (2.19) was computed using the estimator of the conditional density estimate obtained from the paired part of the data, as the ratio of a kernel density estimator of the joint distribution of X and Y and that of the marginal distribution of X. The bandwidths were selected by minimizing the criterion

$$\int_y \tilde{f}^2(y)dy - \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{-i}(Y_i),$$

where $\tilde{f}_{-i}(\cdot)$ is the density estimate obtained by dropping the ith observation. This is an empirical version of a criterion that is equivalent to the IMSE (Silverman, 1986, chapter 3). For comparison, the estimator $\hat{f}_Y(\cdot)$ given by (2.13) was also computed, using the above criterion for bandwidth selection. An 'oracle' estimator, which is similar to (2.13) but is based on $n+n_X$ samples (including the missing data), was also computed.

Table 1 gives the actual IMSE of the three estimators, based on the average of 100 simulation runs. The relative IMSE (given as proportion of $\int f_Y^2(y)dy$) is also given.

Table 1. IMSE of estimators of $f_Y(\cdot)$

		IMSE (relative IMSE) of		
σ^2	n	Marginal data estimator $\hat{f}_Y(\cdot)$	Proposed estimator $\tilde{f}_Y(\cdot)$	Oracle estimator
0.5	15	0.00899 (0.1408)	0.00594 (0.0930)	0.00495 (0.0775)
1.0	15	0.00918 (0.1687)	0.00791 (0.1453)	0.00510 (0.0937)
2.0	15	0.00926 (0.2049)	0.00919 (0.2034)	0.00530 (0.1173)
0.5	35	0.00467 (0.0730)	0.00424 (0.0664)	0.00402 (0.0629)
1.0	35	0.00477 (0.0877)	0.00449 (0.0825)	0.00424 (0.0780)
2.0	35	0.00482 (0.1066)	0.00474 (0.1050)	0.00431 (0.0954)

It is observed that the proposed estimator improves upon the kernel density estimator that does not make use of data on the auxiliary variable (X). The improvement is more when σ^2 is small, i.e., when the modes of the mixed normal distribution are better separated. It should be noted that when the modes are well separated, X carries more information about Y, and it is this information which allows $\tilde{f}_Y(\cdot)$ to be better than $\hat{f}_Y(\cdot)$. When σ^2 is larger, this advantage is reduced.

In another set of simulations, the performance of the proposed kernel classifier was examined. The conditions for this study were as above. The test data set consisted of 250 samples that were drawn from an even mixture of the two populations. The bandwidth was chosen by minimizing the smoothed cross-validation misclassification rate. The empirical misclassification rates based on 100 simulations are reported in Table 2. The different columns of this table represent the following classifiers:
A: Kernel classifier based on paired part of data only (incomplete part of the data ignored);
B: Kernel classifier based on marginal data (ignores dependence of paired data);
C: Proposed Kernel classifier with estimated conditional density;
D: 'Oracle' Kernel classifier that makes use of the 'missing' part of the data.

Table 2. Misclassification errors of different classifiers

		Misclassification rate of classifier			
σ^2	n	A	B	C	D
0.5	15	0.1202	0.0770	0.0621	0.0611
1	15	0.1794	0.1719	0.1703	0.1648
2	15	0.3375	0.3281	0.3293	0.3174
0.5	35	0.1140	0.0762	0.0583	0.0601
1	35	0.1868	0.1741	0.1714	0.1633
2	35	0.3476	0.3327	0.3331	0.3148

It is observed that the proposed classifier generally performs better than the classifiers based on paired data alone or marginal data alone, and only marginally worse than the 'oracle' classifier. As in the case of the estimation, the improvement is more when σ^2 is small.

6 Analysis of tiger pugmark data

As an illustration of the proposed methods, we analyse a data set on left and right hind pug-marks of tigers. In India, tiger population size estimation is carried out on the basis of left hind pug-mark (footprint) measurements. In order to minimize risk, plaster-casts of pug-marks are made in the field, and measurements are taken subsequently. It is generally easier to distinguish between left and right hind pug-marks when the mark is seen in the context of a trail, but mistakes are occasionally made. A classifier is needed for cross-checking.

We considered pairs of left and right hind pug-mark measurements taken from several trails in a field study conducted by the Indian Statistical Institute during 2004-05. The trails were sufficiently

far apart, and it can be safely assumed that each trail corresponds to a different animal. However, the measurements form the left and right pug-marks from a particular trail correspond to the same animal. The training data consisted of seventeen 8-variate observation on pairs of left and right pug-marks, and another fifteen measurements of left pug-marks only. Twenty-seven additional pug-mark measurements were available for validation.

The complete data kernel classifier (classifier A of simulation) misclassified 3 out of 27 observations. The marginal kernel classifier (classifier B of simulation) also misclassified 3 out of 27 observations. The kernel classifier proposed in this paper (classifier C of simulation) misclassified 2 out of 27 observations.

