AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 1, 119–128

Empirical Density Estimation for Interval Censored Data

Eugenia Stoimenova
Bulgarian Academy of Sciences, Sofia

Abstract: This paper is concerned with the nonparametric estimation of a density function when the data are incomplete due to interval censoring. The Nadaraya-Watson kernel density estimator is modified to allow description of such interval data. An interactive R application is developed to explore different estimates.

Keywords: Kernel Density Estimation, Interactive Density Estimation.

1 Incomplete Data Structure

"All distributions of experimental data are discrete"
(Hall, 1982)

Interval censored data are defined to be the result of observing continuous variables only up to the intervals containing the true values. Such data arise in a number of natural ways. Generally speaking, in the real world all continuous variables are observed and recorded in finite precision. Sometimes this depends on the precision of the measurement instruments an the recorded data are given as points from a equispace grid. For any particular recorded value we realize that the true value belongs to some known interval around this point. Since discretization to a fine grid do not move true data very far, we usually apply algorithms without any embarrassment for the result. Unfortunately, this is not always true. Such a typical situation is describing the distribution of earning data (Schweitzer and Severance-Lossin, 1996). Due to some reasons collected data are grouped in intervals with different lengths and moreover several collections might be merged for analysis.

The common feature of the problems we mention is that true data values are not known exactly, but can be identified with a set of possible values. We refer these data as "incomplete" and when the observed sets are intervals we say that they are interval censored. The problem we are interesting here is nonparametric density estimation based on a sample of interval censored data.

To formulate the problem strictly, let X be a random variable with probability density function f in a sample space $\mathcal{X}$, which is assumed to be an interval on the real line. Further, let $\{C_i\}$ be a partition of the sample space, i.e. a collection of disjoint measurable sets verifying $\bigcup C_i = \mathcal{X}$. The partition is assumed independent of X. Suppose, that instead of observing X exactly one received the datum $Y = h(X)$, which is identified with the subset $S_i \in \mathcal{X}$ into which X fallen. The random variable Y is defined on sample space $\mathcal{Y} \subset \mathcal{X}$ and conditionally on $X = x$ has a probability density $h_X(y|X = x)$.

Hence, if f is known, the marginal density of Y_i is

$$g(y|f) = \int_{\mathcal{X}} h_X(y|x)f(x)dx.$$

The observed data $(Y_1, \dots, Y_n)$ are independent realizations of Y conditionally on unobservable complete realizations of X. The goal is to estimate the density f.

The above general framework covers a wide variety of incomplete observing situations. It is studied in the literature mainly in the context of nonparametric estimation of the distribution function of X. Some references include Laird (1978), Liu and Zhu (2007), Goutis (1997).

We relate this problem to the problem of smoothing over a discrete distribution. Our goal is to achieve more precise approximation for small probabilities, at the same time, find good visual representation of the true density.

Existing literature on smoothing over discrete distributions concentrate mainly on asymptotic properties of the proposed estimators, even when considering sparse tables. The Nadaraya-Watson kernel density estimator is modified by Simonoff (1996) for estimating ordered categorical data. Moreover, local polynomial estimators are proposed for smoothing discrete distributions.

Results on the asymptotic behavior of the mean sum of squares of the errors are studied in Simonoff (1983, 1995), or Aerts, Augustyns, and Janssen (1997). The asymptotics of the local polynomials smoothers were studied by Aerts et al. (1997), establishing sufficient conditions this sparse consistency.

A natural extension of these estimators allows interval data. In Section 2 we give an appropriate definition of empirical density function for interval censored data. In Section 2.1 we derive the maximum likelihood density estimate using uniformity condition. In Section 3 the Nadaraya-Watson density estimator is given. Finally, in Section 5 we present a real data example and demonstrate an R-program for interactive control of the smoothing parameter.

2 Empirical Density Function for Interval Censored Data

In the study of the empirical density function (e.d.f.) the first difficulty is to find its best definition. Revesz (1972) and Revesz (1974) gives a number of possible definitions of the e.d.f. based on a complete sample from the underlying distribution. We will follow this construction to define e.d.f. based on a censored sample.

