

Adaptive Nonparametric Density Estimators

by

Alan J. Izenman

Introduction

Theoretical results and practical application of histograms as density estimators usually assume a “fixed-partition” approach, where bin boundaries are set out independent of the data. In certain situations, we would instead like to construct a histogram that would be more sensitive of clumpings and gaps in the data.

Variable-Partition Histograms

Variable-partition histograms (Wegman, 1975) are constructed in a similar manner as fixed-partition histograms, but in this case the partition depends upon the gaps between the order statistics, which we denote by $X_{(1)}, \dots, X_{(n)}$. Choose an integer $m \in [2, n]$ to be the number of bins of the histogram and then set $k = \lceil n/m \rceil$. A partition $\mathbf{P} = \{P_{in}\}$ can be obtained by defining $P_{1n} = [X_{(1)}, X_{(k)}], P_{2n} = (X_{(k)}, X_{(2k)}], \dots, P_{mn} = (X_{((m-1)k}), X_{(n)}]$, so that each interval contains about k observations. Then, for any $x \in [X_{(1)}, X_{(n)}]$, estimate p by

$$\hat{p}_m(x) = \sum_{i=1}^m \frac{k/n}{X_{(ik)} - X_{((i-1)k+1)}} I_{P_{in}}(x). \quad (1)$$

Clearly, \hat{p} is constant on the intervals $\{P_{in}\}$ and is, therefore, a histogram-type estimator of p .

In L_1 -theory, the variable-partition histogram is a strongly-consistent estimator of p if $k = k_n \rightarrow \infty$ and $k_n/n \rightarrow 0$ as $n \rightarrow \infty$; similar results hold for L_2 -theory. The rate of convergence of MISE for variable-partition histograms is $O(n^{-2/3})$, the same order as for the fixed-partition case.

Higher-Order Kernels

More exotic kernels have been studied. The most important of such developments concerns a hierarchy of classes of kernels defined by the existence of certain moments of K .

In this scheme, those univariate symmetric kernels K which integrate to unity are called *order-0 kernels*, while *order- s kernels*, for some positive even

integer s , are those order-0 kernels whose first $s - 1$ moments vanish but whose s th moment is finite. That is, $\int x^j K(x) dx = 0$ for $j = 1, 2, \dots, s - 1$, and $\int x^s K(x) dx = \mu_s(K) \neq 0$. Thus, second-order kernels have zero mean and finite variance ($\mu_2(K) = \sigma_K^2$) and include all compactly-supported kernels. Order- s kernels, for $s \geq 3$, have zero variance, which can be achieved only if K takes on negative values. Such kernels have been promoted as a tool for bias reduction and for improving the MISE convergence rate.

For example, if K is an order- s kernel, then the fastest asymptotic rate of MSE convergence of $\hat{p}(x)$ to $p(x)$, pointwise, is $AMSE^* = O(n^{-2s/(2s+1)})$; thus, for a fourth-order kernel, which cannot be nonnegative, the minimum AMSE convergence rate of $\hat{p}(x)$ to $p(x)$ is of order $O(n^{-8/9})$, which is faster than the best such rate, $O(n^{-4/5})$, for nonnegative kernels.

Such asymptotic results may look very attractive, but they only describe the limiting situation. The pre-limiting behavior (i.e., before the asymptotics become operational) of density estimates which use higher-order kernels shows that they have little or no advantage in either simple or complicated situations over the use of nonnegative kernels (Marron and Wand, 1992). Drawbacks to the use of higher-order kernels include: (1) the obvious interpretation difficulties due to negative ordinates; (2) higher-order kernels require huge sample sizes (in the millions of observations) to get a closer approximation of MISE by AMISE than is obtained using nonnegative kernels; (3) if one works with moderate sample sizes, higher-order kernels perform better than nonnegative kernels only in simple situations where there are few, if any, interesting features in the underlying density; and (4) higher-order kernels tend to produce overly-smooth density estimates.

Locally-Adaptive Kernel Density Estimation

The methods for nonparametric density estimation so far described are quite insensitive to local peculiarities in the data, such as data clumping in certain regions and data sparseness in others, particularly in the tails. We now describe attempts at constructing nonparametric density estimators which are supposed to be more sensitive to the clustering of observations.

