Bin Interval Method of Locally Adaptive Nonparametric Density Estimation

Masahiko Sagae* David W. Scott†

1997.10.7

Abstract

A locally adaptive nonparametric polynomial density estimator is proposed. The main idea is to apply bin-by-bin sample moment information to the estimation of the density function. The density estimate is constrained to match the actual sample moments of observations on each bin. The asymptotic properties of the general order case are derived. The properties of the new estimator are compared with those of other nonparametric density estimators (i.e., histogram, frequency polygon and kernel estimators). Implementation and computational advantages are also discussed. Examples are given to illustrate the method.

KEY WORDS: Nonparametric density estimation, moment-matching constraints, locally adaptive polynomial function, histogram, IMSE, local moment method, polynomial histogram.

1 Introduction

A histogram is the simplest nonparametric density estimator based on counting data in each bin for estimation. However, its optimal integrated mean squared error (IMSE) convergence rate of $O(n^{-2/3})$ is inferior to the rate of other commonly used nonparametric density estimators. The slow rate is one of the main reason why the histogram lacks in theoretical appeal. Scott (1985a, 1985b) proposed the frequency polygon and averaged shifted

*Department of Information Sciences, Gifu University, 1–1 Yanagido, Gifu, Gifu 501–11 Japan. This research was performed while the author was a visiting scholar at Rice University.

†Department of Statistics MS-138, Rice University, 6100 Main Street, Houston, TX 77005-1892 USA. This research was supported in part by NSF grant DMS-9626187.

histogram, which both take advantage of the correlation from bin-to-bin, and achieve an IMSE convergence rate of $O(n^{-4/5})$.

In this paper, we propose a locally adaptive nonparametric polynomial density function, which is estimated by matching actual sample moments of observations in each bin up to the highest order of the polynomial function. That is, the density function is estimated from the data by a technique we call a "local moment method." Our local polynomial density estimator is a natural way to implement the moment-preserving constraints on each bin. The 0-th order case of our new estimator is equivalent to the histogram. In the linear case, useful intuition can be gleaned. Ordinarily to determine if f is increasing, one checks if the bin count is increasing in the adjacent bin (cf. frequency polygon). Alternatively, one may check *within a bin* to see if there are more points to the right of the bin center than to the left; if so, the bin mean will be greater than the midpoint ($\bar{x}_j >$ midpoint). Similarly, local curvature in f is related to the variance within a bin.

For the k-th order polynomial, we present the general (higher-order) asymptotic expressions of the integrated mean square error (IMSE). The main idea is to put the local moment information obtained from samples into the estimating procedure of the nonparametric density estimation. Not only does our estimator use the sample-counting information on bins like a histogram, but also the higher-order sample-moment information on bins. We call our new estimator a "polynomial histogram."

We explore the relationship to kernel estimators based upon binned data as described in Silverman (1982), Scott (1983), Scott and Sheather (1985), and Wand and Jones (1995). These approaches, however, use only counting data. The binned (positive) kernel estimator has the IMSE convergence rate of O($n^{-4/5}$). Properties relating the choice of kernel function and the (global) sample moments of the data and kernel estimator are discussed in Terrell (1984). An extension of the local likelihood method to kernel density estimation is considered by Hjort and Jones (1996) and Loader (1996). Loader supposes that a local log-polynomial form well approximates the unknown density, and that the estimator is a kernel-type density function. The coefficients of the local log-polynomial form are estimated numerically from a local likelihood criterion. Loader demonstrates that solutions of the likelihood equations have a form that incorporates local sample moments, which depend on the kernel window width. A practical problem in the estimation is the need to evaluate the value of the integral of the density estimation for the normalization. The resulting expression with the local sample moments is similar to our approach. Some points of difference between the approach of Loader and ours are as follows: (i) Our local density model has a polynomial rather than a log-polynomial form; hence, our estimate can be negative; (ii) Our estimating technique is a local moment method and not local likelihood method; (iii) Parameters of our estimating density function for a general order have closed-form algebraic solutions rather than iterative ones; (iv) The convergence rate of AMISE for our general-order polynomial histogram is given; (v) Computational cost for estimating the polynomial histogram is less expensive, even if the higher-order polynomial density is selected.

In Section 2, we begin with some notation and definitions. In Section 3, we formally introduce the locally adaptive nonparametric polynomial density estimator and derive its theoretical properties. In Section 4, examples are given illustrating our estimator. In Section 5, we discuss about some properties of this new estimator and describe some advantages.

2 Notation and Definitions

For convenience, we introduce some notation and give some definitions. We consider a piecewise polynomial density function defined on equally spaced bins of width h, and let the number of mesh intervals be denoted by m. Let I_j denote the j-th bin interval, n_j denote the number of sample falling in I_j, and $n = \sum_{j=1}^{m} n_j$, which is the sample size. Let the data $x_1, x_2, \ldots, x_n$ on the overall interval I be independently and identically distributed with true density function $f(x)$. We assume the k-th derivative $f^{(k)}(x)$ is absolutely continuous on I. We define the roughness functional $R(f^{(i)}) = \int_{-\infty}^{\infty} (f^{(i)}(x))^2 dx$.

Without loss of generality, consider the bin centered on the origin $I_0 = (-h/2, h/2)$. Define the k-th order polynomial density function on this bin as

$$f_k(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_k x^k, \tag{1}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_k$ are unknown parameters. Equivalently, we can use a linear combination of orthonormal polynomial functions on $(-h/2, h/2)$ as follows,

$$f_k(x) = a_0 p_0(x) + a_1 p_1(x) + \cdots + a_k p_k(x), \tag{2}$$

where $p_i(x)$ for $x \in (-h/2, h/2)$ satisfy $\int_{I_0} p_i(x) p_j(x)\, dx = \delta_{ij}$, where $\delta_{ii} = 1$ and $\delta_{ij} = 0$ when $i \neq j$. The $p_i(x)$ may be written explicitly as

$$p_i(x) = \begin{cases} c_0^{(i)} + c_2^{(i)} x^2 + \cdots + c_i^{(i)} x^i & \text{for even } i \\ \\ c_1^{(i)} x + c_3^{(i)} x^3 + \cdots + c_i^{(i)} x^i & \text{for odd } i \end{cases} \tag{3}$$

where the sequence $\{c_j^{(i)}\}_{j=0}^{i}$ is constant on the support $(-h/2, h/2)$. We remark that the sequence $\{c_j^{(i)}\}$ contains the coefficients of the normalized Legendre polynomials on $(-h/2, h/2)$.