For interval censored random variable X the subsets $\{S_i\}$ are intervals specified by some known points $\tau_1, \tau_2, \dots$ of the real line which is independent of X. Define

$$R = \inf\{\tau_j : \tau_j \geq X\}, \qquad L = \sup\{\tau_j : \tau_j \leq X\}.$$

The conditional density of X given the interval $Y = [L, R]$ is then

$$f(x|Y) = \frac{f(x)}{\int_Y f(u)du}, \qquad x \in Y. \tag{1}$$

For each X_i the corresponding interval Y_i may be defined on a different partition of $\mathcal{X}$. The conditional density is itself unknown.

Let $Y_1, \dots, Y_n$ be a sample of observed intervals, where $Y_i = [L_i, R_i]$. We assume that each X_i has an equal impact in the sample so the corresponding interval $Y_i = [L_i, R_i]$

should have. A common approach is to assume the conditional distribution over the interval Y_i to be uniform, i.e. we suppose that X_i occurs at any point of the observed interval $[L_i, R_i]$ equally likely and the total mass in the observed interval is $1/n$.

Further, let $[a, b]$ be an interval of the real line and denote by $M_n(a, b)$ the restricted mass of all elements of the sample $Y_1, \dots, Y_n$ into this interval. More precisely, the restriction function is defined as follow:

Notation: For a non-zero interval (c, d) the restriction function over an interval (a, b) is denoted by $J_{(a,b)}(c, d)$ and equals

$$J_{(a,b)}(c,d) = \begin{cases} \dfrac{\min(d,b) - \max(c,a)}{b-a} & \text{if} \quad (a,b) \bigcap (c,d) \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

For two intercept intervals, the restriction function gives the ratio of the length of the common part of the intervals to the length of the second one (here (a, b)). Then $M_n(a, b)$ is the sum of all parts of the observed intervals.

Therefore, the probability $\int_a^b f(t)dt$ can be estimated by the relative mass $\frac{1}{n}M_n(a, b)$ while the value $f(x)$, $a < x \leq b$), can be estimated by $\frac{1}{b-a}\int_a^b f(t)dt$. Therefore, an e.d.f. in the interval $[a, b]$ can be defined by

$$\widehat{f}(x) = \frac{M_n(a,b)}{n(b-a)}, \qquad a < x \leq b.$$

This is the natural way to define the empirical density function. Using $\widehat{f}(x)$ as an estimator of $f(x)$ one makes two errors: the first error is in the fact that $f(x)$ is estimated by its integral mean $\frac{1}{b-a}\int_a^b f(t)dt$; the second is made when the probability $\int_a^b f(t)dt$ is estimated by the relative mass $\frac{1}{n}M_n(a, b)$. The first error is small if the interval $[a, b]$ is short, while the second one is small if the interval contains enough mass of elements, i.e. if the interval $[a, b]$ is not too short.

The nonparametric maximum likelihood estimate for a density with censored data can constructed using this approach.

2.1 Maximum Likelihood Density Estimate from Censored Data

Let $Y_1, \dots, Y_n$ be a sample of observed intervals, where $Y_i = [L_i, R_i]$. Then the mass of the observed intervals $Y_1, \dots, Y_n$ over an interval (a, b) can be expressed by function $J_{(a,b)}$ as a sum of the restricted mass of all Y's.

To construct the maximum likelihood density estimate, let $\mathcal{C}$ be the set of disjoint intervals (cells) defined by the endpoints of all observed intervals $Y_1, \dots, Y_n$, and let $c_1 < c_2 < \dots < c_r$ be the ordered elements of $\mathcal{C}$.

For any observed interval (c, d) the restriction function over an interval $(a, b) \in \mathcal{C}$ is

$$J_{(a,b)}(c,d) = \begin{cases} \dfrac{b-a}{d-c}, & \text{if } c \leq a < b \leq d \\ 0, & \text{otherwise.} \end{cases}$$

That is, any observation covers fully some intervals from the partition $\mathcal{C}$.

