ESTIMATING THE DENSITY OF A FUNCTIONAL OF SEVERAL RANDOM VARIABLES

J. Koronacki *) and U. Luboińska
Institute of Computer Science, Polish Academy of Sciences,
Ordona 21, 01-237 Warsaw, Poland

ABSTRACT

Let $X^{(1)}, \ldots, X^{(m)}$ be m independent, real-valued random variables with unknown densities and let

$$Y = \Psi(X^{(1)}, \ldots, X^{(m)})$$

be their known functional. The aim is to estimate the density of the random variable Y, given n independent realizations of the vector $(X^{(1)}, \ldots, X^{(m)})$.

In the paper, three natural ways of solving the estimation problem, all based on the Rosenblatt-Parzen estimator, are proposed and compared by means of a simulation study. Applicability of each of the three methods is thus assessed. Simulations suggest that the "direct" method based on the observations of Y only, is "almost always" outperformed by the other two methods. Recommendations for choosing between the methods are given.

Key words: density estimation, kernel estimators, estimating density of a functional.

Subject classification: Primary 62G05.

*) The paper was completed while on leave of this author to the Department of Statistics, Rice University, Houston, Texas. Partial support of this author by the Army Research Office (Durham) under DAAL-03-91-G-0210 is acknowledged.

1. Introduction

Let $X^{(1)}, \ldots, X^{(m)}$ be m independent, real-valued random variables (r.v.'s) with unknown densities, $f^{(1)}, \ldots, f^{(m)}$, respectively. Let

$$Y = \Psi(X^{(1)}, \ldots, X^{(m)})$$

be a known functional of $X^{(1)}, \ldots, X^{(m)}$. Our aim is to estimate the density g of the r.v. Y, given n independent realizations of the vector $(X^{(1)}, \ldots, X^{(m)})$ and, thus, of Y.

Practical examples of the problem posed are provided, e.g., by hydrological studies. One of them is that of estimating the probability density of annual maximum river flows at a given location, when each maximum flow observed can be assigned to one of several (also recorded and supposedly mutually independent) sources, such as heavy rain, light rain of long duration, thaw; in this case,

$$Y = MAX\{X^{(1)}, \ldots, X^{(m)}\}. \tag{1}$$

In electronics, densities of such functionals as

$$Y = MAX\{X^{(1)}, \ldots, X^{(m)}\} - MIN\{X^{(1)}, \ldots, X^{(m)}\} \tag{2}$$

or

$$Y = \sum_{i=1}^{m} X^{(i)} \tag{3}$$

happen to be of interest.

In the paper, three natural ways of solving the estimation problem are proposed, all based on using the Rosenblatt-Parzen estimator,

$$f_k(x) \equiv f_k(x; V_1, \ldots, V_k) = \frac{1}{kh} \sum_{s=1}^{k} K\left(\frac{x - V_s}{h}\right), \qquad -\infty < x < \infty, \tag{4}$$

where k is the sample size, $V_1, \ldots, V_k$ are the observed values of a r.v. whose density is being estimated, K is a kernel and $h > 0$ is a bandwidth (smoothing parameter) of the estimator. Throughout, kernel K is assumed to be the standard Gaussian density. The choice of the bandwidth will be discussed later. For later reference note that

$$f_k(x) = \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{x-v}{h}\right) d\, F_k(v), \tag{4'}$$

where F_k is the empirical distribution function (e.d.f.) of the underlying distribution of $V_1, \ldots, V_k$.

The methods of estimating the density g are introduced in Section 2. The simplest of them is based on the observations of the r.v. Y only. The other two make use of the information provided by the observations of the whole vector $(X^{(1)}, \ldots, X^{(m)})$.

In Section 3, a simulation study is performed for assessing the three methods for small sample sizes n. The methods are compared for functionals (1) – (3) and the $X^{(i)}$-s generated from normal and lognormal distributions. It will be seen that the examples considered provide sufficient information for the evaluation of the methods proposed.

In all cases discussed in Section 3, bandwidth h in (4) is of the form

$$h = 0.9\ min\{\hat{\sigma}, q/1.34\}\ k^{-1/5}, \tag{5}$$

where $\hat{\sigma}$ is the sample standard deviation and q is the sample interquartile range of $V_1, \ldots, V_k$. A detailed justification for using this bandwidth is given in Section 3.4.2 of Silverman (1986). A small scale simulation study with other bandwidths is briefly described in Section 4. Also in Section 4, conclusions drawn from all the simulations are summarized.