7 Discussion

The proposed kernel density estimator utilizes the dependence between the samples from the distribution of interest and the auxiliary distribution, as evident from the paired samples. Therefore, the estimator can perform better than a kernel density estimator that does not utilize this information, only if the dependence is strong. Further, this utilization takes place through an estimator of the conditional density of the distribution of interest given a sample from the auxiliary distribution. Hence, the performance of the proposed estimator depends on the quality of the conditional density estimator, which is based on the paired samples. If the paired sample size is small, the conditional density estimator would be poor, and the proposed estimator would not be better than a kernel density estimator that does not utilize the samples from the auxiliary distribution. Therefore, in order that the proposed estimator is useful, there ought to be strong dependence between the distribution of interest and the auxiliary distribution, and one should have a reasonable number of paired samples. The simulation results of Section 5 confirm this fact.

Once these conditions are satisfied, the amount of improvement would depend on the size of the sample from the auxiliary distribution. In some applications, samples from the auxiliary distribution come cheaper or more easily, in comparison with samples from the distribution of interest.

The motivating examples of this paper can lead to further types of dependent data. For instance, in the problem of estimation of pugmark feature distribution, there can multiple measurements of left and right pugmarks taken from a particular trail. In the tissue classification example too, there can be multiple samples of malignant and normal cells of the same subject. These forms of data give rise to the problem of improved estimation through exploitation of more complicated structures of dependence. The work presented in this paper can be a stepping stone for solving that problem.

Appendix

Proof of Theorem 1. We first consider the scalar case ($p = 1$) for ease of visualization. We partition the (X, Y) plane as follows.

(a) The set of points corresponding to paired data.

(b) The set of points of intersections of the lines $X = X_i$, $n + 1 \leq i \leq n + n_X$, with the lines $Y = Y_j$, $n + 1 \leq j \leq n + n_Y$.

(c) The set of lines $X = X_i$, $n+1 \leq i \leq n+n_X$, corresponding to "X only" observations, and the

set of lines $Y = Y_j$, $n+1 \le j \le n+n_Y$, corresponding to "Y only" observations, excluding points described in (b).

(d) The rest of the (X, Y) plane.

Note that any reallocation of mass from (d) to (a) or (b) increases the likelihood. On the other hand, if there is mass in any portion of the line $X = X_i$ excluding the points of intersection of this line with the lines $Y = Y_j$, $n+1 \le j \le n+n_Y$, then its transfer to the intersection points will increase $\prod_{j=n+1}^{n+n_Y} P(Y = Y_j)$ without changing $\prod_{i=n+1}^{n+n_X} P(X = X_i)$ or $\prod_{i=1}^{n} P(X = X_i, Y = Y_i)$. A similar argument holds for any mass allocated on the line $Y = Y_j$, $n+1 \le j \le n+n_Y$. Thus, reallocation of mass from (c) to (b) increases the likelihood. Therefore the optimum allocation of mass on (c) and (d) will be zero.

Thus, (2.1) is maximized when mass is distributed only on all possible intersection points of the lines, $X = X_i$, $n+1 \le i \le n+n_X$, with the lines $Y = Y_j$, $n+1 \le j \le n+n_Y$, together with the points $(X_i, Y_i), 1 \le i \le n$.

The above argument is directly extended to the p-variate case by replacing the (X, Y) plane with the $2p$-dimensional Euclidean space and 'lines' with suitable p-dimensional hyperplanes.

Denoting the mass at the point (X_i, Y_j) by $m_{i,j}$, we have the likelihood profile

$$L = \left\{\prod_{i=1}^{n} m_{i,i}\right\} \left\{\prod_{i=n+1}^{n+n_X} \left(\sum_{j=n+1}^{n+n_Y} m_{i,j}\right)\right\} \left\{\prod_{j=n+1}^{n+n_Y} \left(\sum_{i=n+1}^{n+n_X} m_{i,j}\right)\right\} \tag{A.1}$$

which has to be maximized with respect to the probability masses subject to the condition

$$\sum_{i=1}^{n} m_{i,i} + \sum_{i=n+1}^{n+n_X} \sum_{j=n+1}^{n+n_Y} m_{i,j} = 1.$$