Define the mass $\overline{m}(c_i, c_{i+1})$ of any cell $C_i = (c_j, c_{i+1})$ by

$$\overline{m}(c_i, c_{i+1}) = \sum_{i=1}^{n} J_{(c_i,c_{i+1})}(Y_i). \tag{2}$$

Denoting $\overline{p}_i = \frac{1}{n}\overline{m}_i$ we obtain the vector $\overline{p} = (\overline{p}_1, \ldots, \overline{p}_r)$ that represents the observed probability mass of the sample $Y_1, \ldots, Y_n$ over the support of f and obviously, $\sum \overline{p}_i = 1$.

Thus, the nonparametric maximum likelihood estimate of $f(x)$ is a piecewise function with constant values in each subinterval. It equals to the sum of restricted mass of all observed intervals Y_j $(j = 1, \ldots, n)$ on $[c_i, c_{i+1}]$.

The quantity $\overline{p}_i$ is an estimate of the i-th cell true probability

$$p_i = \frac{1}{c_{i+1} - c_i} \int_{c_i}^{c_{i+1}} f(t)dt,$$

which represents the probability that a random variable with density f takes a value within the interval (c_i, c_{i+1}). It is easy to see that for fixed categories specified by endpoints $\mathcal{C}$, the vector $(\overline{p}_1, \ldots, \overline{p}_r)$ has a multinomial distribution with $\mathrm{E}(\overline{p}_i) = p_i$.

The estimate $\overline{p}_i$ is an accurate estimate of the true probability p_i if the probability mass in interval (c_i, c_{i+1}) is large. However, the number of cells might be large compared with the number of observation, the vector $\overline{p}$ could not be a good estimator for true density f. The sparse asymptotics require the size of cells to become zero as the sample size tends to infinity. Nevertheless, this requirement is not relevant to the multinomial distribution, it is useful for modelling of observations in large number of cells. Under sparse asymptotics the cell probabilities $\overline{p}_i$ are not consistent in the sense that if $n \to \infty$ and $\sup(c_{i+1} - c_i) \to 0$,

$$\sup_{1 \le i \le r} \left| \frac{\overline{p}_i}{p_i} - 1 \right| \nrightarrow 0$$

(assuming f is bounded below, so $\inf_i p_i > 0$).

The sparse asymptotic property is more natural for vector $\overline{p}$ obtained from censored data since they may represent probability mass observed in a sequence of binnings. The simplest relation is a sequence of different roundings (see Hall, 1982 for instance).

Since it is reasonably to assume that the mass changes smoothly as x increases, the mass falling in one particular cell provides information about the probability of falling in its neighbors. Therefore, smoothing makes sense since we assume that the probability function is continuous. The improvement using smoothing is most evident when the distribution is sparse in the sense that mass of Y falling each cell are small.

2.2 Kernel Density Estimates

In density estimation with complete data, one observes a sample $X_1, \ldots, X_n$ from a distribution on the real line admitting a density f. The kernel estimator of f, introduced by Rosenblatt (1956), is defined by

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \tag{3}$$

where $X_1, \ldots, X_n$ are the data, $K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$ is a kernel function and $h = h(n)$ is a sequence of constants tending to zero as $n \to \infty$. Kernel function is defined to satisfy the conditions

$$\int K(u)du = 1, \qquad \int uK(u)du = 0, \qquad \int u^2 K(u)du = \sigma_K^2 > 0.$$

We shall consider both the kernel density estimation and the censored data problems.

Local polynomial smoothers for categorical data is suggested by Simonoff (1995) and further studied by Aerts et al. (1997), Baek (1998), etc. For equispace design $c_j = j/K$ with K be the number of cells, the local polynomial estimator $\widehat{p}_i$ is defined as the constant term of the minimizer $\widehat{\beta}_0$ of

$$\sum_{j=1}^{r} \left[\overline{p} - \beta_0 - \cdots - \beta_t \left(\frac{i}{K} - \frac{j}{K} \right) \right]^2 K_h \left(\frac{i}{K} - \frac{j}{K} \right).$$

Asymptotic behavior of this estimate is derived by Aerts et al. (1997). The advantage of $\widehat{p}_i$ over $\overline{p}_i$ is that it is consistent under sparse asymptotics.

