A Generalization of Histogram Type Estimators *

Pedro Delicado[1] Manuel del Río[2]

[1]Departament d'Economia i Empresa
Universitat Pompeu Fabra

[2] Departamento de Estadística e Investigación Operativa
Universidad Complutense de Madrid

October 20, 1999

*This work was partially supported by DGES PB-960300, 6159/95 UCM projects, and UCM Del Amo Scholarship/97.

Abstract

We introduce simple nonparametric density estimators that generalize the classical histogram and frequency polygon. The new estimators are expressed as linear combination of density functions that are piecewise polynomials, where the coefficients are optimally chosen in order to minimize the integrated square error of the estimator. We establish the asymptotic behaviour of the proposed estimators, and study their performance in a simulation study.

Key-words: Convolution, frequency polygon, nonparametric density estimation, smoothing techniques, splines.
JEL Classification: C13, C14.

1 Introduction

Let $X_1, \ldots, X_n$ be a random sample with unknown density function f. The oldest and widely used estimator of f is the histogram. Consider the set of points $\{a + jh, j \in Z\}, h > 0$; without loss of generality, we can assume that the *anchor* a is 0 (see Simonoff, 1995, and Simonoff and Udina,1997, for discussions about the anchor choice). This set induces the partition $I\!R = \cup_{j \in Z} B_j$, where each interval $B_j = (jh, (j+1)h] = (c_j - h/2, c_j + h/2]$, with length h, is centered at $c_j = hj - (h/2)$. Given this partition and the observed values $X_1, \ldots, X_n$, the histogram estimator of the density f is

$$\hat{f}^H(x) = \frac{n_j}{nh} = \frac{f_j}{h}, \text{ if } x \in B_j,$$

where n_j is the number of observations in B_j, and $f_j = n_j/n$ is the relative frequency of B_j. Note that if we consider $K_U(u) = I_{(-.5,.5]}(u)$, we have

$$\hat{f}^H(x) = \sum_j f_j \frac{1}{h} K_U\left(\frac{x - c_j}{h}\right). \tag{1}$$

Another well known density estimator is the frequency polygon, defined as the function that linearly interpolates the points $\{(c_j, \hat{f}^H(c_j)) : j \in Z\}$. This estimator is given by (see e.g. Simonoff, 1996, p.20),

$$\hat{f}^P(x) = \frac{1}{h^2}(f_j c_{j+1} - f_{j+1} c_j + (f_{j+1} - f_j)x), x \in B_j.$$

If we consider $K_T(u) = (1 - |u|) I_{(-1,1]}(u)$, we have the alternative expression

$$\hat{f}^P(x) = \sum_j f_j \frac{1}{h} K_T\left(\frac{x - c_j}{h}\right). \tag{2}$$

The list of commonly used density estimators also includes *kernel* estimators, defined as

$$\hat{f}_b^K(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right),$$

where the *kernel* function K is usually a symmetric density function, and $b > 0$ is the bandwidth. Kernel density estimation provides smoother estimators with better asymptotic properties than histograms and frequency polygons. Nevertheless, in practice, the most widely used density estimator is still the histogram, mainly because its simplicity and availability in standard statistics packages (Simonoff, 1995).

Expressions (1) and (2) show that histograms and frequency polygons can be expressed as linear combinations of functions K_U and K_T properly translated and scaled. Observe that K_U is the density of a uniform distribution $U([-.5, .5])$, and K_T is the density of the convolution of two of these distributions. This fact suggests generalizing the classical estimators by considering linear combinations of densities K^m corresponding to the convolution of m distributions $U([-.5, .5])$. These basic densities are piecewise degree $(m-1)$ polynomial functions having $(m-2)$ continuous derivatives (i.e., they are *splines*; see e.g. Schumaker, 1981, Chap. 4). Our proposal accomplishes two goals. Firstly, due to its construction, the new estimators are smoother than the histogram and frequency polygon (note that the continuity of the frequency polygon, probably its most appealing advantage over the histogram, is achieved by the convolution of two uniform densities; more regularity is attained convolving more than two uniform random variables). Secondly, good density estimators should be obtained if the linear combination coefficients were chosen accordingly to a spcific criterion.

The rest of the paper is organized as follows. In Section 2 we present a wide class of density estimators, the *generalized histograms*, and we propose a criterion to obtain *optimal* estimators. Section 3 is devoted to prove that such optimal choice is feasible; moreover, the estimators for $m = 1, 2$ are compared with the classical histogram and frequency polygon. The asymptotic properties of the new estimators are studied in Section 4. Simulated data are used to test the proposed estimators in practice, and the results are summarized in Section 5. The proofs are deferred to the Appendix.

2 A Family of Simple Density Estimators

Let K^m be the density function of the sum of m independent distributions $U([-.5, .5])$. We consider the family of density estimators

$$\hat{f}^m_w(x) = \sum_{j \in Z} w_j \frac{1}{h} K^m \left(\frac{x - c_j}{h} \right) = \sum_{j \in Z} w_j K^m_h (x - c_j), \tag{3}$$

depending on $w \in I\!R^Z$, a sequence of real *weights* summing one. We call these estimators *generalized histograms.*

Let us note that the edge frequency polygon (Jones *et al.*, 1998) can be expressed as a *generalized histogram* (m=2) taking $w_j = (f_{j-1} + f_j)/2$ and centering the kernels K^2_h at b_j instead of at c_j. Histosplines defined in Boneva *et al.* (1971) could also be expressed as in (3) with $w_j = f_j$, but using a kernel that is not a density function: the deltaspline, a continuous piecewise quadratic function taking negative values in some intervals. A different estimator family would be obtained by substituting the kernel K^m in the definition (3) by an arbitrary kernel function with an associated bandwidth b not necessarily equal to the length h. For certain weights, this would include the binned kernel estimators: $f_B(x) = \Sigma_j f_j K_b(x - c_j)$, studied by Hall (1982) and Scott and Sheather (1985), and the prebinned kernel estimators (Jones, 1989): $f_{PB}(x) = \Sigma_j d(c_j) K_b(x - c_j)$, where $d(c_j)$ is some discretization of the sample based on B_j.

Our proposal is to improve the histogram-type estimators by means of the estimators $\hat{f}_w^m(x)$. The smoothness increases using K^m instead of K_U or K_T, and the weights in (3) may be chosen accordingly to a sensible criterion.

For the following discussion we can consider in the definition (3) a general kernel function K, denoting the associated estimator by $\hat{f}_w$. Let $K_h(u) = (1/h)K(u/h)$. It is easy to prove that under some mild conditions on K fulfilled by the kernels K^m (for instance, boundedness and monotonicity over $[0, \infty)$, we have that

$$\sum_{j \in Z} K_h(x - c_j) \leq K(0) + \sum_{j \in Z} K(j) \leq K(0) + \int_{I\!R} K(t)dt = K(0) + 1.$$

Thus, the sequence $K_h(x; c) = \{K_h(x - c_j)\}$ belongs to l^1, the set of real sequences $\{a_j\}$ with $\sum_{j \in Z} |a_j| < \infty$; since their elements are in $[0, 1]$, $K_h(x; c)$ it is also in l^2, the Hilbert space of sequences $\{a_j\}$ with $\sum_{j \in Z} a_j^2 < \infty$. Therefore, assuming $w \in l^2$, the estimator $\hat{f}_w(x)$ can be written as the inner product in l^2:

$$\hat{f}_w(x) =< w, K_h(x; c) > .$$

If $a, b \in l^2$, we also write $a^{\mathrm{T}}b$ for $< a, b >$.

We now deal with the election of the weight sequence w based on the sample $X_1, \ldots, X_n$. To evaluate the estimator $\hat{f}_w$, we take the integrated square error, ISE,

$$\mathrm{ISE}(\hat{f}_w) = \int (\hat{f}_w(x) - f(x))^2 dx =$$

$$\int \hat{f}_w^2(x)dx - 2\int f(x)\hat{f}_w(x)dx + \int f(x)^2 dx =$$

$$\int \hat{f}_w^2(x)dx - 2E(\hat{f}_w(X)) + \int f(x)^2 dx.$$

Estimating $E(\hat{f}_w(X))$ by $\sum_{i=1}^n \hat{f}_w(X_i)/n$, we obtain a *feasible* version of ISE,

$$\mathrm{FISE}(\hat{f}_w) = \int \hat{f}_w^2(x)dx - \frac{2}{n}\sum_{i=1}^n \hat{f}_w(X_i) + \int f(x)^2 dx.$$

Disregarding the term $\int f(x)^2dx$, we have the closeness measure:

$$\phi(w) = \int \hat{f}_w^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_w(X_i).$$

The following result characterizes the sequences minimizing $\phi(w)$.