We require of the estimated density function that it preserve the following mass constraint:

$$\int_{I_j} \hat{f}_k(x) dx = n_j/n \equiv m_0'(I_j). \tag{4}$$

If $k > 0$, we also require that it preserve the following local sample moment constraints on I_j around the bin center t_j:

$$\int_{I_j} (x - t_j)^i \hat{f}_k(x) dx \Big/ \int_{I_j} \hat{f}_k(x) dx = m_i(I_j), \tag{5}$$

for $i = 1, \ldots, k$, where the i-th order (central) local sample moment about the bin center is given by

$$m_i(I_j) = \frac{1}{n_j} \sum_{x_k \in I_j} (x_k - t_j)^i \tag{6}$$

and

$$m'_i(I_j) = \frac{1}{n} \sum_{x_k \in I_j} (x_k - t_j)^i \tag{7}$$

for $i \geq 0$. The true local moments are defined by

$$\mu_i(I_j) = \int_{I_j} (x - t_j)^i f(x) dx / \int_{I_j} f(x) dx \tag{8}$$

and

$$\mu'_i(I_j) = \int_{I_j} (x - t_j)^i f(x) dx \tag{9}$$

for $i = 0, \ldots, k$. Note that $\mu_i(\cdot) = \mu'_i(\cdot)/\mu'_0(\cdot)$ and $m_i(\cdot) = m'_i(\cdot)/m'_0(\cdot)$ hold.

By implementing the local sample moment constraints as in equations (4) and (5), we can uniquely determine estimates of the parameters $\alpha_0, \alpha_1, \ldots, \alpha_k$. We actually find the set of equivalent parameters, $a_0, a_1, \ldots, a_k$, which may be expressed in closed form, as shown in the next section.

3 Theoretical Properties of Locally Adaptive Polynomial Density Estimators

We derive an expression for the global integrated mean squared error (IMSE) of our locally adaptive nonparametric polynomial density estimator (i.e., polynomial histogram). To derive the IMSE, we first evaluate the global integrated squared bias (ISB) and the integrated variance (IV) based on a pointwise approximation. Throughout this section, the estimator $f_k(x)$ will be expressed with an orthonormal polynomial basis for convenience of the proof.

Theorem 1: *Let the k-th order local polynomial function be defined by*

$$f_k(x) = a_0 p_0(x) + a_1 p_1(x) + \ldots + a_k p_k(x), \tag{10}$$

on $(-h/2, h/2)$ with the sample moment constraints given in (4) and (5). Then the integrated squared bias is given by

$$\text{ISB} = \{(k+1)!\, d_{k+1}^{(k+1)}\}^{-2} h^{2(k+1)} R(f^{(k+1)}) + O(h^{2(k+2)}) \tag{11}$$

and

$$c_{k+1}^{(k+1)} = d_{k+1}^{(k+1)} h^{-(k+1)-1/2}. \tag{12}$$

Note that $c_{k+1}^{(k+1)}$, which depends on a scaling factor h, is already given in (3) and $d_{k+1}^{(k+1)}$ is the constant part in $c_{k+1}^{(k+1)}$.

Theorem 2: *For the k-th order polynomial function, we have the integrated variance,*

$$\text{IV} = \frac{k+1}{nh} - \frac{R(f)}{n} + O(\frac{h}{n}), \tag{13}$$

where $k+1$ is the number of parameters on each bin in the polynomial histogram.

Before proving these Theorems, we need the following Lemmas. Throughout the following proofs, m_i' refers to $m_i'(I_0)$ and μ_i' refers to $\mu_i'(I_0)$. Recall that $t_0 = 0$.

Lemma 1: *Let the following local sample moment constraints*

$$\int_{I_0} x^i \hat{f}_k(x) dx = m_i',$$

for $i = 0, 1, 2, \ldots, k$, determine the coefficients $\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_k$ in $\hat{f}_k(x)$. Then, the $(2i)$-th coefficient $\hat{a}_{2i}$ for any $2i \le k$ is given by

$$\hat{a}_{2i} = \sum_{j=0}^{i} c_{2j}^{(2i)} m_{2j}'. \tag{14}$$

The $(2i+1)$-th coefficient $\hat{a}_{2i+1}$ for $2i+1 \le k$ is also given by

$$\hat{a}_{2i+1} = \sum_{j=0}^{i} c_{2j+1}^{(2i+1)} m_{2j+1}'. \tag{15}$$

Proof of Lemma 1: We prove the equality in (14) for the even case. By the orthonormality of $p_i(x)$ and $p_j(x)$, we have the following equality for any $2i \leq k$,

$$\begin{aligned}
\hat{a}_{2i} &= \sum_{j=0}^{k} \hat{a}_j \int_{I_0} p_{2i}(x) p_j(x) dx \\
&= \int_{I_0} p_{2i}(x) \hat{f}_k(x) dx \\
&= \sum_{j=0}^{i} c_{2j}^{(2i)} \int_{I_0} x^{2j} \hat{f}_k(x) dx \\
&= \sum_{j=0}^{i} c_{2j}^{(2i)} m'_{2j}.
\end{aligned}$$

The equality in (15) is proved similarly for the odd case. This completes the proof.

Recall that $I_0 = (-h/2, h/2)$ defines the support of orthonormal polynomial $p_i(x)$. Thus, we obtain the following lemma.

Lemma 2: *Let the k-th order orthonormal polynomial function be defined by $p_k(x) = \sum_{j=0}^{k} c_j^{(k)} x^j$. Then, the i-th coefficient $c_i^{(k)}$ is given by*

$$c_i^{(k)} = d_i^{(k)} h^{-i-1/2}, \tag{16}$$

where $d_i^{(k)}$ is constant and does not depend on either the data or h.