3 Modified Nadaraya-Watson Density Estimator

We shall consider the density estimation and the nonlinear regression problems related to density estimation with censored data. We have already mention that for complete data, the kernel estimator of f is defined by (3). In nonparametric regression with complete data, one observes n pairs $(X_1, Z_1), \ldots, (X_n, Z_n)$ obeying the model

$$Z_i = m(X_i) + \epsilon_i,$$

where the regression function $m(x)$ is conditional expectation $m(x) = \mathrm{E}(Z|X = x)$, and the random errors ϵ_i satisfy $\mathrm{E}(\epsilon_i|X = x) = 0$ and $\mathrm{var}(\epsilon_i|X = x) = \sigma^2$ not necessarily constant.

There are many different methods for estimating $m(x)$. Here we focus on a estimator known as Nadaraya-Watson kernel estimator (Nadaraya, 1964). The Nadaraya-Watson kernel estimator for $m(x)$ is defined by

$$\widehat{m}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right) Z_i}{\widehat{f}(x)},$$

where $\widehat{f}(x)$, K and h are as before. Note that this estimator is the local polynomial estimate with $t = 0$

$$\sum_j (Z_j - \beta_0)^2 K_h(x - X_j).$$

When the data are interval censored, so that $X_i \in Y_i$, $i = 1, \ldots, n$, and only $Y_i = [L_i, R_i]$ is observed we define

$$\widehat{p}_i = \frac{\sum_{j=1}^{r} K_h(c_i - c_j)\overline{p}_j}{\sum_{j=1}^{r} K_h(c_i - c_j)}. \tag{4}$$

Therefore $\widehat{p}_i$ is the Nadaraya-Watson kernel regression estimator on the points $(c_j, \overline{p}_j)$, $j = 1, \dots, r$, and $\widehat{p}_i$ is the solution of the natural weighted least squares problem, being the minimizer $\widehat{\beta}_0$ of

$$\sum_j (\overline{p}_j - \beta_0)^2 K_h(c_i - c_j).$$

Thus $\widehat{p}_i$ corresponds to locally approximating p_i with a constant weighting values p_j corresponding to intervals closer to (c_i, c_{i+1}) more heavily.

One of the main problems in density estimation is to find a good connection between the length of the intervals and the size n. This aspect of density estimation is first discussed by Rosenblatt (1956). Since all estimators of the density function under regular assumptions are bias, mean square error is a standard measure that comprising the bias and variance of the estimator.

4 Variable Bandwidth and Interactive Choice

The performance of $\widehat{f}(x)$ depends critically on the choice of h that controls the smoothness of the estimates. There are various methods proposed for automatic selecting an optimal h for uncensored data. Often the optimal value of h minimizes the mean square error or some equivalent functional. Note, that any automatic choice of the smoothing parameter should be view as a benchmark, and need to adjusted based on subjective impression.

In this this study we apply variable bandwidth h_i in (4) depending on the length of cells (c_i, c_{i+1}). Adjusting parameter λ stretch all h_i and gives numerous estimators. The choice of the kernel is known not to have great effect on the practical performance of the estimator. We use Normal kernel everywhere in our examples.

The R implementation of proposed estimators is in interactive graphic environment. Changing options are done in an easy way, getting a rich graphical response. This is R program which uses the tcltk package to provide simple functions for building of a control panel for the graphics.

We remind that in the literature, authors were mainly concerned with the asymptotic behavior when both the number of cells of the distribution and the number of observations were converging to infinity, maintaining some sort of relation implying that the number of cells in the support would not become to small with respect to the size of the sample. In this paper we are mainly interested in the behavior when the sample is fixed and trying to find estimators that have good performance. We have no objective quantification of what good means, so we propose user to explore many different estimator to find appropriate description of the data.