“Locally-adaptive” density estimators are formulated by either using a fixed window width for all n observations, $\mathbf{X}_i, i = 1, 2, \dots, n$, but changing the window width for each estimation point \mathbf{x} , or by choosing a different window width for each observation \mathbf{X}_i , and then using it to produce a global estimate of the density p .

Nearest-Neighbor Methods

The *nearest-neighbor density estimator* was originally introduced in the context of nonparametric discrimination (Fix and Hodges, 1951).

At a fixed point $\mathbf{x} \in \mathfrak{R}^r$ and for a fixed integer k , let $h_k(\mathbf{x})$ denote the Eu-

clidean distance from \mathbf{x} to its k th nearest neighbor amongst the $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$, and let

$$\text{vol}_k(\mathbf{x}) = c_r [h_k(\mathbf{x})]^r \quad (2)$$

be the volume of the r -dimensional sphere of radius $h_k(\mathbf{x})$, where c_r is the volume of the unit r -dimensional sphere. The k th *nearest-neighbor* (k -NN) *density estimator* is then given by

$$\hat{p}_k(\mathbf{x}) = \frac{k/n}{\text{vol}_k(\mathbf{x})}. \quad (3)$$

Tukey and Tukey (1981, Section 11.3.2) called (3) the *balloon density estimate* of p . It can be written as a kernel density estimator by setting

$$\hat{p}_k(\mathbf{x}) = \frac{1}{n[h_k(\mathbf{x})]^r} \sum_{i=1}^r K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_k(\mathbf{x})}\right), \quad (4)$$

where the smoothing parameter is now k and the kernel K is the rectangular kernel on the unit r -dimensional sphere.

An advantage of the k -NN estimator is that it is always positive, even in regions of sparse data. The k -NN estimator is consistent if $k = k_n \rightarrow \infty$ and $k_n/n \rightarrow 0$ as $n \rightarrow \infty$ (Loftsgaarden and Quesenberry, 1965). Abramson (1984) proposed that in the r -dimensional case, k_n should be chosen proportional to $n^{4/(r+4)}$, the constant of proportionality depending upon \mathbf{x} .

Although the k -NN estimator appears reasonable for estimating a density at a point, it is not particularly successful for estimating the entire density function p . Indeed, the estimator is not a bona fide density because (4) is discontinuous and has an infinite integral due to very heavy tails. Studies (see, e.g., Terrell and Scott, 1992) have shown that the k -NN density estimate cannot be recommended in low-dimensional settings, but, in higher dimensions ($r \geq 3$), the k -NN estimator may be more useful.

Generalized Balloon Density Estimators

If we replace $h_k(\mathbf{x})$ in (4) with a more general scale factor $h(\mathbf{x})$, still depending upon the point \mathbf{x} at which estimation takes place, then we have the *generalized balloon density estimator* of p ,

$$\hat{p}(\mathbf{x}) = \frac{1}{n[h(\mathbf{x})]^r} \sum_{i=1}^r K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h(\mathbf{x})}\right), \quad (5)$$

In the univariate case ($r = 1$), if p has continuous derivatives of order- q at \mathbf{x} , which are also absolutely integrable, then the density estimator (5) with pointwise-optimal window width $h^*(\mathbf{x}) = O(n^{-1/(2q+1)})$, has $\text{AMSE}^*(\mathbf{x}) = O(n^{-2q/(2q+1)})$, with the same rate of convergence in n for AMISE^* (Terrell and Scott, 1992). Unfortunately, Terrell and Scott show that the univariate

generalized balloon estimator is not that much better than the Gaussian fixed-window-width kernel density estimator for estimating p .

A more general form of (5) is obtained by using a nonsingular scaling matrix $\mathbf{H}(\mathbf{x})$, which depends upon the estimation point \mathbf{x} . The resulting generalized balloon density estimator is:

$$\hat{p}_h(\mathbf{x}) = \frac{1}{n|\mathbf{H}(\mathbf{x})|} \sum_{i=1}^n K([\mathbf{H}(\mathbf{x})]^{-1}(\mathbf{x} - \mathbf{X}_i)), \quad (6)$$

where K is a multivariate function having mean $\mathbf{0}$ and covariance matrix \mathbf{I}_r . This definition coincides with (5) if $\mathbf{H}(\mathbf{x}) = h(\mathbf{x})\mathbf{I}_r$.