Proof of Lemma 2:

Let $q_k(x)$ denote the k-th order normalized Legendre polynomials on $(-1/2, 1/2)$ and let $q_k(x) = \sum_{\substack{j=0 \\ j\ even}}^{k} d_j^{(k)} x^j$ for even k and $\sum_{\substack{j=1 \\ j\ odd}}^{k} d_j^{(k)} x^j$ for odd k and define $d_0^{(0)} = 1$. We consider expressing the function $p_k(x)$ by using $q_k(x)$. Each domain of $q_k(x)$ and $p_k(y)$ has $x \in (-1/2, 1/2)$ and $y \in (-h/2, h/2)$, respectively. This change of the domain corresponds to the scaling transformation of variable, that is, $y = hx$. The integrals of the square of $p_k(y)$ over the domain equals one and $q_k(x)$ also does. Thus, we have the relation,

$$p_k^2(y) = q_k^2\left(\frac{y}{h}\right)\frac{1}{h}.$$

Noting $q_0(x) = 1$ and $p_0(y) > 0$, the equality

$$p_k(y) = \frac{1}{\sqrt{h}} q_k(\frac{y}{h})$$

holds.

Hence we have the following equality

$$c_i^{(k)} = d_i^{(k)} h^{-i-1/2}.$$

This completes the proof.

Now we prove the theorems. Define the inner product $\langle f(x), g(x)\rangle = \int_{I_0} f(x)g(x)\,dx$.

Proof of Theorem 1: By Lemma 1, for the $2k$-th order polynomial density function $\hat{f}_{2k}(x)$, we have

$$\hat{a}_i = \langle p_i(x), \hat{f}_{2k}(x)\rangle = \sum_{j=0}^{i} c_j^{(i)} \langle x^j, \hat{f}_{2k}(x)\rangle = \sum_{j=0}^{i} c_j^{(i)} m'_j.$$

The expectation of $\hat{a}_i$ with respect to the true density function $f(x)$ is given by

$$E_f \hat{a}_i = \sum_{j=0}^{i} c_j^{(i)} E_f m'_j = \sum_{j=0}^{i} c_j^{(i)} \mu'_j = \langle p_i(x), f(x)\rangle, \tag{17}$$

where m'_j and μ'_j are defined in (7) and (9). Hence, noting that the "x" in $\hat{f}_{2k}(x)$ is not random but is fixed, we have $E_f \hat{f}_{2k}(x) = \sum_{j=0}^{2k} (E_f \hat{a}_j) p_j(x) = \sum_{j=0}^{2k} \langle p_j(t), f(t)\rangle p_j(x)$. We express the above equality by using the Taylor series expansion of $f(x)$,

$$\begin{aligned}
E_f \hat{f}_{2k}(x) &= \sum_{j=0}^{2k} \sum_{l=0}^{\infty} \frac{f^{(l)}(0)}{l!} \langle p_j(t), t^l\rangle p_j(x) \\
&= \sum_{j=0}^{2k} \sum_{l \geq j} \frac{f^{(l)}(0)}{l!} \langle p_j(t), t^l\rangle p_j(x) \\
&= \frac{f^{(0)}(0)}{0!} \langle p_0(t), t^0\rangle p_0(x) \\
&\quad + \frac{f^{(1)}(0)}{1!} \langle p_1(t), t^1\rangle p_1(x) \\
&\quad + \frac{f^{(2)}(0)}{2!} \{ \langle p_0(t), t^2\rangle p_0(x) + \langle p_2(t), t^2\rangle p_2(x) \}
\end{aligned}$$

$$
\begin{aligned}
&+\cdots\\
&+\frac{f^{(2k)}(0)}{2k!}\{\,\langle p_0(t),t^{2k}\rangle\,p_0(x)+\cdots+\,\langle p_{2k}(t),t^{2k}\rangle\,p_{2k}(x)\}\\
&+\frac{f^{(2k+1)}(0)}{(2k+1)!}\{\,\langle p_1(t),t^{2k+1}\rangle\,p_1(x)+\cdots+\,\langle p_{2k}(t),t^{2k+1}\rangle\,p_{2k}(x)\}\\
&+\cdots.
\end{aligned}
$$

Next consider the following equation in $\{\beta_i\}_{i=0}^{k}$:

$$
\beta_0 p_0(x)+\beta_1 p_1(x)+\cdots+\beta_k p_k(x)=x^k. \tag{18}
$$

Solving for the coefficients β_i, we obtain

$$
\beta_i=\ \langle p_i(t),t^k\rangle\,,\ \text{for } i=0,1,\cdots,k. \tag{19}
$$

By the equalities in (18) and (19), we have

$$
\begin{aligned}
\mathrm{E}_f\hat{f}_{2k}(x) \;&=\; f(0)+\frac{f^{(1)}(0)}{1!}x+\frac{f^{(2)}(0)}{2!}x^2+\cdots+\frac{f^{(2k+1)}(0)}{(2k+1)!}x^{2k+1}\\
&\quad -\frac{f^{(2k+1)}(0)}{(2k+1)!}\,\langle p_{2k+1}(t),t^{2k+1}\rangle\,p_{2k+1}(x)+O(x^{2k+2}).
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
\mathrm{Bias}\hat{f}_{2k}(x) \;&=\; \mathrm{E}_f\hat{f}_{2k}(x)-f(x)\\
&=\; -\frac{f^{(2k+1)}(0)}{(2k+1)!}\,\langle p_{2k+1}(t),t^{2k+1}\rangle\,p_{2k+1}(x)+O(x^{2k+2}).
\end{aligned} \tag{20}
$$

Similarly, we can derive Bias $\hat{f}_{2k+1}(x)$ as

$$
\mathrm{Bias}\hat{f}_{2k+1}(x)=-\frac{f^{(2k+2)}(0)}{(2k+2)!}\,\langle p_{2k+2}(t),t^{2k+2}\rangle\,p_{2k+2}(x)+O(x^{2k+3}). \tag{21}
$$

For $x\in I_j$, a similar expression holds but with $f^{(2k+1)}(0)$ replaced by $f^{(2k+1)}(t_j)$. Noting $\langle p_k(t),t^k\rangle\ =\ 1/c_k^{(k)}$, by Lemma 2, (20) and (21), we obtain the integrated squared bias (ISB) of the k-th order polynomial density function $\hat{f}_k(x)\ =\ a_0p_0(x)+a_1p_1(x)+\cdots+$

$a_k p_k(x)$, by squaring the bias, integrating over I_j, summing over all j, and using standard approximations to obtain

$$\mathrm{ISB} = \{(k+1)!\, d_{k+1}^{(k+1)}\}^{-2} h^{2(k+1)} R(f^{(k+1)}) + O(h^{2(k+2)}).$$

This completes the proof.