Theorem 1 *Let $\bar{K}_h(X_1, \ldots, X_n; c)$ denote the sequence*

$$\hat{f}_h^K(c_j) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - c_j), j \in \mathrm{Z}.$$

The sequences of weights w minimizing $\phi(w) = FISE(\hat{f}_w)$ are the solutions of the linear system with infinite equations

$$Mw = h\bar{K}_h(X_1, \ldots, X_n; c), \tag{4}$$

where $M \in \mathbb{R}^{\mathrm{Z}\times\mathrm{Z}}$ has the generic element (k,l) equal to $\eta_{k,l} = \eta(|k-l|)$, being

$$\eta(s) = \int K(u)K(u+s)du. \tag{5}$$

Consequently, we are concerned with the solutions of the system (4). Observe that the infinite matrix M is symmetric and verifies that the element j in row k coincides with the element $(j+l)$ in row $(k+l)$, for all $l \in Z$ (i.e., each row in M is the result of shifting the row above one position); so M is an infinite symmetric Toeplitz matrix (i.e., its entries (i,j) and (k,l) are equal if $j-i = k-l$), and it is easy to see that rows in M are symmetric, in the sense that the elements $(j-l)$ and $(j+l)$ in row j are equal. Therefore, we have that $w = \{w_k\}_{k\in Z}$ verifies (4) if and only if

$$\sum_{l\in Z} \eta(l)w_{k+l} = h\hat{f}_K(c_k), \text{ for all } k \in Z. \tag{6}$$

From (5), we deduce that the difference equation (6) has finite order (i.e., the sum on the left hand side has only a finite number of terms) if and only if the kernel K has compact support.

In the sequel, we restrict ourselves to the study of the proposed density estimators $\hat{f}_w^m(x)$. Note that the kernels K^m have compact support. When a sequence of weights $\hat{w}$ minimizing $\phi(w)$ is used in (3), we call the resulting estimator *optimal generalized histogram*:

$$\hat{f}_{\hat{w}}^m(x) = \sum_{j\in Z} \hat{w}_j K_h^m(x - c_j).$$

The next section is devoted to prove the existence of such estimators.

		Values of s								
m	$s=$	-4	-3	-2	-1	0	1	2	3	4
1						1				
2					1	4	1			
3				1	26	66	26	1		
4			1	120	1191	2416	1191	120	1	
5		1	502	14608	88234	156190	88234	14608	502	1

Table 1: Values of $\eta(s)(2m-1)!$ for some values of m.

3 Optimal Generalized Histograms

Proving the existence of optimal generalized histogram requires to know more about the structure of the matrix M introduced in (4).

Proposition 1 *When the kernel K^m is used, M is a banded matrix with bandwidth equal to $(2m-1)$. The element (k,l) in M is $\eta_m(|k-l|)$, where $\eta_m(s)$ is*

$$\eta_m(s) = \frac{1}{(2m-1)!}\sum_{j=0}^{s+m}\binom{2m}{j}(-1)^j(s+m-j)^{2m-1}$$

for $s = -m+1, \ldots, -1, 0, 1, \ldots, m-1$, and it is equal to zero for any other integer value s. Moreover, the sum of the elements of any row (or column) in M is equal to one.

Table 1 shows the basic row of matrix M (multiplied by $(2m-1)!$) for some values of m.

Before dealing with the solution of system (4) for a generic m, we will obtain the optimal generalized histogram estimators for the particular cases $m = 1$ and $m = 2$. We will see that for $m = 1$ the optimal choice of the weights w leads us just to the usual histogram, but for $m = 2$ the resulting piecewise linear estimator is different to the frequency polygon.

3.1 Optimal estimator for $m = 1$ and $m = 2$.

For $m = 1$, it is easy to see that matrix M is the identity operator (M has ones in its diagonal and zeros outside), so the optimal weights are

$$\hat{w}_j = \frac{h}{n}\sum_{i=1}^{n} K_h^1(X_i - cj) = \frac{1}{n}\sum_{i=1}^{n} I_{[c_j-.5h, c_j+.5h]}(X_i) = f_j,$$

and the optimal generalized histogram for $m = 1$ is the classical histogram.

For $m = 2$ the elements of matrix M are given by

$$\eta_{k,l} = \begin{cases} 4/6 & \text{if } l = k \\ 1/6 & \text{if } |l-k| = 1 \\ 0 & \text{in other cases.} \end{cases}$$

It is easily checked that weights $w_j^P = f_j$ used to build the usual frequency polygon are not optimal. Let L_j be the generic left hand side term in the equation (4) when w is replaced by w^P, and let R_j be the corresponding right hand term; we have

$$L_j = \frac{1}{6}(f_{j-1} + 4f_j + f_{j+1}),$$

$$R_j = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|X_i - c_j|}{h}\right) I_{[c_j-h,c_j+h]}(X_i).$$

These expressions are different in general: for instance, if we move an observation X_i from $(c_{j-1} - h/2, c_{j-1})$ to $(c_{j-1}, c_{j-1} + h/2)$, the L_j value does not change, but R_j does. Thus, the frequency polygon is not the best piecewise linear density estimator in terms of ISE.

The computation of the inverse of the operator M can be handled as follows. Assume by now —see the next subsection— that the inverse operator of M exists and can be represented by an infinite symmetric Toeplitz matrix $N = M^{-1} = \{\nu_{k,l}\}_{k,l\in Z}$, with rows in l^2. Then,

$$\sum_{j\in Z}\eta_{k,j}\nu_{j,l} = \begin{cases} 1 & \text{if } l = k \\ 0 & \text{in other case.} \end{cases}$$

Therefore,

$$(1/6)\nu_{k-2,l} + (4/6)\nu_{k-1,l} + (1/6)\nu_{k,l} = 0,\ k \neq l+1,$$

which is a difference equation with the initial condition

$$(1/6)\nu_{k-1,k} + (4/6)\nu_{k,k} + (1/6)\nu_{k+1,k} = 1.$$

Since N is a Toeplitz symmetric matrix ($\nu_{l+k,l} = \nu_{l-k,l}$) with rows in l^2, the above equation can be solved and the solution sequence is

$$\nu_{k,l} = \sqrt{3}(-2 + \sqrt{3})^{|k-l|}.$$

Observe that the sum of the elements in each row or column of N is one. Moreover $\nu_{k,k+l}$ only depends on l, so, defining $a_l = \nu_{k,k+l}$, we have $a_l = a_{-l}$, and the optimal weights are:

$$\hat{w}_j = h\sum_{l=-\infty}^{\infty} a_l \hat{f}_{K^2}(c_{j+l}).$$

Thus, the weight of the box centered at c_j is a weighted mean of the values of the kernel density estimator at points c_{j+l}, where the weights are given by the columns of $N = M^{-1}$. The optimal estimator is

$$\hat{f}_{\hat{w}}(x) = h \sum_{j=[x/h-.5]}^{[x/h-.5]+1} \left\{ \sum_{l=-\infty}^{\infty} a_l \hat{f}_{K^2}(c_{j+l}) \right\} K_h^2(x - c_j).$$

Example 1. Figure 1 compares the classical frequency polygon with the optimal generalized histogram estimator corresponding to $m = 2$, piecewise linear. Two samples of size $n = 100$ were simulated. For the first one (upper panel), data follow a standard normal distribution, and the length was $h = .8$. The model used in the second case (lower panel) is a mixture of normal data:

$$X_i \sim .6N(0,1) + .4N(3, \sigma^2 = .64),$$

and $h = 1.2$.