The numerical results included were obtained from KT-DigiCult-Bg project collection of metadata on mediaeval manuscripts studied for another purpose by Stoimenova, Mateev, and Dobreva (2006).

5 Application to Chronologcal Data

Temporal data is often subject to dilatation and translation effects. The most of manuscripts, chronicles and other historical documents that we have today, related to ancient and medieval times, were not dated exactly but dated as a likely periods. The two values "Notbefore" and "Notafter" are common in catalogue descriptions and specifies the interval $[L, R]$ for possible origin of the item and can be also extracted from other types of description (see Stoimenova et al., 2006).

In this section the methods considered in this paper are applied to a sequence of 807 intervals of "Notbefore" and "Notafter" dates of origin of mediaeval manuscripts . Figure 1 represents the data as line segments with length equal to the observed interval length $R-L$ and height approximately proportional to $log(1/R-L)$. The exact dated documents are less that 25% of the data and are depicted at the top level on the Figure 1.

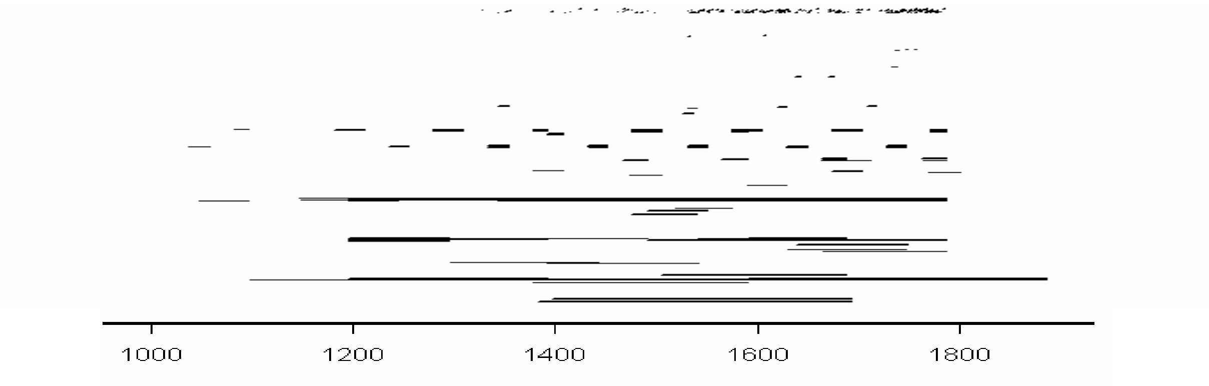


Figure 1: Raw interval data for medieval manuscripts.

A very natural desire is in the informal investigation of the distribution of manuscripts over time. Here density estimates can give valuable indication of such features as skewness and multimodality in the data. Multimodality of an estimate is of interest in describing chronology data, since they specify possible historical upward and downward trend periods of society development.