The multivariate k -NN density estimator can be obtained by substituting $\mathbf{H}_k = h_k \mathbf{A}$ for $\mathbf{H}(\mathbf{x})$ in (6), where $|\mathbf{A}| = 1$, and $(\mathbf{x} - \mathbf{X}_i)^\tau (\mathbf{H}_k \mathbf{H}_k)^{-1} (\mathbf{x} - \mathbf{X}_i) \leq 1$ is the smallest sphere which contains k observations.

Variable-Kernel Estimators

The *variable-kernel density estimator* attempts to avoid the problems associated with the k -NN estimator. It is defined as

$$\hat{p}_{h,k}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{H_{ik}^r} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{H_{ik}}\right), \quad (7)$$

where the variable window-width $H_{ik} \propto h_k(\mathbf{X}_i)$ now depends upon the observation \mathbf{X}_i and not upon the estimation point \mathbf{x} , and k controls the local behavior of H_{ik} . The estimator (7) is a bona fide density for an appropriate kernel K .

The variable-kernel density estimator was apparently first considered by Meisel in 1973 in the context of pattern recognition, and was then studied empirically by Breiman, Meisel, and Purcell (1977), who listed its advantages as having the smoothness properties of kernel estimators, the data-adaptive character of the k -NN approach, and very little computational penalty. In their simulation studies, however, the estimator (7) was found to perform very poorly unless k was large, on the order of $n/10$.

Adaptive-Kernel Estimators

One way of generalizing (7) is to replace H_{ik} by a more general scale factor $h_i = h(\mathbf{X}_i)$; this yields the *adaptive-kernel density estimator*,

$$\hat{p}_h(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i^r} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_i}\right). \quad (8)$$

Various studies have proposed methods for defining the adaptive window widths h_i .

All proposals start by computing a “pilot” (nonadaptive) kernel density estimate, p_h^0 , with fixed window width h . (Some definitions add a second step in which the pilot density is “clipped” away from zero, but this step is not usually adopted in practice.) Then, h_i is defined according to any of the following proposals:

- $h_i \propto [p_h^0(\mathbf{X}_i)]^{-1/r}$ (Breiman, Meisel, and Purcell, 1977);
- $h_i = h[p_h^0(\mathbf{X}_i)]^{-1/2}$ (Abramson, 1982a,b);
- $h_i = hg^{-\alpha}[p_h^0(\mathbf{X}_i)]^\alpha$, where g is a scale factor (e.g., the geometric mean of the $p_h^0(\mathbf{X}_i)$, $i = 1, 2, \dots, n$) and $\alpha \in [0, 1]$ reflects the sensitivity of the window width to variations in the pilot estimate (Silverman, 1986, Section 5.3);
- $h_i = h_F[p_h^0(\mathbf{X}_i)]^{-1/2}$, where h_F is the window width of the final estimate (Hall and Marron, 1988).

One of the most popular of these suggestions has been that of Abramson, who justified his choice of window width by focussing on pointwise estimation. Abramson showed that his adaptive choice of window width (which did not depend upon the dimensionality, r , of the data) performed better (in terms of MSE) than the fixed (nonadaptive) window width for pointwise kernel density estimation. Unfortunately, it has been difficult, if not impossible, to extend Abramson’s “square-root” rule to global estimation of the entire density function.

FP and ASH

By modifying the block-like shape of the histogram, a faster rate of IMSE convergence to $O(n^{-4/5})$ (or close to it) can be reached. This has been done in a number of ways, including the *frequency polygon (FP)* (Scott, 1985b) and the *average shifted histogram (ASH)* (Scott and Thompson, 1983; Scott, 1985a).