3.2 Optimal estimation for an arbitrary m

Theorem 1 translates the minimization of $\phi(w)$ to the solution of the system (4), that can be written as a difference equation of order $(2m - 2)$ with constant coefficients:

$$\eta_m(m-1)w_{k-m+1} + \cdots \eta_m(1)w_{k-1} + \eta_m(0)w_k +$$
$$\eta_m(1)w_{k+1} + \cdots \eta_m(m-1)w_{k+m-1} = h\hat{f}_K(c_j), \forall k \in Z.$$

Let P_m be the characteristic polynomial of this equation:

$$P_m(w) = p_{m-1} + p_{m-2}w + \cdots + p_0 w^{m-1} + \cdots + p_{m-2}w^{2m-3} + p_{m-1}w^{2m-2},$$

where $p_k = \eta_m(k)$ for $k = 0, \ldots, m-1$. Some properties of P_m can be easily derived from Proposition 1, for instance: $p_k \leq p_{k-1}$, $k = 1, \ldots, m-1$, $P_m(1) = 1$, the real roots of P_m are negative, and if v is a real root of P_m then $1/v$ is also a root of P_m. A numerical study developed with Mathematica (some of its results are summarized in Table 2) shows that all roots of P_m are real and different for $m = 2, \ldots, 43$, and that some complex roots appear for $m \geq 44$. Apparently, these complex roots do not seem to arise due to numerical instability. In practice, only small values of m will be used; therefore, the assumption of real and different roots for P_m is not restrictive. The following result gives the solutions of the equations (6) under such condition.

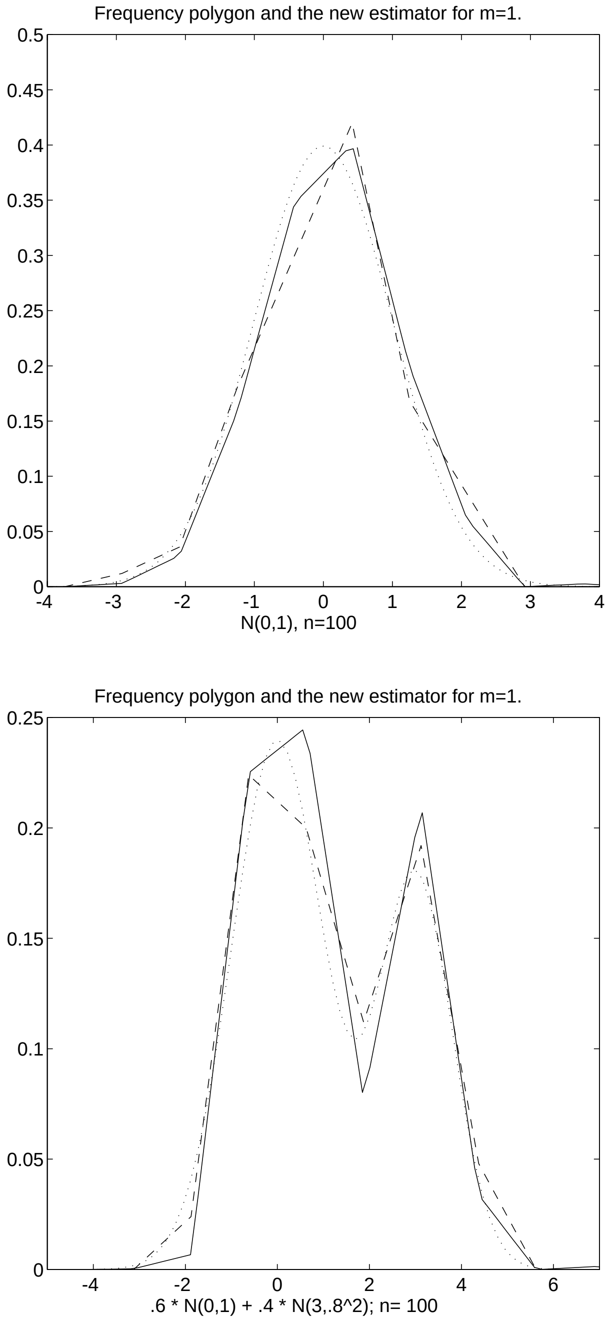

Figure 1: Frequency polygon (dashed line) and the new estimator for $m = 2$ (solid line). The true density is represented by a dotted curve.

m	Absolute value of some roots of P_m							
2	.2679							
3	.4306	.0431						
4	.5353	.1226	.9149 E-2					
5	.6080	.2018	.4322 E-1	.2121 E-2				
6	.6613	.2722	.8976 E-1	.1667 E-1	.5106 E-3			
7	.7019	.3331	.1389	.4321 E-1	.6738 E-2	.1251 E-3		
8	.7339	.3856	.1865	.7591 E-1	.2175 E-1	.2801 E-2	.3094 E-4	
9	.7597	.4309	.2311	.1108	.4321 E-1	.1126 E-1	.1186 E-2	.7688 E-5

Table 2: Absolute value of the roots of P_m polynomials for some values of m. Only absolute values of roots $v \in (-1, 0)$ are shown.

Proposition 2 *Let us assume that P_m has real roots with multiplicity one. Let us call $v_1, \ldots, v_{m-1}$ the roots belonging to $(-1, 0)$. Define the sequences $w^{(j)}$ with $w_k^{(j)} = v_j^{|k|}, k \in \mathrm{Z}$. Let e^0 be the sequence in l^2 having $e_k^0 = 0$ for $k \neq 0$ and $e_0^0 = 1$. Consider the difference equation associated to the modified version of system (4)*

$$Mw = e^0.$$

Then, this equation has a unique solution ν in l^2, that have the form

$$\nu = \sum_{j=1}^{m-1} \alpha_j w^{(j)},$$

where the real numbers α_j, $j = 1, \ldots, m-1$, are the solutions to the $(2m-1)$ linear system

$$\left\{ \begin{array}{l} \sum_{l=-m+1}^{m-1} \left(\eta(|l|) \sum_{j=1}^{m-1} \alpha_j v_j^{|l|} \right) = 1, \\ \sum_{l=-m+1}^{m-1} \left(\eta(|l|) \sum_{j=1}^{m-1} \alpha_j v_j^{|k+l|} \right) = 0, \ k = -m+2, \ldots, -1, 1, \ldots, m-2 \end{array} \right.$$

Consequently, the linear operator characterized by the infinite matrix M is invertible and, therefore, the system (4) has a unique solution.

Proposition 3 *Let us assume that P_m has all its roots real with multiplicity one. Consider kernel $K = K^m$. Let $M \colon l^2 \longrightarrow l^2$ be the linear operator characterized by the infinite symmetric Toeplitz banded matrix M defined in Theorem 1.*

(i) There exists a linear operator $N \colon l^2 \longrightarrow l^2$ such that both $M \circ N$ and $N \circ M$ are equal to the identity operator on l^2. Denote N by M^{-1}. The operator M^{-1} is given by an infinite symmetric Toeplitz matrix, say M^{-1}, with generic row (and column) ν defined in Proposition 2. The sequence $\nu \in l^2$ and the sum of its elements is equal to one.

(ii) Let ν^j be the sequence $\nu^j_k = \nu_{k-j}$. The vector

$$\hat{w} = hM^{-1}\bar{K}_h(X_1, \ldots, X_n; c) = h \sum_{j \in \mathrm{Z}} \hat{f}(c_j)\nu^j$$

is the unique solution in l^2 to the system (4). The sum of the elements of this optimal sequence, $\hat{w}$, is equal to one.

(iii) The optimal density estimator,

$$\hat{f}^m_{\hat{w}}(x) = \hat{w}^{\mathrm{T}} K^m_h(x; c),$$

can be written as

$$\hat{f}^m_{\hat{w}}(x) = l(x)^{\mathrm{T}} \bar{K}_h(X_1, \ldots, X_n; c), \tag{7}$$

a linear combination of kernel density estimation at points c_j with coefficient vector given by

$$l(x) = hM^{-1}K^m_h(x; c). \tag{8}$$

Example 1 (cont.). Figure 2 completes the Figure 1 illustrating the behaviour of optimal generalized histogram estimators corresponding for $m = 1, 2, 3$.

4 Asymptotic Behaviour

The next result gives the asymptotic properties of the optimal generalized histogram density estimators.