2. The Estimators

The most obvious method of estimation of g, to be referred to as the direct method or method 1, is obtained from (4) via substituting $k = n$, and $V_s = Y_s$, $s = 1, \ldots, n$, where

$$Y_s = \Psi(X_s^{(1)}, \ldots, X_s^{(m)}).$$

Thus, the method is based on the observed values of Y only. It is intuitively clear that better methods should be available because of an obvious loss of information (included in the vectors $(X_s^{(1)}, \ldots, X_s^{(m)})$, $s = 1, \ldots, n$) by the direct method.

At least in principle, given densities $f^{(1)}, \dots, f^{(m)}$, one can compute g as a suitable function of the $f^{(i)}$-s,

$$g = h_\Psi(f^{(1)}, \dots, f^{(m)}), \tag{6}$$

where h_Ψ is determined by the functional Ψ. It is, therefore, natural to use samples $X_1^{(i)}, \dots, X_n^{(i)}, i = 1, \dots, m$, to estimate independently densities $f^{(i)}$ and then to substitute the estimates $f_n^{(1)}, \dots, f_n^{(m)}$ thus obtained in (6),

$$g_n^{(A)} = h_\Psi(f_n^{(1)}, \dots, f_n^{(m)}).$$

The estimator $g_n^{(A)}$ of g will be referred to as the analytical method or method 2. It is a different question whether one can give a closed form formula for h_Ψ in a particular case considered. Moreover, even if this question can be answered in the affirmative, it may occur that the formula obtained is numerically intractable. For instance, the forms of h_Ψ are obvious for the Ψ-s given by (1) – (3) but, while $g_n^{(A)}$ can be easily calculated for (1) and (2), it is virtually an impossible task in the case of Ψ given by (3) and m greater than, say, 3. Actually, the analytical estimator can be used rather rarely. Another problem is that, with m increasing, the properties of $g_n^{(A)}$ may be expected to worsen due to the "cumulative effect" of errors associated with the estimation of each of the densities $f^{(i)}, i = 1, \dots, m$.

In order to introduce the third method, let us first refer to the direct one (see definition (4′)) and recall that the e.d.f. of Y is of the form

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \chi(\Psi(X_i^{(1)}, \dots, X_i^{(m)}) \le y),$$

where $\chi(\Psi(X_i^{(1)}, \dots, X_i^{(m)}) \le y)$ denotes the indicator function of the event

$$\Psi(X_i^{(1)}, \dots, X_i^{(m)}) \le y.$$

Now, let us replace $F_n(y)$ by a suitable U-statistic. To motivate such an approach note that the data matrix for the problem considered,

$$\mathbf{X}_n = \begin{bmatrix} X_1^{(1)} & \dots & X_1^{(m)} \\ \vdots & \ddots & \vdots \\ X_n^{(1)} & \dots & X_n^{(m)} \end{bmatrix},$$

has a special structure, namely the data are both row–wise and column–wise independent. Thus, when constructing an estimator of the distribution function F_Y of Y, there is no need to confine to the values of Ψ corresponding to the arguments from the same row of $\mathbf{X}_n$. In fact, instead of using $F_n(y)$, the following generalized U-statistic may be used

$$U_n(y) = \frac{1}{n^m} \sum_c \chi(\Psi(X_{i_1}^{(1)}, X_{i_2}^{(2)}, \ldots, X_{i_m}^{(m)}) \leq y), \tag{7}$$

where i_j denotes an element of the set $\{1, 2, \ldots, n\}, j = 1, 2, \ldots, m$, and $\sum_c$ denotes summation over all n^m possible sequences of indices $(i_1, i_2, \ldots, , i_m)$. It is easy to verify that $U_n(y)$ is an unbiased, strongly consistent and asymptotically normal estimator of F_Y at y, $F_Y(y)$; see Serfling (1980), Sec. 5.1.3 and Problems 5.P.14 and 5.P.19. Moreover, following the lines of the proof of the Theorem on p. 176 of Serfling (1980), one readily obtains that

$$Var\ U_n(y) < Var\ F_n(y) \tag{8}$$

(although, as should be expected, $Var\ U_n(y) = O(n^{-1})$, as is the case for $F_n(y)$).