By using the fact that arithmetic mean dominates geometric mean, we have

$$\begin{aligned}
L &\le \left\{\prod_{i=1}^{n} m_{i,i}\right\} \left\{\frac{1}{n_X} \sum_{i=n+1}^{n+n_X} \sum_{j=n+1}^{n+n_Y} m_{i,j}\right\}^{n_X} \left\{\prod_{j=n+1}^{n+n_Y} \left(\sum_{i=n+1}^{n+n_X} m_{i,j}\right)\right\} \\
&\le \left\{\prod_{i=1}^{n} m_{i,i}\right\} \left\{\frac{1}{n_X} \sum_{i=n+1}^{n+n_X} \sum_{j=n+1}^{n+n_Y} m_{i,j}\right\}^{n_X} \left\{\frac{1}{n_Y} \sum_{j=n+1}^{n+n_Y} \sum_{i=n+1}^{n+n_X} m_{i,j}\right\}^{n_Y} \\
&\le \left\{\frac{1}{n} \sum_{i=1}^{n} m_{i,i}\right\}^{n} \left\{\frac{1}{n_X} \sum_{i=n+1}^{n+n_X} \sum_{j=n+1}^{n+n_Y} m_{i,j}\right\}^{n_X} \left\{\frac{1}{n_Y} \sum_{j=n+1}^{n+n_Y} \sum_{i=n+1}^{n+n_X} m_{i,j}\right\}^{n_Y} \\
&= \left\{\frac{m}{n}\right\}^{n} \left\{\frac{1-m}{n_X}\right\}^{n_X} \left\{\frac{1-m}{n_Y}\right\}^{n_Y},
\end{aligned}$$

where $m = \sum_{i=1}^{n} m_{i,i}$. It can be easily verified by differentiation that the last expression is maximized with respect to m when $m = n/(n + n_X + n_Y)$. The last of the three preceding inequalities holds with equality if and only if all the $m_{i,i}$'s are equal, i.e., (2.2) holds. The other two inequalities hold if and only if

$$\begin{aligned}
\sum_{j=n+1}^{n+n_Y} m_{i,j} &= \frac{n_X + n_Y}{n_X(n + n_X + n_Y)}, \qquad n+1 \le i \le n+n_X, \\
\sum_{i=n+1}^{n+n_X} m_{i,j} &= \frac{n_X + n_Y}{n_Y(n + n_X + n_Y)}, \qquad n+1 \le j \le n+n_Y.
\end{aligned}$$

The stated result follows. □

Proof of Theorem 2. Using a similar argument as in Theorem 1, the optimum arrangement of masses that maximizes the likelihood allocates masses on the points corresponding to the paired data and the lines $X = X_i$, $n+1 \le i \le n+n_X$. Thus, the likelihood reduces to $\prod_{i=1}^{n} m_{i,i} \prod_{i=n+1}^{n+n_X} \mu_i$ where $m_{i,i}$ is the mass on the point (X_i, Y_i), $1 \le i \le n$, and μ_i is the mass on the line $X = X_i$, $n+1 \le i \le n+n_X$. This expression is maximized subject to the condition $\sum_{i=1}^{n} m_{i,i} + \sum_{i=n+1}^{n+n_X} \mu_i = 1$ for $m_{i,i} = \frac{1}{n+n_X}$, $\mu_i = \frac{1}{n+n_X}$. This leads us to the form of NPMLE given by (2.5). □

Proof of Theorem 3. (i) Let us use write $\phi_p(y; \mu, \Sigma)$ for the density (at point y) of the p-variate normal distribution with mean μ and dispersion matrix Σ. Let $F_{X,Y}$ have mean vector $\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$ and dispersion matrix $\begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}$. Let $\Sigma_{Y \cdot X} = \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$. Since $f_{X,Y}$ and K are both assumed to be normal, we can write, using (2.19),

$$
\begin{aligned}
&\tilde{f}_Y(y) \\
&= \frac{1}{(n+n_X)h^p} \sum_{i=1}^{n} \prod_{j=1}^{p} \phi\left(\frac{y_j - Y_{i,j}}{h}\right) \\
&\quad + \frac{1}{(n+n_X)h^p} \sum_{i=n+1}^{n+n_X} \int f_{Y|X=X_i}(z) \left\{ \prod_{j=1}^{p} \phi\left(\frac{y_j - z_j}{h}\right) \right\} dz_1 \cdots dz_p \\
&= \frac{1}{n+n_X} \sum_{i=1}^{n} \phi\left(y; Y_i, h^2 I\right) \\
&\quad + \frac{1}{n+n_X} \sum_{i=n+1}^{n+n_X} \int \phi\left(y; z, h^2 I\right) \phi\left(z; \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(X_i - \mu_X), \Sigma_{Y \cdot X}\right) dz \\
&= \frac{1}{n+n_X} \sum_{i=1}^{n} \phi\left(y; Y_i, h^2 I\right) + \frac{1}{n+n_X} \sum_{i=n+1}^{n+n_X} \phi\left(y; \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(X_i - \mu_X), \Sigma_{Y \cdot X} + h^2 I\right).
\end{aligned}
$$