Theorem 2 *Assume that the density function f has an absolutely continuous m-th derivative $f^{(m)}$. The mean integrate squared error of the estimator $\hat{f}^m_{\hat{w}}$ is*

$$MISE(\hat{f}^m_{\hat{w}}) = \frac{1}{nh} + \frac{h^{2m}}{(m!)^2} S^2_M(K^m) R(f^{(m)}) + O\left(\frac{1}{n}\right) + o\left(h^{2m}\right),$$

where $S^2_M(K^m) = \int_{-.5}^{.5} S^2_m(t)dt$, $S_m(t) = \int K^m(t; \mathrm{Z})^{\mathrm{T}} M^{-1} K^m(u; \mathrm{Z})(u-t)^m du$, $K^m(t; \mathrm{Z})$ is the sequence with generic term $K^m(t-k)$, $k \in Z$, and $R(g) = \int g(y)^2 dy$.

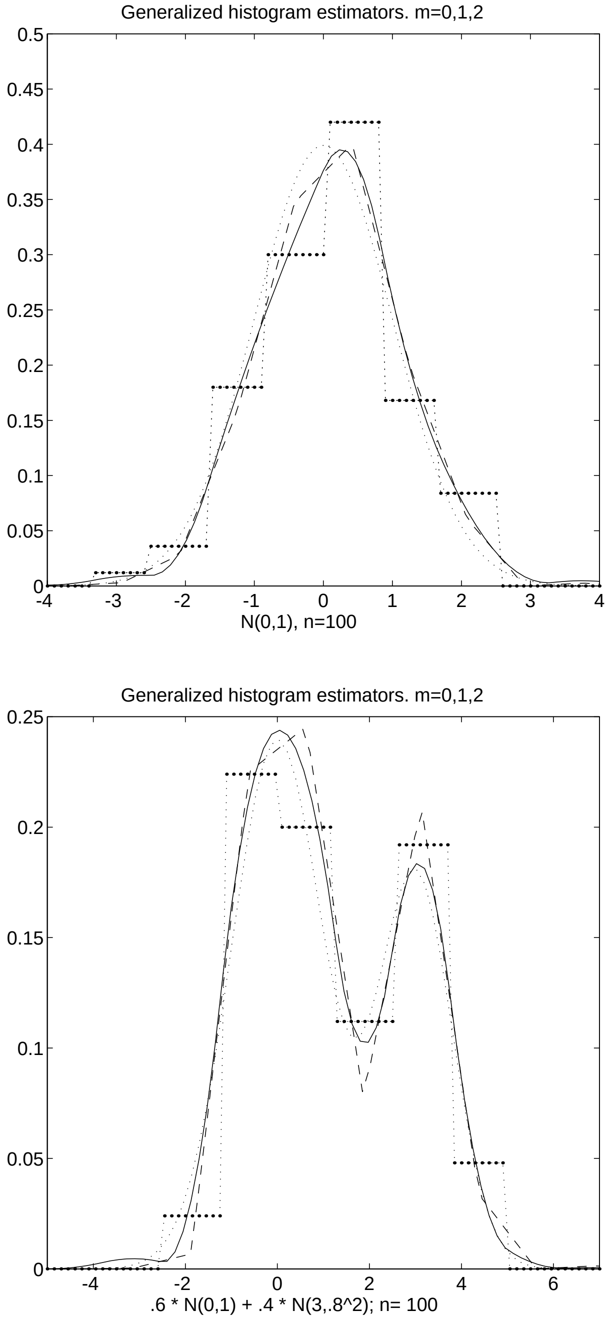

Figure 2: Optimal generalized histogram estimators for m equal to 1 (histogram), 2 (polygonal line) and 3 (solid line). The true density is represented by a dotted curve.

The minimizer of asymptotic MISE, AMISE, is

$$h_m = \left(\frac{(m!)^2}{2m S_M^2(K^m) R(f^{(m)})} \right)^{\frac{1}{2m+1}} n^{-\frac{1}{2m+1}} = c_{h,m} (R(f^{(m)})n)^{-\frac{1}{2m+1}},$$

and the minimal AMISE, AMISE$_m$ is

$$\frac{2m+1}{2m} \left(\frac{2m S_M^2(K^m) R(f^{(m)})}{(m!)^2} \right)^{\frac{1}{2m+1}} n^{-\frac{2m}{2m+1}} = c_{A,m} R(f^{(m)})^{\frac{1}{2m+1}} n^{-\frac{2m}{2m+1}}.$$

Table 3 shows the values of $S_M^2(K^m)$ and the constant multiplicative factors, depending on m, for the optimal bandwidth h_m and the minimum AMISE$_m$. In the same table, the constants for AMISE$_m$ are compared with that of some known estimators with the same rate of convergence: the frequency polygon (FP) and the Epanechnikov kernel estimator (EpK) for $m = 2$, and the order 4 polynomial kernel (PK$_4$) for $m = 4$ (see Scott, 1992, p. 134). Based on these asymptotic results, and as expected, the optimal generalized frequency polygon, $\hat{f}^2_{\hat{w}}$, performs better than the FP estimator, and very similar to EpK estimator. Also, the optimal generalized histogram estimator for $m = 4$ behaves better than the PK$_4$ estimator.

The values in Table 3 allow quick and simple rules for selecting the bandwidth replacing f by a reference density. For instance, using the gaussian density as reference we obtain the results in Table 4; in this case, $h_m = c^{\phi}_{h,m} \sigma n^{-1/(2m+1)}$ and AMISE$_m = c^{\phi}_{A,m} \sigma^{-1} n^{-2m/(2m+1)}$.

Let us note that, arguing as Jones *et al.* (1998) do in relation with the practice of their edge frequency polygon, we could use methods of bandwidth selection in kernel density estimation for selecting the length h of the optimal generalized histograms. For instance, the constants appearing in Table 3 indicate that, for $m = 2$, we may take a length equal to $(2.8832/1.7188) = 1.6774$ times the bandwidth we would use for the Epanechnikov kernel estimator.

5 Some simulation results and conclusions

We study the behaviour of the optimal generalized histograms for finite samples and their agreement with the asymptotic properties studied above. A Fortran-90 program was written for this purpose.

Samples were generated from two populations: the standard normal, and the mixture of normal populations used in previous examples: $N(0, 1)$ with probability .6 and $N(3, \sigma^2 = .64)$ with probability .4. For each of them, samples of sizes $n = 50$, 200 and 1000 were taken. We consider the seven estimators included in Tables 3 and 4: optimal generalized histograms for

m	$S_M^2(K^m)$	$c_{h,m}$	$c_{A,m}$
1	.08333	1.8171	.8255
2	.005556	2.8252	.4424
3	.001192	3.3794	.3452
4	.0006136	3.6582	.3075
Constants for some comparable estimators			
FP		1.5784	.5280
EpK		1.7188	.4364
PK_4		3.2431	.4501

Table 3: $S_M^2(K^m)$ and constants giving h_m and AMISE_m for $m = 1, 2, 3, 4$. Constants for some comparable estimators when $m = 2$ (FP and EpK) or $m = 4$ (PK_4).

m	c_h^ϕ	c_A^ϕ
1	3.4908	.4297
2	3.8545	.3243
3	3.7013	.3152
4	3.4163	.3293
Constants for some comparable estimators		
FP	2.15	.387
EpK	2.35	.320
PK_4	3.03	.482

Table 4: Constants giving h_m and AMISE_m in the $N(\mu, \sigma)$ case. Constants for some comparable estimators when $m = 2$ (FP and EpK) or $m = 4$ (PK_4).

$m = 1, 2, 3, 4$, noted by GH(m), frequency polygon (FP), Epanechnikov kernel estimator (EpK) and the order 4 polynomial kernel (PK$_4$). The density function is estimated in a grid of 400 equispaced points covering the interval $[-4, 4]$ for the standard normal and $[-3, 5]$ for the mixture case. For each combination of population and sample size, 2000 simulations were done. Numerical integration based on Simpson's method was used to compute MISEs values.

Bandwidth parameters h are determined by the normal reference rule, using the values of constant c_h^{ϕ} shown in Table 4. This is the asymptotic optimal choice for the normal case. Nevertheless, as could be expected, this rule leads to oversmoothing when data are generated from the mixture of normals population. Therefore, for non normal simulated data we use also the ad hoc rule: *choose h as 1/2 times the quantity given by the normal reference rule.*

The values of the mean integrated squared error (MISE) are shown in Table 5. The proportion of MISE due to the integrated squared bias (the rest is due to the integrated variance) is indicated in brackets. This information is useful for detecting oversmoothing.