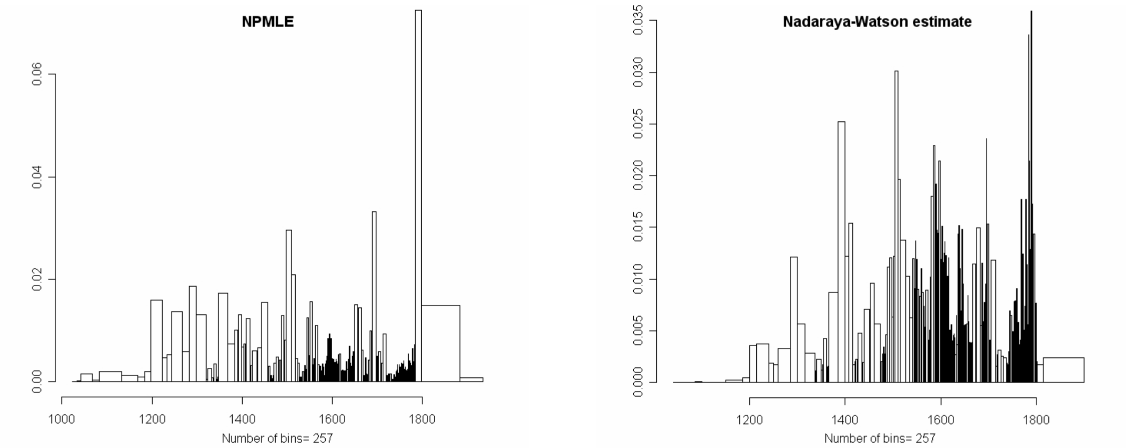
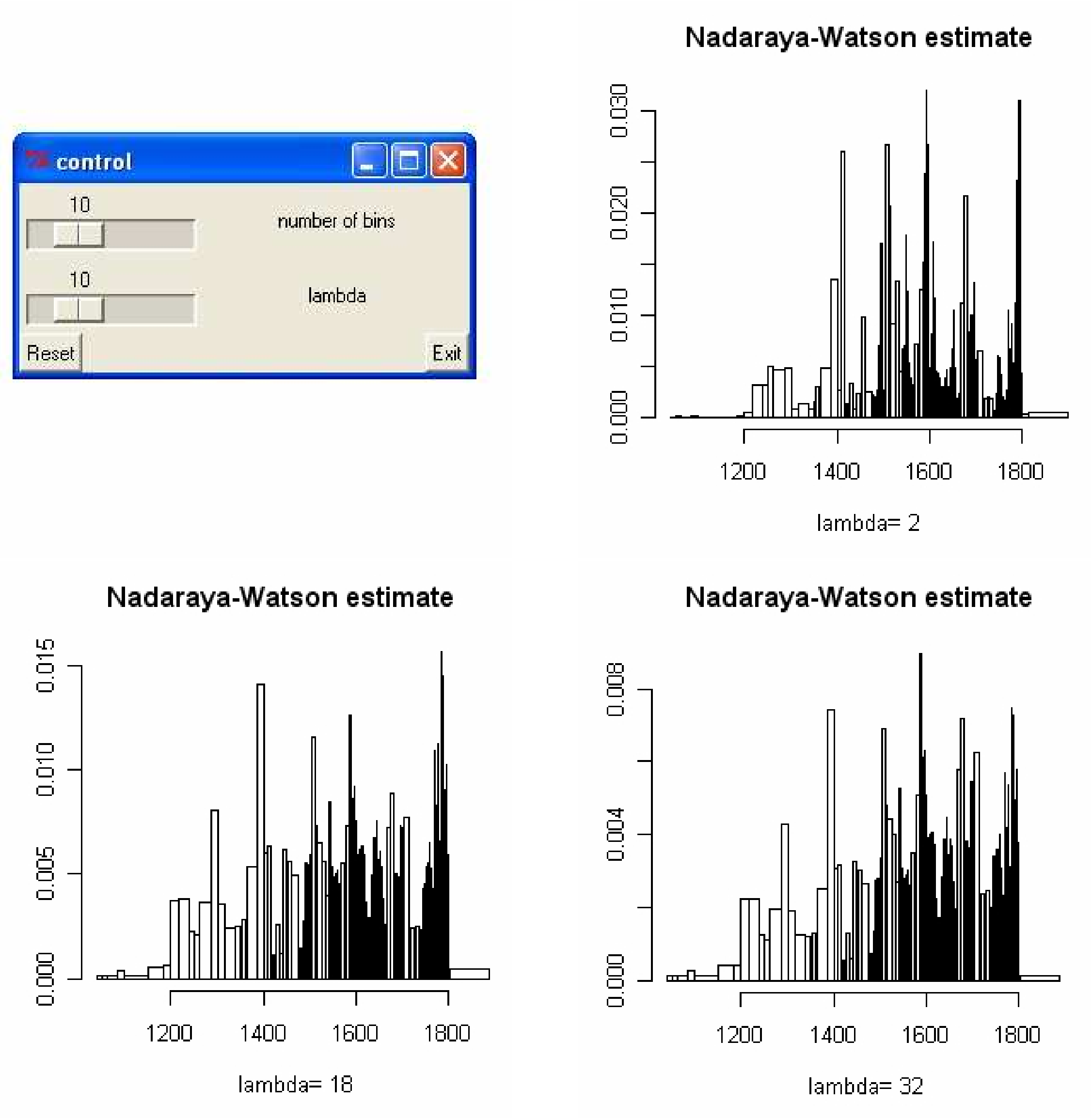


Figure 2: Maximum likelihood density estimate and Nadaraya-Watson estimate.