Considering the first two moments of the summands, note that

$$
\begin{aligned}
E_{Y_i} \phi\left(y; Y_i, h^2 I\right) &= \phi\left(y; \mu_Y, \Sigma_Y + h^2 I\right), \\
E_{X_i} \phi\left(y; \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(X_i - \mu_X), \Sigma_{Y \cdot X} + h^2 I\right) & \\
&= \phi\left(y; \mu_Y, \Sigma_Y + h^2 I\right), \\
E_{Y_i} \phi^2\left(y; Y_i, h^2 I\right) &= \frac{1}{(4\pi)^{p/2} h^p} \phi\left(y; \mu_Y, \Sigma_Y + (h^2/2) I\right), \\
E_{X_i} \phi^2\left(y; \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(X_i - \mu_X), \Sigma_{Y \cdot X} + h^2 I\right) & \\
&= \frac{1}{(4\pi)^{p/2}} |\Sigma_{Y \cdot X} + h^2 I|^{-1/2} \times \phi\left(y; \mu_Y, \tfrac{1}{2}(\Sigma_Y + h^2 I + \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})\right).
\end{aligned}
$$

It follows that

$$
E\tilde{f}_Y(y) = \phi\left(y; \mu_Y, \Sigma_Y + h^2 I\right),
$$

$$
\begin{aligned}
var\{\tilde{f}_Y(y)\} &= \frac{1}{(n+n_X)^2}\sum_{i=1}^{n} var\left\{\phi\left(y; Y_i, h^2 I\right)\right\} \\
&\quad + \frac{1}{(n+n_X)^2}\sum_{i=n+1}^{n+n_X} var\left\{\phi\left(y; \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(X_i - \mu_X), \Sigma_{Y\cdot X} + h^2 I\right)\right\} \\
&= \frac{n}{(n+n_X)^2}\frac{1}{(4\pi)^{p/2}h^p}\phi\left(y; \mu_Y, \Sigma_Y + (h^2/2)I\right) \\
&\quad + \frac{n_X}{(n+n_X)^2}\frac{1}{(4\pi)^{p/2}}|\Sigma_{Y\cdot X} + h^2 I|^{-1/2}\phi\left(y; \mu_Y, \tfrac{1}{2}(\Sigma_Y + h^2 I + \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})\right) \\
&\quad - \frac{1}{n+n_X}\phi^2\left(y; \mu_Y, \Sigma_Y + h^2 I\right).
\end{aligned}
$$

Therefore, the IMSE of $\tilde{f}_Y$ is

$$
\begin{aligned}
&IMSE(\tilde{f}_Y) \\
&= \int E\{\tilde{f}_Y(y) - f_Y(y)\}^2 dx \quad = \quad \int \left[Bias\{\tilde{f}_Y(y)\}\right]^2 dy + \int var\{\tilde{f}_Y(y)\}dy \\
&= \int \left\{\phi\left(y; \mu_Y, \Sigma_Y + h^2 I\right) - \phi\left(y; \mu_Y, \Sigma_Y\right)\right\}^2 dy + \frac{n}{(n+n_X)^2}\frac{1}{(4\pi)^{p/2}h^p}\int \phi\left(y; \mu_Y, \Sigma_Y + \tfrac{1}{2}h^2 I\right) dy \\
&\quad + \frac{n_X}{(n+n_X)^2}\frac{1}{(4\pi)^{p/2}}|\Sigma_{Y\cdot X} + h^2 I|^{-1/2}\int \phi\left(y; \mu_Y, \tfrac{1}{2}(\Sigma_Y + h^2 I + \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})\right) dy \\
&\quad - \frac{1}{n+n_X}\int \phi^2\left(y; \mu_Y, \Sigma_Y + h^2 I\right) dy \\
&= \int \phi^2\left(y; \mu_Y, \Sigma_Y + h^2 I\right) dy + \int \phi^2\left(y; \mu_Y, \Sigma_Y\right) dy - 2\int \phi\left(y; \mu_Y, \Sigma_Y + h^2 I\right)\phi\left(y; \mu_Y, \Sigma_Y\right) dy \\
&\quad + \frac{n}{(n+n_X)^2 2^p h^p \pi^{p/2}} + \frac{n_X}{(n+n_X)^2 2^p \pi^{p/2}|\Sigma_{Y\cdot X} + h^2 I|^{1/2}} - \frac{1}{n+n_X}\int \phi^2\left(y; \mu_Y, \Sigma_Y + h^2 I\right) dy \\
&= \frac{1}{(4\pi)^{p/2}|\Sigma_Y + h^2 I|^{1/2}}\int \phi\left(y; \mu_Y, \tfrac{1}{2}(\Sigma_Y + h^2 I)\right) dy + \frac{1}{(4\pi)^{p/2}|\Sigma_Y|^{1/2}}\int \phi\left(y; \mu_Y, \tfrac{1}{2}\Sigma_Y\right) dy \\
&\quad - \frac{2}{(2\pi)^{1/2}|2\Sigma_Y + h^2 I|^{1/2}}\int \phi\left(y; \mu_Y, \left(\Sigma_Y^{-1} + (\Sigma_Y + h^2 I)^{-1}\right)^{-1}\right) dy \\
&\quad + \frac{n}{(n+n_X)^2(4\pi)^{p/2}h^p} + \frac{n_X}{(n+n_X)^2(4\pi)^{p/2}|\Sigma_{Y\cdot X} + h^2 I|^{1/2}} \\
&\quad - \frac{1}{(n+n_X)(4\pi)^{p/2}|\Sigma_Y + h^2 I|^{1/2}}\int \phi\left(y; \mu_Y, \tfrac{1}{2}(\Sigma_Y + h^2 I)\right) dy \\
&= \frac{1}{(4\pi)^{p/2}}\left[\frac{1}{|\Sigma_Y + h^2 I|^{1/2}} + \frac{1}{|\Sigma_Y|^{1/2}} - \frac{2}{|\Sigma_Y + \frac{1}{2}h^2 I|^{1/2}}\right. \\
&\quad \left. + \frac{n}{(n+n_X)^2 h^p} + \frac{n_X}{(n+n_X)^2|\Sigma_{Y\cdot X} + h^2 I|^{1/2}} - \frac{1}{(n+n_X)|\Sigma_Y + h^2 I|^{1/2}}\right].
\end{aligned}
$$