We start our comments by the rows of Table 5 relative to standard normal data. The generalized version GH(2) of the frequency polygon gives always better results than the classical version FP. For $n = 200$ and 1000, the EpK estimator beats slightly to GH(2), as it happens asymptotically (see Table 4). The performances of GH(4) and PK$_4$ also agree with the asymptotic results: GH(4) gives lower MISEs values than PK$_4$. On the other hand, GH(3) performs better than EpK and PK$_4$. Asymptotic results do not allow to compare these three estimators.

The second set of rows in Table 5 corresponds to the mixture population and the choice of bandwidth based on the normal reference rule. In these cases classical estimators are in general preferred to optimal generalized histograms, specially for $n = 1000$. If we look at proportions of MISE due to estimators bias, we deduce that oversmoothing occurs in all the estimators, being more important for GE(m), $m = 2, 3, 4$. Therefore we conclude, as should be expected, that the normal reference rule for choosing the bandwidth h is not appropriate for mixture populations.

In order to reduce oversmoothing, we implemented the above ad hoc rule. The results of this procedure form the last part of Table 5. All the estimators perform better now than they previously did. The MISEs values for GH(3) (even for $m = 3, 4$) are now comparable with other estimators.

As a general conclusion, we point out three positive aspects of generalized histograms. Firstly, their asymptotic properties overcome those of histograms and some other known density estimators. Secondly, they are

Population		GH(m)				FP	EpK	PK$_4$
and h choice	n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$(m = 2)$	$(m = 2)$	$(m = 4)$
	50	.0261	.0070	.0045	.0055	.0119	.0092	.0088
		(.16)	(.18)	(.08)	(.27)	(.18)	(.21)	(.05)
$N(0,1)$	200	.0111	.0036	.0015	.0017	.0042	.0033	.0027
		(.18)	(.35)	(.03)	(.23)	(.20)	(.22)	(.07)
$h = c_h^\phi n^{-1/(2m+1)}$	1000	.0041	.0012	.0006	.0003	.0013	.0010	.0007
		(.20)	(.35)	(.33)	(.02)	(.22)	(.24)	(.09)
	50	.0320	.0252	.0325	.0326	.0234	.0296	.0284
		(.94)	(.99)	(.99)	(.99)	(.98)	(.99)	(.99)
$.6N(0,1) + .4N(3,.8^2)$	200	.0158	.0201	.0150	.0170	.0136	.0128	.0172
$\sigma^2 = 3.016$		(.80)	(.97)	(.99)	(.99)	(.92)	(.94)	(.96)
$h = c_h^\phi \sigma n^{-1/(2m+1)}$	1000	.0035	.0079	.0147	.0162	.0022	.0019	.0032
		(.41)	(.94)	(.98)	(.99)	(.74)	(.77)	(.88)
	50	.0247	.0151	.0147	.0145	.0136	.0110	.0120
		(.05)	(.37)	(.51)	(.55)	(.08)	(.10)	(.03)
$.6N(0,1) + .4N(3,.8^2)$	200	.0100	.0047	.0069	.0073	.0047	.0037	.0035
$\sigma^2 = 3.016$		(.06)	(.25)	(.65)	(.71)	(.09)	(.10)	(.03)
$h = .5c_h^\phi \sigma n^{-1/(2m+1)}$	1000	.0036	.0013	.0016	.0028	.0014	.0011	.0009
		(.07)	(.17)	(.56)	(.80)	(.09)	(.10)	(.04)

Table 5: Simulation results. MISE values for seven density estimators and some combinations of populations and sample sizes. Proportion of MISE due to bias is given in brackets.

less computationally demanding than kernel estimators, as equation 3 indicates and simulations confirm. Finally, optimal generalized histograms seem to give very good results in practice, when h is appropriately chosen.

Future work should be addressed in bandwidth selection other than the normal reference rule. In this regard, the solution pointed out at the end of Section 4, taken from a proposal by Jones *et al.* (1998), and based on adapting with known constants any reasonable bandwidth selector in kernel density estimation, could be a good practical solution.

Appendix: Proofs

We begin this section recalling two results on kernels K^m. The first lemma (see e.g. Feller, 1971, p. 28) gives an explicit expression for K^m, and the second one (see e.g. Schumaker, Secs. 4.3-4.4) establishes the properties of the family of splines derived from K^m.

Lemma 1 *The function K^m is given by*

$$K^m(x) = \frac{1}{(m-1)!} \sum_{j=0}^{[x+\frac{m}{2}]} \binom{m}{j} (-1)^j \left(x + \frac{m}{2} - j\right)^{m-1} I_{[-m/2,m/2]}(x).$$

Lemma 2 *The set $\{K_h^m(x-c_j) = (1/h)K^m((x-c_j)/h), j \in \mathrm{Z}\}$ is a basis of the space of piecewise polynomial functions with $(m-1)$ continuous derivatives (*splines*) having knots at points $\{c_j\}_j$.*

Theorem 1: Noting $\hat{f}_w(x) = w^{\mathrm{T}} K_h(x;c)$, we have

$$\int \hat{f}_w^2(x)dx = \int w^{\mathrm{T}} K_h(x;c) K_h^{\mathrm{T}}(x;c) w dx = w^{\mathrm{T}} A(c) w,$$

where $A(c) = \{a_{k,l}(c)\}_{k,l \in Z} \in I\!R^{Z \times Z}$, and

$$a_{k,l}(c) = \frac{1}{h^2} \int K\left(\frac{x-c_k}{h}\right) K\left(\frac{x-c_l}{h}\right) dx = \frac{1}{h} \int K(u) K\left(u - (l-k)\right)) du.$$

Thus, the elements $a_{k,l}(c)$ of matrix $A(c)$ only depend on kernel K and are equal to $h^{-1}\eta_{k,l} = h^{-1}\eta(|k-l|)$, where $\eta(s)$ is given by (5). Recalling the definition of M, we have that

$$\int \hat{f}_w^2(x)dx = \frac{1}{h} w^{\mathrm{T}} M w.$$

Since,

$$\frac{1}{n} \sum_{i=1}^{n} \hat{f}_w(X_i) = \frac{1}{n} \sum_{i=1}^{n} w^{\mathrm{T}} K_h(X_i;c) = w^{\mathrm{T}} \bar{K}_h(X_1, \ldots, X_n; c),$$

where $\bar{K}_h(X_1, \ldots, X_n; c)$ is the sequence with j-th term equal to $(\sum_{i=1}^n K_h(X_i - c_j))/n$, $j \in Z$, we have

$$\phi(w) = \frac{1}{h} w^{\mathrm{T}} M w - 2 w^{\mathrm{T}} \bar{K}_h(X_1, \ldots, X_n; c).$$

Observe that for any pair of weight sequences w and w_o,

$$\phi(w) = \phi(w_o + (w - w_o)) =$$

$$\phi(w_o) + \frac{1}{h}(w - w_o)^{\mathrm{T}} M(w - w_o) - \frac{2}{h}(w - w_o)^{\mathrm{T}}(h\bar{K}_h(X_1, \ldots, X_n; c) - Mw_o).$$

Hence, for any solution w_0 to (4), we conclude $\phi(w_o) \leq \phi(w)$ for all w. □

Proposition 1: The first part follows directly from Lemma 1, and from the fact that $\eta_m(x)$ is the convolution of two densities K^m, so $\eta_m(x)$ equals $K^{2m}(x)$, the density function of the sum of $2m$ independents $U([-.5, .5])$.

Let f_X be now the density of a continuous distribution with compact support. The density, f_Y, of the convolution of this distribution with the $U([-.5, .5])$ is

$$f_Y(y) = \int_{-.5}^{.5} f_X(y - u)du = \int_{y-.5}^{y+.5} f_X(x)dx,$$

then

$$\sum_{j \in Z} f_Y(j) = \sum_{j \in Z} \int_{j-.5}^{j+.5} f_X(x)dx = \int f_X(x)dx = 1, \tag{9}$$

proving the second part. □

Proposition 2: The system $Mw = e^0$ is equivalent to the difference equation

$$\sum_{l=-m+1}^{m-1} \eta(|l|)w_{k+l} = \delta_0(k),\ k \in Z, \tag{10}$$

where $\delta_0(k)$ is 0 for $k \neq 0$ and $\delta_0(0) = 1$.