Proof of Theorem 2: We prove Theorem 2 by induction. We assume that the integrated variance for the k-th order polynomial density function is given by

$$\mathrm{IV}(\hat{f}_k) = \frac{k+1}{nh} - \frac{R(f)}{n} + O(\frac{h}{n}). \tag{22}$$

For the 0-th order polynomial density function, the integrated variance is given by

$$\mathrm{IV}(\hat{f}_0) = \frac{1}{nh} - \frac{R(f)}{n} + O(\frac{h^2}{n}).$$

Hence the assertion for $k = 0$ is true. We consider the $(k+1)$-th order case for odd k and the density estimator is expressed by

$$\hat{f}_{k+1}(x) = \hat{f}_k(x) + \hat{a}_{k+1} p_{k+1}(x). \tag{23}$$

Noting the relation between $\hat{f}_k$ and $\hat{f}_{k+1}$, the variance of $\hat{f}_{k+1}$ is given by

$$V(\hat{f}_{k+1}) = V(\hat{f}_k) + V(\hat{a}_{k+1} p_{k+1}(x)) + 2Cov(\hat{f}_k, \hat{a}_{k+1} p_{k+1}(x)). \tag{24}$$

By Lemma 1, $\hat{a}_{k+1}$ is expressed as

$$\hat{a}_{k+1} = \sum_{j \le (k+1)/2} c_{2j}^{(k+1)} m'_{2j}, \tag{25}$$

and its expectation on the true density function $f(x)$ is given by

$$\mathrm{E}_f(\hat{a}_{k+1}) = \sum_{j \le (k+1)/2} c_{2j}^{(k+1)} \mu'_{2j},. \tag{26}$$

In the right hand side of (24), we need to evaluate $V(\hat{a}_{k+1}p_{k+1}(x))$ and $Cov(\hat{f}_k, \hat{a}_{k+1}p_{k+1}(x))$. Firstly, we evaluate the term $V(\hat{a}_{k+1}p_{k+1}(x))$, i.e.,

$$V(\hat{a}_{k+1}p_{k+1}(x)) = p_{k+1}(x)^2 V(\hat{a}_{k+1}) = p_{k+1}(x)^2\{E_f(\hat{a}_{k+1}^2) - E_f(\hat{a}_{k+1})^2\}. \tag{27}$$

The first term in (27) is given by

$$\mathrm{E}(\hat{a}_{k+1}^2) = \sum_{0\le j\le (k+1)/2} (c_{2j}^{(k+1)})^2 E_f(m_{2j}'^2) + \sum_{0\le i\ne j\le (k+1)/2} c_{2i}^{(k+1)} c_{2j}^{(k+1)} E_f(m_{2i}' m_{2j}'), \tag{28}$$

and we have

$$\mathrm{E}(m_{2j}'^2) = \frac{\mu_{4j}'}{n} + \frac{n-1}{n}\mu_{2j}'^2, \tag{29}$$

and

$$\mathrm{E}(m_{2i}' m_{2j}') = \frac{\mu_{2(i+j)}'}{n} + \frac{n-1}{n}\mu_{2i}'\mu_{2j}', \tag{30}$$

where $\mu_j'(\cdot)$ is defined in (9). By substituting (29) and (30) into (28), we have

$$\begin{aligned} \mathrm{E}(\hat{a}_{k+1}^2) &= \sum_{0\le j\le (k+1)/2} (c_{2j}^{(k+1)})^2 \left(\frac{\mu_{4j}'}{n} + \frac{n-1}{n}\mu_{2j}'^2\right) \\ &\quad + \sum_{0\le i\ne j\le (k+1)/2} c_{2i}^{(k+1)} c_{2j}^{(k+1)} \left(\frac{\mu_{2(i+j)}'}{n} + \frac{n-1}{n}\mu_{2i}'\mu_{2j}'\right). \end{aligned} \tag{31}$$

Hence, noting here that

$$\{\mathrm{E}_f(\hat{a}_{k+1})\}^2 = \sum_{0\le j\le (k+1)/2} (c_{2j}^{(k+1)})^2 \mu_{2j}'^2 + \sum_{0\le i\ne j\le (k+1)/2} c_{2i}^{(k+1)} c_{2j}^{(k+1)} (\mu_{2i}'\mu_{2j}'),$$

we have

$$\begin{aligned} V(\hat{a}_{k+1}) &= \sum_{0\le j\le (k+1)/2} (c_{2j}^{(k+1)})^2 (\mu_{4j}' - \mu_{2j}'^2)/n \\ &\quad + \sum_{0\le i\ne j\le (k+1)/2} c_{2i}^{(k+1)} c_{2j}^{(k+1)} (\mu_{2(i+j)}' - \mu_{2i}'\mu_{2j}')/n \\ &= \sum_{0\le j\le (k+1)/2} (c_{2j}^{(k+1)})^2 \left(\frac{h^{4j+1}}{(4j+1)2^{4j}} + O(\frac{h^{4j+2}}{n})\right) \\ &\quad + \sum_{0\le i\ne j\le (k+1)/2} c_{2i}^{(k+1)} c_{2j}^{(k+1)} \left(\frac{h^{2(i+j)+1}}{(2(i+j)+1)2^{2(i+j)}}\frac{f(0)}{n} + O(\frac{h^{2(i+j)+2}}{n})\right), \end{aligned} \tag{32}$$

by using the approximation of the integration, that is,

$$\mu'_{4j}/n \approx \frac{h^{4j+1}}{(4j+1)2^{4j}}\frac{f(0)}{n} + O(\frac{h^{4j+3}}{n})$$

and

$$\mu'^2_{2j}/n \approx \left\{\frac{h^{2j+1}}{(2i+1)2^{2j}}f(0) + O(h^{2j+3})\right\}^2/n = O(\frac{h^{4j+2}}{n}).$$