The IMSE of $\hat{f}_Y$ is obtained by substituting $n_X = 0$ in the above expression, i.e.,

$$
\begin{aligned}
IMSE(\hat{f}_Y) &= \int E\{\hat{f}_Y(y) - f_Y(y)\}^2 dy \\
&= \frac{1}{(4\pi)^{p/2}}\left[\frac{1}{|\Sigma_Y + h^2 I|^{1/2}} + \frac{1}{|\Sigma_Y|^{1/2}} - \frac{2}{|\Sigma_Y + \frac{1}{2}h^2 I|^{1/2}}\right. \\
&\quad \left. + \frac{1}{nh^p} - \frac{1}{n|\Sigma_Y + h^2 I|^{1/2}}\right].
\end{aligned}
$$

By comparing the IMSE expressions, we have

$$\begin{aligned}
&IMSE(\hat{f}_Y) - IMSE(\tilde{f}_Y) \\
&= \frac{1}{(4\pi)^{p/2}}\left[\frac{1}{nh^p} - \frac{1}{n|\Sigma_Y + h^2 I|^{1/2}} - \frac{n}{(n+n_X)^2 h^p}\right. \\
&\quad \left. - \frac{n_X}{(n+n_X)^2|\Sigma_{Y\cdot X} + h^2 I|^{1/2}} + \frac{1}{(n+n_X)|\Sigma_Y + h^2 I|^{1/2}}\right] \\
&= \frac{1}{(4\pi)^{p/2}}\left[n\left(\frac{1}{n^2} - \frac{1}{(n+n_X)^2}\right)\frac{1}{h^p} - \frac{n_X}{(n+n_X)^2|\Sigma_{Y\cdot X} + h^2 I|^{1/2}}\right. \\
&\quad \left. - \left(\frac{1}{n} - \frac{1}{n+n_X}\right)\frac{1}{|\Sigma_Y + h^2 I|^{1/2}}\right].
\end{aligned}$$

The Löwner order among the positive definite matrices $h^2 I \le \Sigma_{Y\cdot X} + h^2 I \le \Sigma_Y + h^2 I$ implies the algebraic order of the corresponding determinants. Therefore,

$$\begin{aligned}
&IMSE(\hat{f}_Y) - IMSE(\tilde{f}_Y) \\
&\ge \frac{1}{(4\pi)^{p/2}}\left[n\left(\frac{1}{n^2} - \frac{1}{(n+n_X)^2}\right)\frac{1}{|h^2 I|^{1/2}} - \frac{n_X}{(n+n_X)^2|h^2 I|^{1/2}} - \left(\frac{1}{n} - \frac{1}{n+n_X}\right)\frac{1}{|h^2 I|^{1/2}}\right] \\
&= \frac{1}{(4\pi)^{p/2}|h^2 I|^{1/2}}\left[\frac{1}{n} - \frac{n}{(n+n_X)^2} - \frac{1}{n+n_X} + \frac{n}{(n+n_X)^2} - \frac{1}{n} + \frac{1}{n+n_X}\right] \\
&= 0.
\end{aligned}$$

This completes the proof of (i).