First we will prove that there exists a solution of (10), and then we will deal with uniqueness. Consider the sequence

$$\nu = \sum_{j=1}^{m-1} \alpha_j w^{(j)}.$$

For any coefficients α_j, ν verifies (10) for $k \geq m - 1$ because for these values of k, ν is a linear combination of sequences $\{v_j^k\}_k$, and the standard difference equation theory assures that these sequences solve the equation

$$\sum_{l=-m+1}^{m-1} \eta(|l|)w_{k+l} = 0,\ k \in Z.$$

Similarly, ν verifies (10) for $k \leq -m+1$ and any choice of α_j, $j = 1, \ldots, m-1$. So, we only need to prove that the $(2m + 1)$ linear system

$$\sum_{l=-m+1}^{m-1} \left\{ \eta(|l|) \sum_{j=1}^{m-1} \alpha_j v_j^{|k+l|} \right\} = \delta_0(k),\ k = -m + 2, \ldots, -1, 0, 1, \ldots, m - 2,$$

has a unique solution $(\alpha_1, \ldots, \alpha_{m-1})$. Observe that the equation corresponding to $k \in \{1, \ldots, m-1\}$ coincides with that corresponding to $-k$, because of the symmetry of η. So we have in fact a linear system with $(m-1)$ unknowns and $(m-1)$ equations. Let A be the coefficients matrix:

$$A = \begin{pmatrix} \sum_{l=-m+1}^{m-1} \eta(|l|) v_1^{|l|} & \ldots & \sum_{l=-m+1}^{m-1} \eta(|l|) v_{m-1}^{|l|} \\ \vdots & \ddots & \vdots \\ \sum_{l=-m+1}^{m-1} \eta(|l|) v_1^{|m+l-2|} & \ldots & \sum_{l=-m+1}^{m-1} \eta(|l|) v_{m-1}^{|m+l-2|} \end{pmatrix}.$$

We know that

$$b_{ij} = \sum_{l=-m+1}^{m-1} \eta(|l|) v_j^{l+i} = 0, \; i = 0, \ldots, m-2, \; j = 1, \ldots m-1,$$

by standard difference equations results. So $B = (b_{ij})$ is the null $(m-1) \times (m-1)$ matrix and, writing $A = A - B$, we have

$$A = \begin{pmatrix} \sum_{l=1}^{m-1} \eta(l) \left(v_1^{m-l} - \frac{1}{v_1^{m-l}} \right) & \ldots & \sum_{l=1}^{m-1} \eta(l) \left(v_{m-1}^{m-l} - \frac{1}{v_{m-1}^{m-l}} \right) \\ \vdots & \ddots & \vdots \\ \sum_{l=m-2}^{m-1} \eta(l) \left(v_1^{m-l} - \frac{1}{v_1^{m-l}} \right) & \ldots & \sum_{l=m-2}^{m-1} \eta(l) \left(v_{m-1}^{m-l} - \frac{1}{v_{m-1}^{m-l}} \right) \\ \eta(m-1) \left(v_1 - \frac{1}{v_1} \right) & \ldots & \eta(m-1) \left(v_{m-1} - \frac{1}{v_{m-1}} \right) \end{pmatrix}.$$

A is non singular if and only if matrix $\tilde{A}_m$ is also non singular, where $\tilde{A}_m$ is obtained subtracting to each row the preceding one, dividing the resulting j-th row by $\eta(j)$, $j = 1, \ldots, m-1$, and reversing the order of rows; that is,

$$\tilde{A}_m = \begin{pmatrix} v_1 - \frac{1}{v_1} & \ldots & v_{m-1} - \frac{1}{v_{m-1}} \\ \vdots & \ddots & \vdots \\ v_1^{m-1} - \frac{1}{v_1^{m-1}} & \ldots & v_{m-1}^{m-1} - \frac{1}{v_{m-1}^{m-1}} \end{pmatrix}$$

Let d_m be the determinant of $\tilde{A}_m$. We will prove that

$$d_m = \prod_{j=1}^{m-1} \left(v_j - \frac{1}{v_j} \right) \cdot \prod_{i,j=1; i>j}^{m-1} \left(v_i + \frac{1}{v_i} - v_j - \frac{1}{v_j} \right), \tag{11}$$

for all $m \geq 2$, by induction on m. It is easy to check the proposed expression for $m = 2$ and $m = 3$. Let us assume that it is also true for $\tilde{A}_m$. Note that

$$\left(v_k^l - \frac{1}{v_k^l} \right) = \left(v_k^{l-1} - \frac{1}{v_k^{l-1}} \right) \left(v_k + \frac{1}{v_k} \right) + \left(v_k^{l-2} - \frac{1}{v_k^{l-2}} \right)$$

for all $l \geq 2$. So

$$\left(v_k^l - \frac{1}{v_k^l}\right) - \left\{\left(v_k^{l-1} - \frac{1}{v_k^{l-1}}\right)\left(v_1 + \frac{1}{v_1}\right) + \left(v_k^{l-2} - \frac{1}{v_k^{l-2}}\right)\right\} =$$

$$\left(v_k^{l-1} - \frac{1}{v_k^{l-1}}\right)\left(v_k + \frac{1}{v_k} - v_1 - \frac{1}{v_1}\right)$$

We transform $\tilde{A}_m$ as follows: for $k = m-1, \ldots, 3$, we add to the row k the row $(k-1)$ multiplied by $-(v_1 + 1/v_1)$, plus the row $(k-2)$ multiplied by -1; we also subtract the first row to the second one. Thus, d_m is equal to

$$\det\begin{pmatrix} v_1 - \frac{1}{v_1} & v_2 - \frac{1}{v_2} & \cdots & v_{m-1} - \frac{1}{v_{m-1}} \\ 0 & \left(v_2 - \frac{1}{v_2}\right)\left(v_2 + \frac{1}{v_2} - v_1 - \frac{1}{v_1}\right) & \cdots & \left(v_{m-1} - \frac{1}{v_{m-1}}\right)\left(v_{m-1} + \frac{1}{v_{m-1}} - v_1 - \frac{1}{v_1}\right) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \left(v_2^{m-2} - \frac{1}{v_2^{m-2}}\right)\left(v_2 + \frac{1}{v_2} - v_1 - \frac{1}{v_1}\right) & \cdots & \left(v_{m-1}^{m-2} - \frac{1}{v_{m-1}^{m-2}}\right)\left(v_{m-1} + \frac{1}{v_{m-1}} - v_1 - \frac{1}{v_1}\right) \end{pmatrix} =$$

$$\left(v_1 - \frac{1}{v_1}\right)\prod_{j=2}^{m-1}\left(v_j + \frac{1}{v_j} - v_1 - \frac{1}{v_1}\right) \times$$

$$\det\begin{pmatrix} v_2 - \frac{1}{v_2} & \cdots & v_{m-1} - \frac{1}{v_{m-1}} \\ \vdots & \ddots & \vdots \\ v_2^{m-1} - \frac{1}{v_2^{m-1}} & \cdots & v_{m-1}^{m-1} - \frac{1}{v_{m-1}^{m-1}} \end{pmatrix} =$$

(by the induction hypothesis)

$$\left(v_1 - \frac{1}{v_1}\right)\prod_{j=2}^{m-1}\left(v_j + \frac{1}{v_j} - v_1 - \frac{1}{v_1}\right) \cdot \prod_{j=2}^{m-1}\left(v_j - \frac{1}{v_j}\right) \cdot \prod_{i,j=2; i>j}^{m-1}\left(v_i + \frac{1}{v_i} - v_j - \frac{1}{v_j}\right) =$$

$$\prod_{j=1}^{m-1}\left(v_j - \frac{1}{v_j}\right) \cdot \prod_{i,j=1; i>j}^{m-1}\left(v_i + \frac{1}{v_i} - v_j - \frac{1}{v_j}\right),$$

proving (11). So we conclude that matrix A has rank equal to $(m-1)$ because d_m is different from 0 if and only if all the roots of P_m are different. Then, the sequence ν is uniquely determined.