Hence, the variance is given as,

$$\begin{aligned}
V(\hat{a}_{k+1}) &\approx \sum_{0\le j\le (k+1)/2}\left\{(c_{2j}^{(k+1)})^2\frac{h^{4j+1}}{(4j+1)2^{4j}}\frac{f(0)}{n} + O(\frac{h^{4j+2}}{n})\right\} \\
&\quad + \sum_{0\le i\ne j\le (k+1)/2}\left\{c_{2i}^{(k+1)}c_{2j}^{(k+1)}\frac{h^{2(i+j)+1}}{(2(i+j)+1)2^{2(i+j)}}\frac{f(0)}{n} + O(\frac{h^{2(i+j)+2}}{n})\right\} \\
&= \int_{I_0}\left(\sum_{0\le i\le (k+1)/2} c_{2i}^{(k+1)}x^{2i}\right)^2 dx\ \frac{f(0)}{n} + O(\frac{h^2}{n}) \\
&= \int_{I_0} p_{k+1}^2(x)\,dx\ \frac{f(0)}{n} + O(\frac{h^2}{n}), \\
&= \frac{f(0)}{n} + O(\frac{h^2}{n}).
\end{aligned} \tag{33}$$

Therefore, we have

$$\mathrm{IV}(\hat{a}_{k+1}p_{k+1}(x)) = \frac{f(0)}{n} + O(\frac{h^2}{n}), \tag{34}$$

on the interval $(-h/2, h/2)$. Hence, we have

$$\mathrm{IV}(\hat{a}_{k+1}p_{k+1}(x)) = \frac{1}{nh} + O(\frac{h}{n}), \tag{35}$$

on the global interval.

Next we evaluate $Cov(\hat{f}_k, \hat{a}_{k+1}p_{k+1}(x))$ in the right hand side of (24). The integration of the covariance term on I_0 is zero by the orthogonality between $f_k(x)$ and $p_{k+1}(x)$.

By the assumption in (22) and (35), we have the global integrated variance of $\hat{f}_{k+1}(x)$ is

$$\mathrm{IV}(\hat{f}_{k+1}) = \frac{(k+2)}{nh} - \frac{R(f)}{n} + O(\frac{h}{n}). \tag{36}$$

Note that $\mathrm{IV}(\hat{f}_{k+1})$ for even k is proved similarly, completing the proof. From Theorems 1 and 2, the following corollaries are easily derived, which we give without proof.

Corollary 1: *For the k-th order polynomial density function, we have*

$$\mathrm{IMSE}(\hat{f}_k) = \frac{k+1}{nh} + \{(k+1)!\, d_{k+1}^{(k+1)}\}^{-2} h^{2(k+1)} R(f^{(k+1)}). \tag{37}$$

By minimizing the IMSE in (37) with respect to h, we have the following two results.

Corollary 2: *The optimal bin width h^* is given by*

$$h^* = \{\frac{((k+1)!)^2 (d_{k+1}^{(k+1)})^2}{2R(f^{(k+1)})}\}^{\frac{1}{2k+3}} n^{-\frac{1}{2k+3}}. \tag{38}$$

Corollary 3: *The minimum* IMSE *is given by*

$$\mathrm{IMSE}^* = (2k+3)\{(k+1)!\, d_{k+1}^{(k+1)}\}^{-\frac{2}{2k+3}} R(f^{(k+1)})^{\frac{1}{2k+3}} (2n)^{-\frac{2k+2}{2k+3}}. \tag{39}$$

4 Examples

We show polynomial histogram estimates for two simulated data sets and one real data set in this section. In our simple implementation, we do the polynomial estimation bin-by-bin so that the final estimate is not even continuous across the bin boundaries. We take it as a positive sign if the polynomial estimates in two adjacent bins are nearly continuous across the bin boundaries.

Figure 1 displays four polynomial histogram estimates of 10,000 Normal observations $N(0,1)$ with the order numbers $k = 0, 1, 2, 3$. For each estimate, $h = 8/9$ and the range is $(-4, 4)$. Figure 2 displays four polynomial histograms of 10,000 normal mixture observations from $0.5N(-1.5, 0.5) + 0.5N(1, 1)$. For each figure, $h = 8/11$ and the range is again $(-4, 4)$. In each case, the estimates show some (visual) improvement as k is increased, but the case $k = 3$ shows signs of high variance for each example.

Finally, we show the numerical example obtained by applying our first and second order estimators to the actual data which have been analyzed in the literature (Good and Gaskins, 1980; Scott, 1992). The so-called LRL data have a sample size of 25,752. Figure 3 displays two of our first order and two of our second order estimators each with $h = 60$ and $h = 90$ MeV.

The bin-by-bin construction of a higher-order density estimate is particularly attractive for massive data set analysis. However, the lack of global smoothness is a distraction. In another article, we show how additional constraints may be added to the polynomial histogram so that the entire estimate is, for example, continuous. For example, we illustrate a locally fifth degree polynomial estimate which is constrained to: (i) match 2 local moments as in (4) and (5); (ii) have 2 continuous derivatives (as in a cubic spline); and (iii) minimize the roughness $\int \hat{f}''(x)^2 dx$. This estimator for the LRL data is shown in Figure 4 with $h = 100$ and should be compared directly with the $(h, k) = (100, 1)$ frame in Figure 3. We remark that the solution is obtained from a linear system of equations of size $6m \times 6m$. Software to generate these and similar figures is available at ftp.stat.rice.edu in pub/scottdw/PH.code.

5 Discussion and Concluding Remarks

5.1 Discussion

We discuss the new estimator (i.e., polynomial histogram) and compare it with other nonparametric density estimators.

5.1.1 Higher order efficiency

We discuss the higher order properties in comparison with kernel estimator. Bartlett (1963) pointed out that a kernel of order p, which satisfies $\int_S K(x)\,dx = 1$, $\int_S x^i K(x)\,dx = 0$ for $i = 1, 2, \ldots, p-1$, and $\int_S x^p K(x)\,dx = \mu_p$ where $\mu_p \neq 0$, reduces the asymptotic bias to

$O(h^p)$. Then the AIMSE, the optimal h and the minimal AIMSE are given as,

$$\begin{aligned} \text{AIMSE}(h) &= \frac{R(K)}{nh} + \frac{1}{(p!)^2} h^{2p} \mu_p^2 R(f^{(p)}), & (40) \\ h^* &= \{\frac{(p!)^2 R(K)}{2p\mu_p^2 R(f^{(p)})}\}^{\frac{1}{2p+1}} n^{-\frac{1}{2p+1}}, & (41) \end{aligned}$$

and

$$\text{AIMSE}^* = \{\mu_p^2 R(K)^{2p} R(f^{(p)})\}^{\frac{1}{2p+1}} n^{-\frac{2p}{2p+1}}. \quad (42)$$

Though the polynomial histogram estimator is classified into a bin-by-bin estimator like a histogram, by applying a locally $(p-1)$-th order polynomial function, the AIMSE convergence rate achieves $O(n^{-\frac{2p}{2p+1}})$. The density function is estimated matching by the local sample moments.