(ii) Let us write the estimators (2.10), (2.15) and (2.13) as $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ respectively. Further, define $\hat{f}_4$ by

$$\hat{f}_4(y) = \frac{1}{n_Y h^p}\sum_{i=n+1}^{n+n_Y}\prod_{j=1}^{p} K\left(\frac{y_j - Y_{i,j}}{h}\right).$$

Note that

$$\begin{aligned}
\hat{f}_1 &= \frac{n}{n+n_X+n_Y}\hat{f}_3 + \frac{n_X+n_Y}{n+n_X+n_Y}\hat{f}_4, \\
\text{while } \hat{f}_2 &= \frac{n}{n+n_Y}\hat{f}_3 + \frac{n_Y}{n+n_Y}\hat{f}_4.
\end{aligned}$$

It follows from the proof of (i) that

$$\begin{aligned}
E\hat{f}_3(y) &= \phi(y;\mu_Y, \Sigma_Y + h^2 I), \\
E\hat{f}_4(y) &= \phi(y;\mu_Y, \Sigma_Y + h^2 I), \\
var\{\hat{f}_3(y)\} &= \frac{1}{n}A(y), \\
var\{\hat{f}_4(y)\} &= \frac{1}{n_Y}A(y), \\
\text{where } A(y) &= \left[\frac{1}{(4\pi)^{p/2}h^p}\phi\left(y;\mu_Y, \Sigma_Y + (h^2/2)I\right) - \phi^2\left(y;\mu_Y, \Sigma_Y + h^2 I\right)\right].
\end{aligned}$$

Consequently, for $\alpha \in (0,1)$,

$$\begin{aligned}
E\{\alpha\hat{f}_3(y) - (1-\alpha)\hat{f}_4(y)\} &= \phi(y;\mu_Y, \Sigma_Y + h^2 I), \\
var\{\alpha\hat{f}_3(y) - (1-\alpha)\hat{f}_4(y)\} &= \left[\frac{\alpha^2}{n^2} + \frac{(1-\alpha)^2}{n_Y^2}\right]A(y),
\end{aligned}$$

and the corresponding IMSE is

$$\begin{aligned} & IMSE\{\alpha\hat{f}_3 - (1-\alpha)\hat{f}_4\} \\ = & \int \Big(\phi(y;\mu_Y,\Sigma_Y+h^2I) - \phi(y;\mu_Y,\Sigma_Y)\Big)^2 dy + \left[\frac{\alpha^2}{n^2} + \frac{(1-\alpha)^2}{n_Y^2}\right]\int A(y)dy \\ = & \int \Big(\phi(y;\mu_Y,\Sigma_Y+h^2I) - \phi(y;\mu_Y,\Sigma_Y)\Big)^2 dy \\ & + \left[\frac{\alpha^2}{n^2} + \frac{(1-\alpha)^2}{n_Y^2}\right]\frac{1}{(4\pi)^{p/2}}\left[\frac{1}{|h^2I|^{1/2}} - \frac{1}{|\Sigma_Y+h^2I|^{1/2}}\right]. \end{aligned}$$

The last expression is minimized when $\alpha = n/(n+n_Y)$, which corresponds to $\hat{f}_2$. □

Proof of Theorem 5. (i) For the given choice of G_i,

$$\hat{F}_{X,Y}(x,y) = \frac{1}{n+n_X}\sum_{i=1}^{n} I(X_i \le x, Y_i \le y) + \frac{1}{n+n_X}\sum_{i=n+1}^{n+n_X} I(X_i \le x)F_{Y|X=X_i}(y).$$

Let $\xi_i = I(X_i \le x)F_{Y|X=X_i}(y)$. Note that, for fixed x and y, the ξ_i's are i.i.d. random variables with

$$E(\xi_i) = E_X(I(X_i \le x)F_{Y|X=X_i}(y)) = \int\limits_{u\le x}\int\limits_{v\le y} f_{Y|X=u}(v)f_X(u)dudv = F_{X,Y}(x,y).$$

Hence, by strong law of large number (SLLN),

$$\frac{1}{n+n_X}\sum_{i=n+1}^{n+n_X} I(X_i \le x)F_{Y|X=X_i}(y) \to (1-k)F_{X,Y}(x,y) \text{ almost surely.}$$

Again by SLLN, $\frac{1}{n+n_X}\sum_{i=1}^{n} I(X_i \le x, Y_i \le y) \to kF_{X,Y}(x,y)$ almost surely. Hence (i) is proved.
(ii) We have

$$\hat{F}_{X,Y}(x,y) = \frac{1}{n+n_X}\sum_{i=1}^{n} I(X_i \le x, Y_i \le y) + \frac{1}{n+n_X}\sum_{i=n+1}^{n+n_X} I(X_i \le x)\hat{F}_{Y|X=X_i}(y).$$