The left histogram in Figure 2 is the maximum likelihood density estimate calculated using (2) and $\overline{p}_i = \frac{1}{n}\overline{m}_i$. The set $\mathcal{C}$ of endpoints consists of 258 points. The right histogram is a Nadaraya-Watson estimate calculated using (4). Here the variable bandwidths h_i are calculated by $(R_i - L_i)\lambda^2/24$ with $\lambda = 2$.



10 number of bins

10 lambda

Figure 3: Exploring different bandwidths.

Figure 3 illustrates the real-time evolution of the modified Nadaraya-Watson density estimator. The computation begins with a given adjusting parameter λ (the stretch of the bandwidths). Then the user can change it by a slider and explore different shapes of the distribution. Here λ varies in $[2, 40]$. A snapshot of the estimators obtained by using the interactive program is shown. The values of λ used was 2, 18, and 32. The top left picture shows the control panel.

It is clear that using all endpoints $\mathcal{C}$ for visual representation of the estimator does not give an easy interpretative estimator. Second interactive adjusting parameter choose a subset from $\mathcal{C}$ and the modified Nadaraya-Watson estimator is calculated using corresponding intervals in (4).

Figure 4 illustrates different shapes of the distribution for different subsets of $\mathcal{C}$. Slider values varies in $[2,\ 40]$. A snapshot of the estimators obtained by using the interactive program is shown. The number of intervals used was 42, 28, 18, and 9.