Hence, if we need higher order asymptotic convergence rate with respect to IMSE, then it is easy to get the estimator with higher order preferable property by using the higher order local sample moments. On the other hand, the constraints of the higher order kernel must be determined *a priori* and their constraints do not have any relationships with sample information. Our method improves bias but not the integrated variance. It is interesting that IV is related to the number of the parameters of the polynomial histogram, that is,

$$\text{IV} = \frac{\#parameters}{nh}.$$

Clearly, the IV grows worse as the order of the polynomial increases.

We remark that the increased variance can be virtually eliminated by allowing for the computation of the higher-order local moments in ever wider (and hence overlapping) intervals. Our current discussion limits the computation of all local moments to the same interval. We do not present details here. However, for the particular application of density estimation to terabyte-sized massive data sets, we believe the increased computational efficiency of the higher-order polynomial histogram far outweighs the loss of statistical efficiency due to inflated variance.

5.1.2 Optimal bandwidth of polynomial histogram

In this section, we compare the optimal h for various estimators under the assumption that the samples are independently and identically distributed as $N(\mu, \sigma^2)$.

Table 5.1.2 Comparison of Optimal h^*'s. The final column gives the ratio compared to h^*_{1st} of the first-order polynomial histogram.

density estimator	*the value of the optimal h*	*ratio*
First-order polynomial histogram	$4.43\sigma n^{-1/5}$	1.00
Normal kernel estimator	$1.06\sigma n^{-1/5}$	0.24
Epanechnikov kernel estimator	$2.35\sigma n^{-1/5}$	0.53
Frequency Polygon (FP) estimator	$2.15\sigma n^{-1/5}$	0.49
Histogram	$3.49\sigma n^{-1/3}$	$0.78n^{-2/15}$
		0.31 (for $n = 1000$)
		0.22 (for $n = 100000$)

Table 5.1.2 gives optimal h^* for the first order polynomial histogram and four typical density estimators. Our estimator has the widest optimal bin width, that is, $h^*_{1st} \approx 4h^*_{normal}$, $h^*_{1st} \approx 2h^*_{Epan.}$ and $h^*_{1st} \approx 2h^*_{FP}$. For the histogram, we have $h^*_{1st} > h^*_{hist}$ for any n such as $h^*_{1st} \approx 3h^*_{hist}$ for $n = 1,000$ and $h^*_{1st} \approx 5h^*_{hist}$ for $n = 10,000$. We note that h_{1st} and h_{hist} correspond to bin widths but they are somewhat different from scale parameter of kernel estimators. This property is preferable for most statistician with applications, because we can use more data collection than others for the limited sample size.

5.1.3 Location of bin point

In general, bin-by-bin density estimators are largely influenced by the location of bin points, even though the bin width and the number of bin are well chosen. Our estimator alleviates vital influence of the location of bin points by using the piece-wise polynomial function and also by matching the local sample moments from actual observations. In bin-by-bin density estimators, the FP and averaged shifted histogram improve the AIMSE convergence rate by the effect of interpolation among neighboring bins. It seems that the most advantageous effect of its interpolation is to alleviate the influence of the location of bin points.

5.2 Concluding Remarks

Briefly, we conclude that the local sample moments of observations are very informative with respect to the estimation of parameters of the polynomial histogram. However, we need to carefully choose the number of polynomial order depending on data structure and the number of data. This method is easy to compute the value of the parameters regardless of the order of polynomial, because their parameters have formulaic answers with respect to local sample moments in Lemma 1. It seems that this local moment method has the wide range of application for other nonparametric settings.

Just as ordinary parametric estimation can be accomplished by either the method of maximum likelihood or the method of moments, we see that nonparametric density estimators can also be formulated with either technique. In our growing work with massive data sets, we see a large interest in techniques that avoid numerical optimization. In such a setting, we believe that the accuracy in the local moments will be more than sufficient to offset any theoretical loss of efficiency compared to likelihood-based numerical algorithms. We will explore that further in the accompanying paper (Scott and Sagae, 1997). The structure of these equations makes leave-one-out cross-validation feasible as the specific data values only change the right-hand side of the defining linear equations. We find it interesting that we could explicitly demonstrate that matching local moments up to order k gives the same global higher-order performance of choosing kernels K that have all moments of order $1, \ldots, k-1$ vanish so that the bias is $O(h^k)$ as well. As Terrell and others have noticed, with such a kernel the moments $\int x^i \hat{f}(x)\, dx$ exactly match the sample moments for $i < k$. Of course, matching all local sample moments guarantees that the global sample moments will match as well. We find such parallels interesting.

References:

Bartlett, M.S. (1963). Statistical Estimation of density estimation. *Sankhya Ser.A*, **25**, 245-254.

Good, I.J. and Gaskins, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by the scattering and meteorite data (with discussion). *J. Amer. Statist. Assoc.*, **75**, 42-73.

Hjort, L.N. and Jones, M.C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.*, **24**, 1619-1647.

Loader, C.R. (1996). Local likelihood density estimation. *Ann. Statist.*, **24**, 1602-1618.

Scott, D.W. (1983). Nonparametric probability density estimation for data analysis in several dimensions. *Proceedings of the twenty-eighth conference on the design of experiments in army research development and testing*, pp.387-397.

Scott, D.W. (1985a). Frequency Polygon. *J. Amer. Statist. Assoc.*, **80**, 348-354.

Scott, D.W. (1985b). Average shifted histograms: Effective nonparametric density estimators in several dimensions. *Ann. Statist.*, **13**, 1024-1040.

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

Scott, D.W. and Sagae, M. (1997). In preparation.

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

Silverman, B.W. (1982). Algorithm AS176. Kernel density estimation by using the fast Fourier transformation. *Appl. Statist.*, **31**, 93-99.