Note that under assumption $n/(n+n_X) \to k$, the SLLN ensures that

$$\frac{1}{n+n_X}\sum_{i=1}^{n} I(X_i \le x, Y_i \le y) \to kF_{X,Y}(x,y).$$

Now write the second term as

$$\frac{1}{n+n_X}\sum_{i=n+1}^{n+n_X} I(X_i \le x)[\hat{F}_{Y|X=X_i}(y) - F_{Y|X=X_i}(y)] + \frac{1}{n+n_X}\sum_{i=n+1}^{n+n_X} I(X_i \le x)F_{Y|X=X_i}(y).$$

Note that the first term of the above expression is bounded by $\sup_x |\hat{F}_{Y|X=x}(y) - F_{Y|X=x}(y)|$ and under assumption stated in the theorem, the bound and hence the first term goes to zero in limit. Applying SLLN again, the second term converges to $(1-k)E[I(X_i \le x)F_{Y|X=X_i}(y)] = (1-k)F_{X,Y}(x,y)$. Hence the proof. □

Proof of Theorem 6. (i) We have

$$\begin{aligned}\tilde{f}_Y(y) &= \frac{1}{(n+n_X)h^p}\sum_{i=1}^{n}\prod_{j=1}^{p}K\left(\frac{y_j-Y_{i,j}}{h}\right)\\ &\quad+\frac{n_X}{(n+n_X)h^p}\int f_{Y|X=x}(z)\left\{\prod_{j=1}^{p}K\left(\frac{y_j-z_j}{h}\right)\right\}d\hat{F}_{X;n_X}(x)dz_1\cdots dz_p,\end{aligned}$$

where $\hat{F}_{X;n_X}$ is the EDF based on X_i, $n+1\le i\le n+n_X$. Since $n/(n+n_X)\to k$, convergence (in probability) of the first term in the expression of $\tilde{f}_Y(y)$ to $kf_Y(y)$ is established by standard techniques (Silverman, 1986, chapter 3). We need to show that

$$\frac{n_X}{(n+n_X)h^p}\int f_{Y|X=x}(z)\left\{\prod_{j=1}^{p}K\left(\frac{y_j-z_j}{h}\right)\right\}d\hat{F}_{X;n_X}(x)dz_1\cdots dz_p\xrightarrow{P}(1-k)f_Y(y).$$

Write

$$\begin{aligned}&\frac{1}{h^p}\int f_{Y|X=x}(z)\left\{\prod_{j=1}^{p}K\left(\frac{y_j-z_j}{h}\right)\right\}d\hat{F}_{X;n_X}(x)dz_1\cdots dz_p\\ &\quad=\frac{1}{h^p}\int f_{Y|X=x}(z)\left\{\prod_{j=1}^{p}K\left(\frac{y_j-z_j}{h}\right)\right\}d(\hat{F}_{X;n_X}(x)-F_X(x))dz_1\cdots dz_p\\ &\qquad+\frac{1}{h^p}\int f_{Y|X=x}(z)\left\{\prod_{j=1}^{p}K\left(\frac{y_j-z_j}{h}\right)\right\}dF_X(x)dz_1\cdots dz_p.\end{aligned}$$

The second term simplifies to $\int f_Y(y-ht)K(t_1)\cdots K(t_p)dt_1\cdots dt_p$ which, under the assumption and by standard techniques, converges to $f_Y(y)$.

To show that the first term converges to zero, write it as $\int g_h(x)d(\hat{F}_{X;n_X}(x)-F_X(x))$ where $g_h(x)=\int f_{Y|X=x}(y-ht)K(t_1)\cdots K(t_p)dt_1\cdots dt_p$. Under the assumption of uniform (in x) equicontinuity (in h) of $g_h(x)$, we have

$$\int g_h(x)d(\hat{F}_{X;n_X}(x)-F_X(x))\to 0\quad\text{almost surely.}$$

Hence the result.

(ii) We have

$$\begin{aligned}\tilde{f}_Y(y) &= \frac{1}{(n+n_X+n_Y)h^p}\sum_{i=1}^{n+n_Y}\prod_{j=1}^{p}K\left(\frac{y_j-Y_{i,j}}{h}\right)\\ &\quad+\frac{n_X}{(n+n_X+n_Y)h^p}\int f_{Y|X=x}(z)\left\{\prod_{j=1}^{p}K\left(\frac{y_j-z_j}{h}\right)\right\}d\hat{F}_{X;n_X}(x)dz_1\cdots dz_p,\end{aligned}$$

where $\hat{F}_{X;n_X}$ is the EDF based on X_i, $n+1\le i\le n+n_X$. Since $(n+n_Y)/(n+n_X+n_Y)\to k$, convergence (in probability) of the first term in the expression of $\tilde{f}_Y(y)$ to $kf_Y(y)$ is established by standard techniques. The rest of the proof follows along the lines of the proof of part (i), with $n+n_Y$ and $n+n_X+n_Y$ in places of n and $n+n_X$, respectively. □

Proof of Theorem 7. (i) Let, for population i, $i = 1, 2$, $BMISP_i$ be the average Bayes misclassification probability and $MISP_i$ be the average misclassification probability of rule given in (4.2). Note that

$$\begin{aligned} BMISP_1 &= \int P(f_X(z) < cf_Y(z)) f_X(z) dz, \\ \text{and } MISP_1 &= \int P(\tilde{f}_X(z) < c\tilde{f}_Y(z)) f_X(z) dz, \end{aligned}$$

where $\tilde{f}_X$ and $\tilde{f}_Y$ are given by (2.14) and (2.19).