Finally, we will show that ν is the only sequence in l^2 that solves the difference equation (10). Observe first that solutions of the homogeneous version of (10) are not in l^2, because they should be linear combinations of $\{v_j^k\}_{k \in Z}$, that are not in l^2. Moreover, if $\tilde{\nu}$ is another sequence in l^2 solving (10), the sequence $\gamma = \nu - \tilde{\nu}$ belongs to l^2 and it is a solution of the homogeneous difference equation, so γ is a linear combination of sequences

$\{v_j^k\}_{k \in Z}$. Therefore, $\gamma = 0$ and $\tilde{\nu} = \nu$, what proves the uniqueness of ν. $\square$

Proposition 3: The former parts of (i) follow directly from Proposition 2, defining the infinite matrix N as that having the 0-th column equal to ν and the column j equal to ν shifted j positions, for all $j \in Z$. Observe also that $\nu_{-k} = \nu_k$, because the definition of ν_k only depends on $|k|$. Then N is a Toeplitz symmetric matrix. Also, Proposition 2 implies $\nu \in l^2$. To see that $\sum_{j \in Z} \nu_j = 1$, we have to define some new elements. For $h \in Z$, let $1_{[h]}$ be the sequence in l^2 having elements equal to 1 in positions $-|h|, \ldots, |h|$, and equal to 0 otherwise, and let $1_{\{h\}}$ be the sequence that has a one in position h and zeros otherwise. Let be $S_h = \nu^{\mathrm{T}} 1_{[h]}$. Our goal is to prove that

$$\lim_{h \to \infty} S_h = 1.$$

Observe that, for $h \geq 0$,

$$M 1_{[h+m-1]} = 1_{[h]} + \sum_{j=1}^{m-1} \sum_{l=1}^{j} \eta(l) \left(I_{\{-h-l\}} + I_{\{h+l\}} \right).$$

Multiplying both sides by N, and equalling the element at position 0 of the obtained sequences, we have

$$1 = S_h + \sum_{j=1}^{m-1} \sum_{l=1}^{j} 2\eta(l) \nu_{h+l}.$$

Therefore,

$$|S_h - 1| \leq \sum_{j=1}^{m-1} \sum_{l=1}^{j} 2\eta(l) |\nu_{h+l}| \leq$$

(because $\eta(l) \leq 1$)

$$\leq (m-1)m \max_{l=h,\ldots,h+m-1} |\nu_{h+l}| \leq (m-1)m \max_{l \geq h} |\nu_{h+l}|$$

and this last expression goes to zero as h goes to infinity because ν belongs to l^2, so $|\nu_h|$ converges to zero.

The first part of (ii) is the result of applying operator $N = M^{-1}$ to both sides of (4). Note that only a finite number of c_j have kernel density estimation $\hat{f}_K(c_j)$ different from 0 because the used kernel $K = K^m$ has compact support. Then, the resulting sequence $\hat{w}$ is well defined and belongs to l^2. To prove the last part, observe that

$$\sum_{j \in Z} \hat{w}_j = h \sum_{j \in Z} \sum_{l \in Z_S} \nu_{j,j-l} \hat{f}_K(c_{l-j}) =$$

(the sum on l is over a finite subset of Z denoted by Z_S)

$$h \sum_{l \in Z_S} \hat{f}_K(c_l) \sum_{j \in Z} \nu_{j,l} = h \sum_{l \in Z_S} \hat{f}_K(c_l) =$$

$$h \frac{1}{nh} \sum_{l \in Z_S} \sum_{i=1}^{n} K\left(\frac{c_l - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l \in Z_S} K(l),$$

and, recalling (9), this is equal to

$$\frac{1}{n} \sum_{i=1}^{n} 1 = 1.$$

Finally, (iii) follows directly. □

Theorem 2: The proof of this result is divided into two parts, the first one refers to the estimator bias, and the second one deals with the variance.
Estimator bias.
For simplicity, we write K instead of K^m. From (7),

$$E(\hat{f}_{\hat{w}}(y)) = l(y)^{\mathrm{T}} E(\bar{K}_h(X_1, \ldots, X_n; c)) = l(y)^{\mathrm{T}} \int K_h(z; c) f(z) dz,$$

where $\int K_h(z;c)f(z)dz$ is the element in $\mathbb{R}^Z$ with j-th term equal to

$$\int \frac{1}{h} K\left(\frac{z - c_j}{h}\right) f(z) dz.$$

Firstly, we prove that the new estimator is unbiased when the density is a piecewise polynomial function with degree lower than or equal to $(m-1)$. Let g be that polynomial. By Lemma 2, there exist $\lambda_k \in \mathbb{R}$, $k \in Z$, such that

$$g(z) = \sum_k \lambda_k K_h(z - c_k) = \lambda^{\mathrm{T}} K_h(z; c),$$

and, therefore, recalling the definition of $l(y)$ in equation (8),

$$E(\hat{g}_w(y)) = l(y)^{\mathrm{T}} \int K_h(z; c) g(z) dz = l(y)^{\mathrm{T}} \int K_h(z; c) K_h(z; c)^{\mathrm{T}} \lambda dz =$$

$$= \frac{1}{h} l(y)^{\mathrm{T}} M \lambda = (K_h(y; c))^{\mathrm{T}} \lambda = g(y).$$

As we have, for all $z \in \mathbb{R}$,

$$f(z) = f(y) + (z - y) f^{(1)}(y) + \ldots + f^{(m-1)}(y) \frac{(z - y)^{m-1}}{(m-1)!} +$$

$$f^{(m)}(y)\frac{(z-y)^m}{m!}+o((z-y)^m),$$

the preceding result gives

$$\mathrm{Bias}(\hat{f}_{\hat{w}}(y)) = E(\hat{f}_{\hat{w}}(y)) - f(y) =$$

$$l(y)^{\mathrm{T}}\int K_h(z;c)\left(f^{(m)}(y)\frac{(z-y)^m}{m!}+o((z-y)^m)\right)dz =$$

$$h\frac{f^{(m)}(y)}{m!}(K_h(y;c))^{\mathrm{T}}M^{-1}\int K_h(z;c)(z-y)^m dz +$$

$$l(y)^{\mathrm{T}}\int K_h(z;c)o((z-y)^m)dz. \qquad (12)$$

Let us assume that $y \in B_l$ and put $t = (y-c_l)/h \in [-.5,.5)$. For an arbitrary c_j we have that $y-c_j = y-c_l-(c_j-c_l) = ht-h(j-l)$, and then $(y-c_j)/h = t-(j-l)$. So the generic element of sequence $hK_h(y;c)$ is $K(t-(j-l)) = K((t+l)-j)$, $j \in Z$, and, substituting $u = (z-c_l)/h$, the first adding term in (12) is

$$\frac{f^{(m)}(y)}{m!}(K(t+l;Z))^{\mathrm{T}}M^{-1}\int K_h(z;c)(z-y)^m dz =$$

$$h^m\frac{f^{(m)}(y)}{m!}\int (K(t+l;Z))^{\mathrm{T}}M^{-1}K(u+l;Z)(u-t)^m du =$$

$$h^m\frac{f^{(m)}(y)}{m!}\int (K(t;Z))^{\mathrm{T}}M^{-1}K(u;Z)(u-t)^m du = h^m\frac{f^{(m)}(y)}{m!}S_m(t);$$

the second equality follows from the fact that $N = M^{-1}$ is a Toeplitz matrix: $N_{i,j} = N_{i-l,j-l}$. The second adding term in (12) is trivially $o(h^m)$, so we have for $y \in B_l$,

$$\mathrm{Bias}(\hat{f}_{\hat{w}}(y)) = h^m\frac{f^{(m)}(y)}{m!}S_m\left(\frac{y-c_l}{h}\right)+o(h^m).$$

Writing g for $(f^{(m)})^2$, we have that

$$\int_{B_l}(f^{(m)}(y))^2 S_m^2\left(\frac{y-c_l}{h}\right)dy = h\int_{-.5}^{.5} g(c_l+ht)S_m^2(t)dt =$$

$$h\int_{-.5}^{.5} g(c_l)S_m^2(t)dt + h\int_{-.5}^{.5} h\ t\ g^{(1)}(\theta(t))S_m^2(t)dt = h\int_{-.5}^{.5} g(c_l)S_m^2(t)dt + O(h^2),$$

being $\theta(t)$ an intermediate point between c_l and c_l+ht, and being the term $O(h^2)$ independent of l. Therefore,

$$\int_{B_l}(\mathrm{Bias}(\hat{f}_{\hat{w}}(y)))^2 dy = \frac{h^{2m}}{(m!)^2}hg(c_l)\int_{-.5}^{.5} S_m^2(t)dt + o(h^{2m}).$$