Terrell, G.R. (1984). Efficiency of nonparametric density estimators. Technical Report, Department of Statistics, Rice University.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

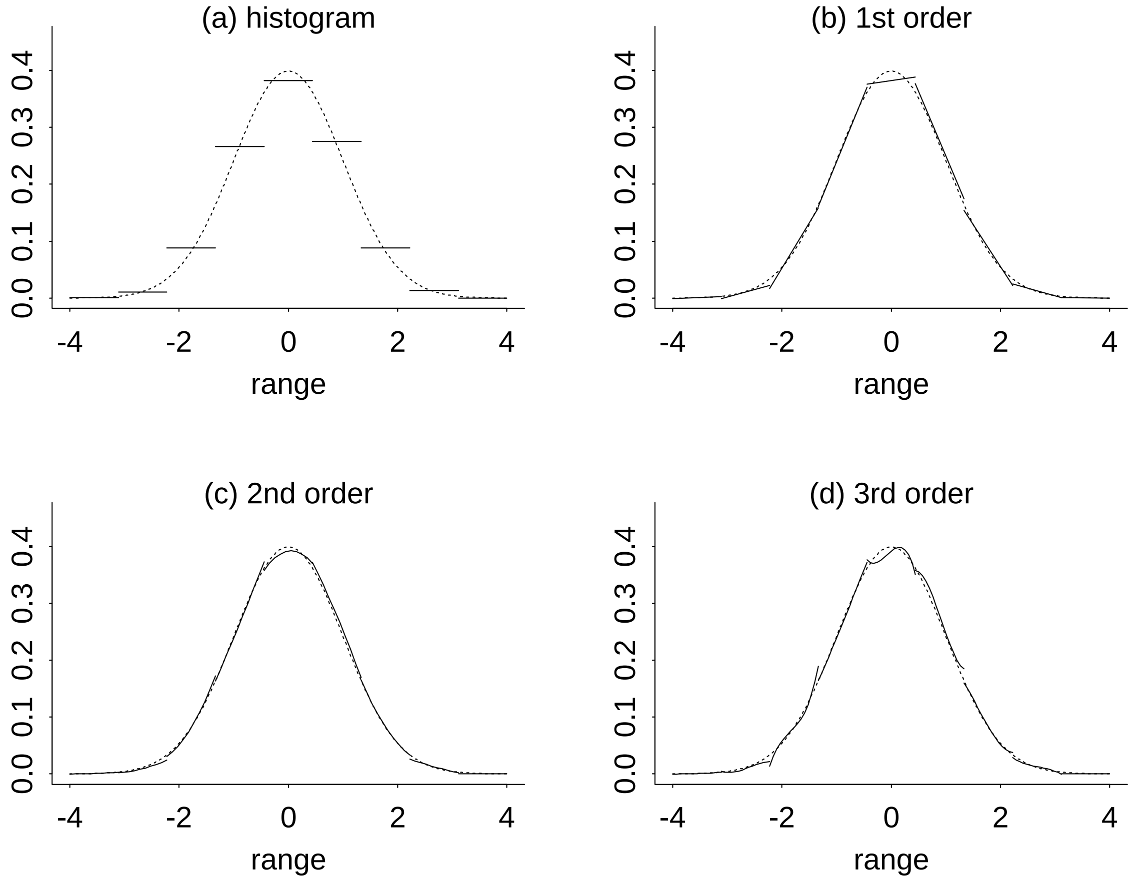

Figure 1: Four polynomial histograms of 10,000 normal observations $N(0,1)$ with the orders k=0,1,2,3

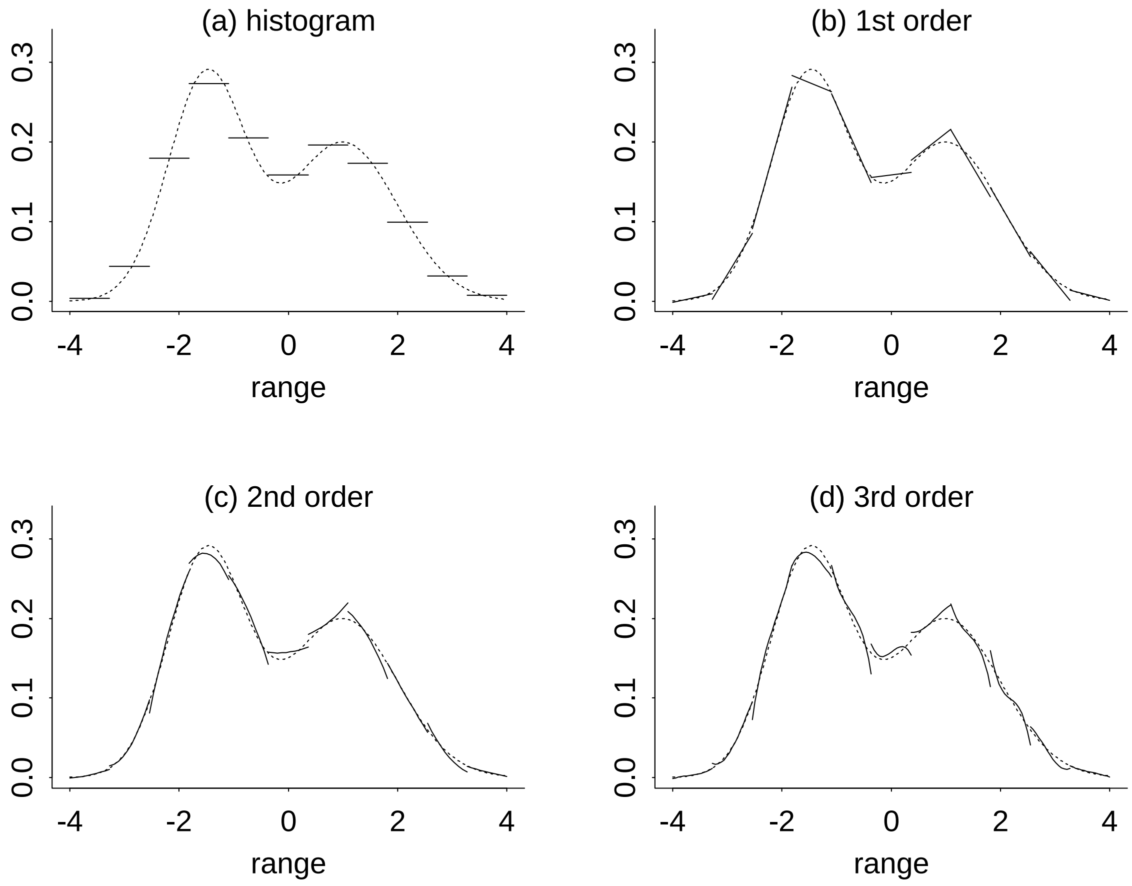

Figure 2: Four polynomial histograms of 10,000 normal mixture observations $\frac{1}{2}N(-1.5, 0.5)$ and $\frac{1}{2}N(1, 1)$ with the orders k=0,1,2,3.