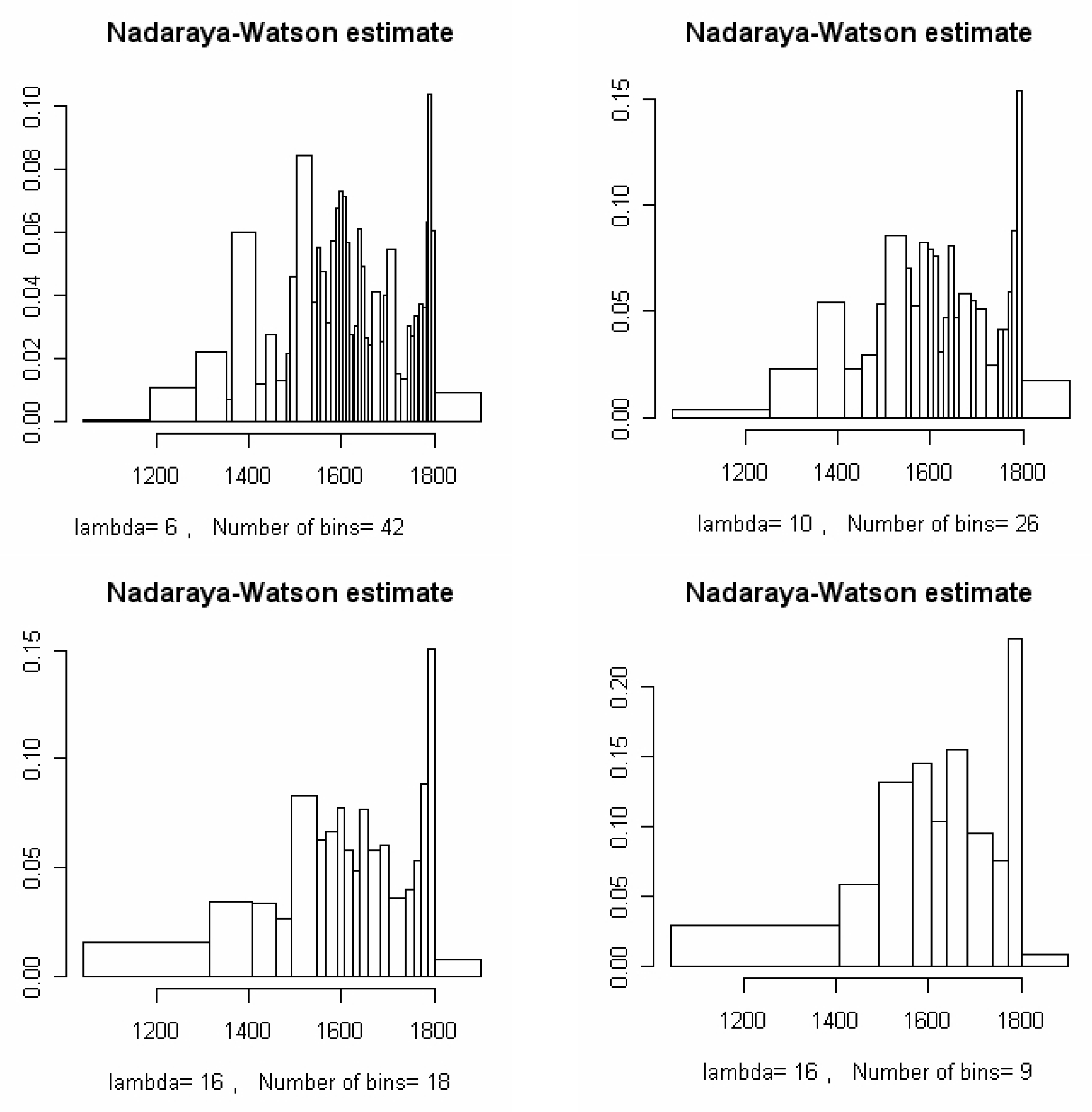


Figure 4: Exploring different endpoints subsets.

References

Aerts, M., Augustyns, I., and Janssen, P. (1997). Smoothing sparse multinomial data using local polynomial fitting. *Journal of Nonparametric Statistics*, *8*, 127-147.

Baek, J. (1998). A local linear kernel estimator for sparse multinomial data. *Journal of the Korean Statistical Society*, *27*, 515-529.

Goutis, C. (1997). Nonparametric estimation of a mixing density via the kernel method. *Journal of the American Statistical Association*, *92*, 1445-1450.

Hall, P. (1982). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *IAM Journal of Applied Mathematics*, *42*, 390-399.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*, 805-811.

Liu, L., and Zhu, Y. (2007). Partially projected gradient algorithms for computing nonparametric maximum likelihood estimates of mixing distributions. *Journal of Statistical Planning and Inference*, *137*, 2509-2522.

Nadaraya, E. A. (1964). On estimating regression. *Teor. Veroyatn. Primen.*, *9*, 157-159.

Revesz, P. (1972). On empirical density function. *Period. Math. Hung.*, *2*, 85-110.

Revesz, P. (1974). On empirical density function. In A. Obretenov (Ed.), *Verojatn. stat. metody; mezdunar. letn. sk. teor. verojatn. mat. stat.* (p. 63-88). Varna: Bulgarian Academy of Sciences.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, *27*, 832-837.

Schweitzer, M. E., and Severance-Lossin, E. K. (1996). *Rounding in earnings data.* (Working Paper 9612, Federal Reserve Bank of Cleveland)

Simonoff, J. S. (1983). A penalty function approach to smoothing large sparse contingency tables. *Annals of Statistics*, *11*, 208–218.

Simonoff, J. S. (1995). Smoothing categorical data. *Journal of Statistical Planning and Inference*, *47*, 41-69.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* New York: Springer.

Stoimenova, E., Mateev, P., and Dobreva, M. (2006). Outlier detection as a method for knowledge extraction from digital resources. *Rev. Nat. Center Digitization*, *9*, 1-11.

Author's Address:

Eugenia Stoimenova
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G.Bontchev str., bl.8
1113 Sofia
Bulgaria

E-mail: `jeni@math.bas.bg`