Now

$$\begin{aligned} |BMISP_1 - MISP_1| &= \left| \int \{P(f_X(z) < cf_Y(z)) - P(\tilde{f}_X(z) < c\tilde{f}_Y(z))\} f_X(z) dz \right| \\ &\leq \int |P(f_X(z) < cf_Y(z)) - P(\tilde{f}_X(z) < c\tilde{f}_Y(z))| f_X(z) dz. \end{aligned}$$

Theorem 6 ensures point-wise convergence of $\tilde{f}_Y$ to f_Y in probability, while standard results on the kernel density estimator (Silverman, 1986, chapter 3) ensure point-wise convergence of $\tilde{f}_X$ to f_X in probability. Thus, $\tilde{f}_X - c\tilde{f}_Y$ converges point-wise to $f_X - cf_Y$ in probability, and hence in distribution. Therefore, for every fixed z, $P(\tilde{f}_X(z) < c\tilde{f}_Y(z))$ converges to $P(f_X(z) < cf_Y(z))$. It follows from the dominated convergence theorem that $|BMISP_1 - MISP_1| \longrightarrow 0$. Similarly $|BMISP_2 - MISP_2| \longrightarrow 0$ and hence the result.

(ii) The proof follows exactly along the lines of the proof of part (i) and is omitted. □

References

ALR – Alliance for Lupus Research (2009). Information about lupus. URL `http://lupusresearch.org/about/about_lupus.html?gclid=CLibirXC05cCFSIgDQod4FEzDA`, accessed on 20-05-2009.

Bandyopadhyay, S. (1979). Two population classification in Gaussian process. *J. Statist. Plann. Inf.* **3** 225–233.

Bandyopadhyay, S. and Bandyopadhyay, S. (2003). Choosing better training sample for classifying an individual into one of two correlated normal populations. *Calcutta Statist. Assoc. Bull.* **54**, 167–180.

Cheng, P.E. (1994). Nonparametric estimation of mean functional with data missing at random. *J. Amer. Statist. Assoc.* **89**, 81–87.

Cheng, P.E. and Chu, C.K. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statist. Sinica* **6**, 63–78.

Cheng, P.E. and Wei, L.J.(1986). Nonparametric inferences under ignorable missing data process and treatment assignment. In *Proceedings of the International Statistical Symposium, Taipei, Vol. 1*, Institute of Statistical Science, Academia Sinica, Taipei, 97–111.

Dasgupta, S. and Bandyopadhyay, S. (1977). Asymptotic expansions of the distributions of some classification statistics and the probabilities of misclassification when the training samples are dependent. *Sankhyā Ser. B* **39**, 12–25.

Hazelton, M.L. (2000). Marginal density estimation from incomplete bivariate data. *Statist. Probab. Letters* **47**, 75–84.

Jones M.C., J.S. Marron and S.J. Sheather (1996). A Brief Survey of Bandwidth Selection for Density Estimation. *J. Amer. Statist. Assoc.* **91**, 401–407.

Leung, C.Y. and Shrivastava, M.S. (1983). Covariate classification for two correlated populations. *Comm. Statist. Theory Methods* **12**, 223–241.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Loader, C.R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist.* **27**, 415-438.

Ong, S.H., Yap, B.W., Ng, K.H. and Bradley, D.A. (2005). Discriminant analysis involving dependence and censoring: trace and minor elemental concentrations of normal and malignant breast tissues. *J. Science and Technology in Tropics* **1**, 87–91.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London.

Sharma, S, Jhala, Y. and Sawarkar, V.B. (2005). Identification of individual tigers (Panthera tigris) from their pugmarks. *J. Zoology* **267**, 9–18.

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.

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

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley, New York.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38**, 290–295.

Titterington, D.M. and Mill, G.M. (1983). Kernel-based density estimates from incomplete data. *J. Roy. Statist. Soc., Ser. B* **45**, 258–266.

Address for correspondence:

Debasis Sengupta
Applied Statistics Unit
Indian Statistical Institute
203 B T Road
Kolkata 7000108
INDIA
email: sdebasis@isical.ac.in