Recalling the notation used in Theorem 2, we have

$$\int_{I\!R} (\mathrm{Bias}(\hat{f}_{\hat{w}}(y)))^2 dy = \frac{h^{2m}}{(m!)^2} S_M^2(K) \sum_l hg(c_l) + o(h^{2m}) =$$

$$\frac{h^{2m}}{(m!)^2} S_M^2(K) \left(\int g(y)dy + O(h) \right) + o(h^{2m}) =$$

$$\frac{h^{2m}}{(m!)^2} S_M^2(K) \int (f^{(m)}(y))^2 dy + o(h^{2m}) = \frac{h^{2m}}{(m!)^2} S_M^2(K) R(f^{(m)}(y)) + o(h^{2m}). \tag{13}$$

<u>*Estimator variance.*</u>
According to (7), the random component in $\hat{f}_{\hat{w}}(y)$ is given by the sequence $\bar{K}_h(X_1, \ldots, X_n; c)$. We firstly consider its covariance structure. Let

$$V_{jj} = \mathrm{Var}\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - c_j}{h} \right) \right) = \frac{1}{nh^2} \mathrm{Var}\left(K\left(\frac{X_i - c_j}{h} \right) \right).$$

We have

$$\int K\left(\frac{z - c_j}{h} \right)^2 f(z)dz = \int_{c_j - \frac{mh}{2}}^{c_j + \frac{mh}{2}} K\left(\frac{z - c_j}{h} \right)^2 f(z)dz =$$

$$\int_{\frac{-mh}{2}}^{\frac{mh}{2}} K\left(\frac{t}{h} \right)^2 f(t + c_j)dt = \int_{\frac{-mh}{2}}^{\frac{mh}{2}} K\left(\frac{t}{h} \right)^2 (f(c_j) + tf^{(1)}(\theta_j(t)))dt =$$

$$\int_{\frac{-mh}{2}}^{\frac{mh}{2}} K\left(\frac{t}{h} \right)^2 f(c_j)dt + O(h^2) = hf(c_j) \int K(u)^2 du + O(h^2).$$

Also

$$\left[\int K\left(\frac{z - c_j}{h} \right) f(z)dz \right]^2 = \left[hf(c_j) \int K(u)du + O(h^2) \right]^2 = O(h^2),$$

and we obtain

$$V_{jj} = \frac{1}{nh} \left(f(c_j) \int K(u)^2 du \right) + O\left(\frac{1}{n} \right). \tag{14}$$

Consider now

$$V_{jk} = \mathrm{Cov}\left(\frac{1}{nh} \sum_i K\left(\frac{X_i - c_j}{h} \right), \frac{1}{nh} \sum_i K\left(\frac{X_i - c_k}{h} \right) \right) =$$

$$\frac{1}{nh^2} \mathrm{Cov}\left(K\left(\frac{X_i - c_j}{h} \right), K\left(\frac{X_i - c_k}{h} \right) \right).$$

Since

$$E\left(K\left(\frac{X_i - c_k}{h}\right)\right) = \int K\left(\frac{z - c_k}{h}\right) f(z)dz = O(h),$$

it follows that

$$\mathrm{Cov}\left(K\left(\frac{X_i - c_j}{h}\right), K\left(\frac{X_i - c_k}{h}\right)\right) =$$

$$\int K\left(\frac{z - c_j}{h}\right) K\left(\frac{z - c_k}{h}\right) f(z)dz + O(h^2) =$$

$$h\int K(u)K(u + (j-k))f(c_j + uh)du + O(h^2) = hf(c_j)\eta_{k-j} + O(h^2).$$

Therefore

$$V_{jk} = \frac{1}{nh} f(c_j)\eta_{k-j} + O\left(\frac{1}{n}\right). \tag{15}$$

Observe that this covariance term can be also written as

$$\frac{1}{nh} f(c_k)\eta_{j-k} + O\left(\frac{1}{n}\right),$$

because $c_k = c_j + h(k-j)$ and then

$$f(c_k) = f(c_j) + h(k-j)f^{(1)}(\theta_{jk}) = f(c_j) + O(h).$$

From (14) and (15) we obtain that

$$\mathrm{V}(\bar{K}_h(X_1, \ldots, X_n; c)) = \frac{1}{nh} DM + O\left(\frac{1}{n}\right),$$

where D is the infinite dimensional square diagonal matrix with j-th diagonal term equal to $f(c_j)$, $j \in Z$. Recalling (7), we have that

$$\mathrm{Var}\left(\hat{f}_{\hat{w}}(y)\right) = h^2 K_h(y;c)^{\mathrm{T}} M^{-1}\left[\frac{1}{nh} DM + O\left(\frac{1}{n}\right)\right] M^{-1} K_h(y;c) =$$

$$\frac{h}{n} K_h(y;c)^{\mathrm{T}} M^{-1} D K_h(y;c) + h^2 K_h(y;c)^{\mathrm{T}} M^{-1} O\left(\frac{1}{n}\right) M^{-1} K_h(y;c),$$

that are terms with order $1/(nh)$ and $1/n$, respectively. Therefore,

$$\mathrm{Var}\left(\hat{f}_{\hat{w}}(y)\right) = \frac{h}{n} \mathrm{tr}\left(K_h(y;c)K_h(y;c)^{\mathrm{T}} M^{-1} D\right) +$$

$$h^2 \mathrm{tr}\left(K_h(y;c)K_h(y;c)^{\mathrm{T}} M^{-1} O\left(\frac{1}{n}\right) M^{-1}\right),$$

where $\mathrm{tr}(A)$ is the trace of matrix A. Since

$$M = \int hK_h(y;c)K_h(y;c)^{\mathrm{T}} dy = O\left(\frac{1}{h}\right),$$

it follows that

$$\int \mathrm{Var}\left(\hat{f}_{\hat{w}}(y)\right) dy = \frac{1}{n}\mathrm{tr}(D) + h\mathrm{tr}\left(O\left(\frac{1}{n}\right) M^{-1}\right) =$$

$$\frac{1}{n}\mathrm{tr}(D) + O\left(\frac{1}{n}\right) = \frac{1}{nh}\mathrm{tr}(hD) + O\left(\frac{1}{n}\right) = \frac{1}{nh}\left(\sum_j f(c_j)h\right) + O\left(\frac{1}{n}\right) =$$

$$\frac{1}{nh}\left(\int f(z)dz + O(h)\right) + O\left(\frac{1}{n}\right) = \frac{1}{nh} + O\left(\frac{1}{n}\right). \tag{16}$$

Finally, from (13) and (16), we conclude

$$\mathrm{MISE}\left(\hat{f}_{\hat{w}}\right) = \frac{1}{nh} + \frac{h^{2m}}{(m!)^2} S_m^2(K) R(f^{(m)}) + O\left(\frac{1}{n}\right) + o(h^{2m}).$$

□

References

Boneva, L. I., D. Kendall, and I. Stefanov (1971). Spline transformations: Three new diagnostic aids for the statistical data-analyst. *J. Roy. Statist. Soc. Ser. B*, **33**, 1–70.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications. Vol. II, 2 ed.* New York: J.Wiley.

Hall, P. (1982). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *SIAM J. Appl. Math.*, **42**, 390–399.

Jones, M. C. (1989). Discretized and interpolated kernel density estimates. *J. Amer. Statist. Assoc.*, **84**, 733–741.

Jones, M. C., M. Samiuddin, A. H. Al-Harbey, and T. A. H. Maatouk (1998). The edge frequency polygon. *Biometrika*, **85**, 235–239.

Schumaker, L.L. (1981). *Spline Functions: Basic Theory*. New York: J. Wiley.

Scott, D. W. (1992). *Multivariate Density Estimation*. New York: J. Wiley.

Scott, D. W. and S. J. Sheather (1985). Kernel density estimation with binned data. *Comm. Statist. Theory Meth.*, **14**, 1353–1359.

Simonoff, J.S. (1995). The anchor position of histograms and frequency polygons: quantitative and qualitative smoothing. *Comm. Statist. Simul. Computat.*, **24**, 691–710.

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

Simonoff, J. S. and F. Udina (1997). Measuring the stability of histogram appearance when the anchor position is changed. *Comput. Statist. Data Anal.*, **23**, 335–353.