In view of the properties of $U_n(y)$, in particular in view of (8), it is indeed natural to replace $F_n(y)$ by $U_n(y)$ in (4′). However, the number n^m is usually astronomically large and, hence, it is reasonable to approximate $U_n(y)$ by a simpler statistic, $U_{n^*}^*(y)$, obtained by means of a suitable "Monte Carlo" experiment. The method based on $U_{n^*}^*(y)$ in lieu of $F_n(y)$, and with n replaced by n^*, will be referred to as the Monte Carlo (MC) method or method 3. The MC method is constructed in two steps. First, a pseudosample of size n^* is generated through resampling from the original data matrix and then estimator (4) is used:

• Choose at random (i.e., according to the uniform distribution on $\{1, 2, \ldots, n\}$) the index $I_1^{(1)}$ of $X_{I_1^{(1)}}^{(1)}$, an element of the first column of the data matrix $\mathbf{X}$; likewise, obtain the index $I_1^{(2)}$ of $X_{I_1^{(2)}}^{(2)}$, an element of the second column of $\mathbf{X}$, and so on until the m elements from m columns of $\mathbf{X}$, $X_{I_1^{(1)}}^{(1)}$, $X_{I_1^{(2)}}^{(2)}$, $\ldots$, $X_{I_1^{(m)}}^{(m)}$, are obtained; the vector $\left(X_{I_1^{(1)}}^{(1)}, \ldots, X_{I_1^{(m)}}^{(m)}\right)$ determines the first "observation",

$$Y_1^* = \Psi\left(X_{I_1^{(1)}}^{(1)}, \ldots, X_{I_1^{(m)}}^{(m)}\right),$$

of the pseudosample $Y_1^*, \ldots, Y_{n^*}^*$ of the values of Ψ. Repeat the procedure described to obtain the second element of the pseudosample,

$$Y_2^* = \Psi\left(X^{(1)}_{I_2^{(1)}}, \ldots, X^{(m)}_{I_2^{(m)}}\right),$$

and so on until the last element of the pseudosample,

$$Y_{n^*}^* = \Psi\left(X^{(1)}_{I_{n^*}^{(1)}}, \ldots, X^{(m)}_{I_{n^*}^{(m)}}\right),$$

is obtained.

• Given the pseudosample $Y_1^*, \ldots, Y_{n^*}^*$, use (4) with n replaced by n^* and with $V_s = Y_s^*$, $s = 1, \ldots, n^*$, to obtain the MC estimator

$$g_{n^*}^{(MC)}\,(x; Y_1^*, \ldots, Y_{n^*}^*)\,.$$

If we show that $U_{n^*}^*$ approximates U_n, the use of the former in the definition of the MC estimator will be justified. One way of achieving this, in fact of achieving more than is needed here, is to use the "special construction" for the empirical process corresponding to $U_{n^*}^*$ (see Shorack and Wellner (1986), in particular Section 23.1). It follows from the special construction that, for almost every sequence $(X_1^{(1)}, \ldots, X_1^{(m)}), \ldots\,,\ (X_n^{(1)}, \ldots, X_n^{(m)}), \ldots$, conditionally upon $(X_1^{(1)}, \ldots, X_1^{(m)}), \ldots, (X_n^{(1)}, \ldots, X_n^{(m)})$,

$$\sqrt{n^*}\,(U_{n^*}^* - U_n) \xrightarrow{D} B(F_Y) \quad \text{as} \quad n,\ n^* \to \infty,$$

where B is the Brownian bridge and $\xrightarrow{D}$ denotes convergence in distribution. Thus, in particular, for each y, $\sqrt{n^*}\,(U_{n^*}^*(y) - U_n(y))$ is (conditionally) asymptotically normal and, in the given sense, $U_{n^*}^*$ approximates U_n. It is also worthwhile to note that the MC estimator is intuitively appealing in its own right, without any reference to the theory of U-statistics.

3. Simulation Results

Please see Computational Statistics & Data Analysis 18 (1994), 317-330.

References

Ciesielski, Z. (1991), Asymptotic Nonparametric Spline Density Estimation, Probab. Math. Statist., **12**, 1-24.

Hall, P., Sheather, S., Jones, M. and Marron, J.S. (1991), On Optimal Data-based Bandwidth Selection in Kernel Density Estimation, Biometrika, **78**, 263-269.

Jones, M.C. and Kappenman, R.F. (1992), On a Class of Kernel Density Estimate Bandwidth Selectors, Scand. J. Statist., **19**, 337-349.

Park, B.U. and Turlach, B.A. (1992), Practical Performance of Several Data Driven Bandwidth Selectors, Comp. Statist., **7**, 251-270 (see also discussion).

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Sheather, S.J. (1992), The Performance of Six Popular Bandwidth Selection Methods on Some Real Data Sets, Comp. Statist., **7**, 225-250 (see also discussion).

Shorack, G.R. and Wellner, J.A. (1986), *Empirical Processes with Applications to Statistics*, Wiley, New York.

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