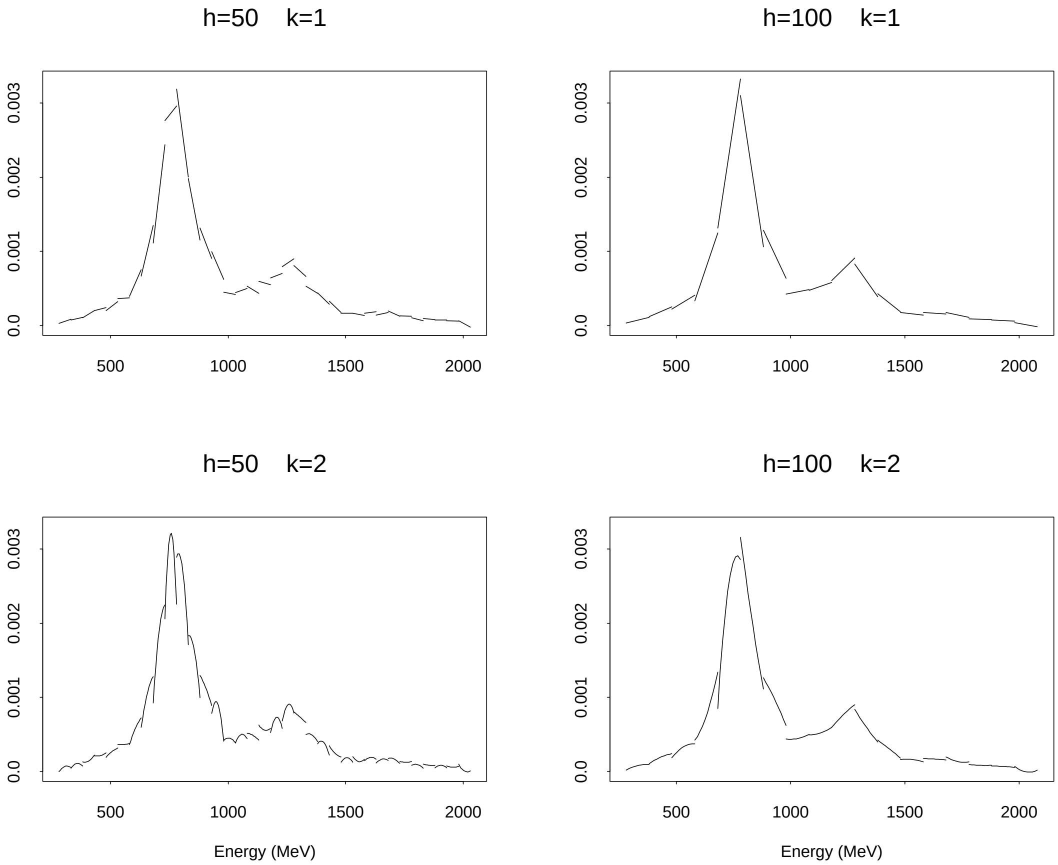

Figure 3: For the LRL particle physics data, two linear and two quadratic polynomial histogram estimates.

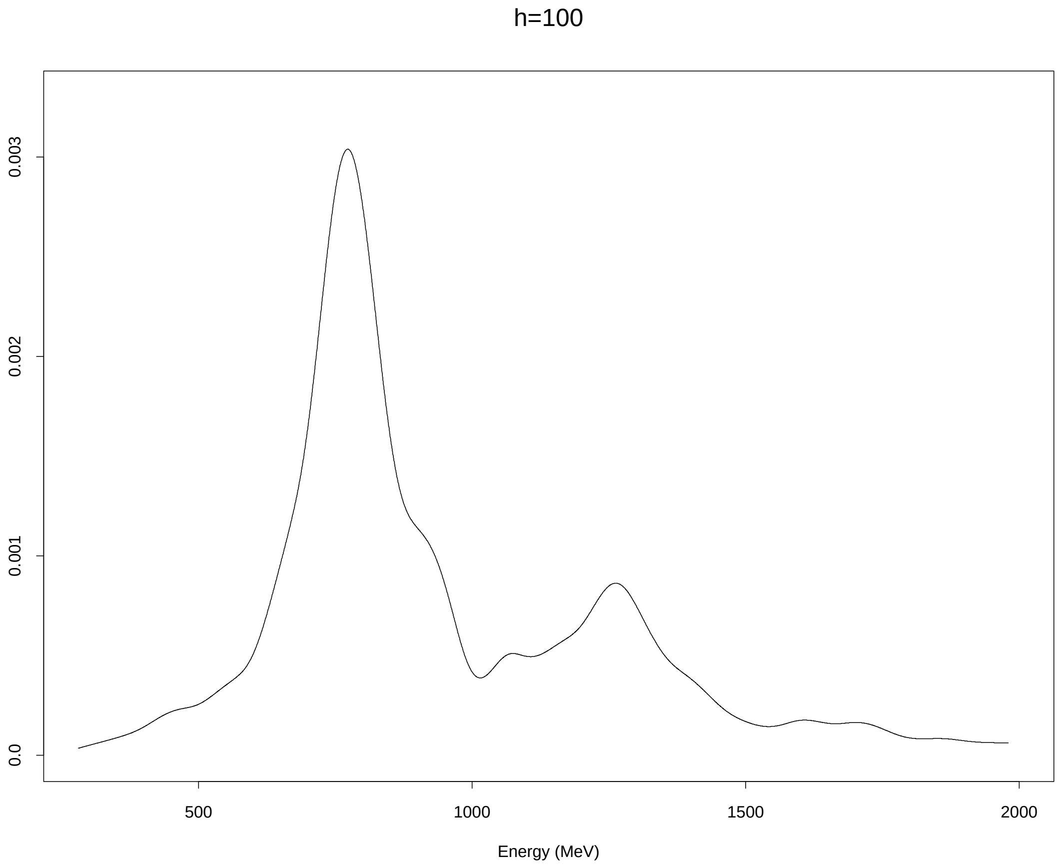

Figure 4: For the LRL particle physics data, a version of the linear polynomial histogram estimate which is both continuous and smoothed by a roughness penalty.