The Frequency Polygon

The FP connects the center of each pair of adjacent histogram bin-values with a straight line. If two adjacent bin-values are $\hat{p}_\ell = N_\ell/nh_n$ and $\hat{p}_{\ell+1} = N_{\ell+1}/nh_n$, then the value of the FP at $x \in [(\ell - \frac{1}{2})h_n, (\ell + \frac{1}{2})h_n]$ is

$$\hat{p}_{\text{FP}}(x) = \left(\left(\ell + \frac{1}{2} \right) - \frac{x}{h_n} \right) \hat{p}_\ell + \left(\frac{x}{h_n} - \left(\ell - \frac{1}{2} \right) \right) \hat{p}_{\ell+1}. \quad (9)$$

While the histogram is discontinuous, the FP is a continuous density estimator. Under certain continuity and integrability conditions on the derivatives of p , the optimal bin width for the FP is $h_n^* = O(n^{-1/5})$, which is wider than that for the histogram, which has optimal bin width $O(n^{-1/3})$. For p a Gaussian density,

$\mathcal{N}(0, \sigma^2)$, the optimal bin width becomes $h_n^* = 2.15\sigma n^{-1/5}$. The optimal FP has $\text{AMISE}^* = O(n^{-4/5})$, which is a substantial improvement over the $O(n^{-2/3})$ rate of the histogram.

The Average Shifted Histogram

The ASH, which was motivated by the need to resolve the problem of choice of bin origin for the histogram and frequency polygon, is constructed by taking m histograms, $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_m$, say, each of which has the same bin width h_n , but with different bin origins, $0, h_n/m, 2h_n/m, \dots, (m-1)h_n/m$, respectively, and then averaging those histograms,

$$\hat{p}_{\text{ASH}}(x) = m^{-1} \sum_{k=1}^m \hat{p}_k(x). \quad (10)$$

The resulting ASH is piecewise constant over intervals $[k\delta, (k+1)\delta)$ of width $\delta = h_n/m$; it has a similar block-like structure as a histogram, but is defined over narrower bins. The most important choice for ASH as a density estimator is h_n ; choice of m does not appear as important. For large m , the optimal bin width for the ASH is $h^* = O(n^{-1/5})$. See Exercise 4.4 for the AMISE. A weighted version of the ASH has also been studied. Just like the histogram, the ASH is also not continuous; however, it can be made continuous by linearly interpolating the ASH using the FP approach to yield an FP-ASH density estimator.

The FP, ASH, and FP-ASH modifications of the histogram have been generalized to two and higher dimensions. See Scott (1992, Chapters 4, 5) for details.

Orthogonal Series Estimators

Orthogonal series density estimators (Cencov, 1962) have been applied to several different areas, especially pattern recognition and classification problems.

Arbitrary Orthogonal Expansions

The method assumes that a square-integrable function p can be represented as a convergent orthogonal series expansion,

$$p(x) = \sum_{k=-\infty}^{\infty} a_k \psi_k(x), \quad x \in \Omega, \quad (11)$$

where $\{\psi_k\}$ is a complete orthogonal system of functions on a set Ω of the real line. That is,

$$\int_{\Omega} \psi_j^*(x) \psi_k(x) dx = \delta_{jk}, \quad (12)$$

where the Kronecker delta $\delta_{jk} = 1$ or 0 according as $j = k$ or $j \neq k$, respectively. The coefficients $\{a_k\}$ are defined by

$$a_k = \mathbb{E}_p\{\psi_k^*(X)\} = \int_{\Omega} \psi_k^*(x)p(x)dx, \quad (13)$$

where ψ_k^* is the complex conjugate of ψ_k . This formulation allows for systems of real- or complex-valued orthonormal functions.

Orthonormal systems proposed for $\{\psi_k\}$ are those with compact support (e.g., Fourier, trigonometric, and Haar systems on $[0, 1]$, and Legendre systems on $[-1, 1]$) and those with unbounded support (e.g., Hermite system on \Re and Laguerre system on $[0, \infty)$). For a given system $\{\psi_k\}$, the problem is to estimate the density p by estimating the coefficients $\{a_k\}$.

The *Hermite series estimator* is the most popular orthogonal series estimator for densities with unbounded support, usually \Re or $[0, \infty)$. Let $D^k = d^k/dx^k$. The Hermite series is derived from the result that $D^k e^{-x^2} = (-1)^k H_k(x)e^{-x^2}$, where the k th Hermite polynomial, $H_k(x)$, is defined by

$$H_k(x) = (-1)^k e^{x^2} D^k e^{-x^2}, \quad k = 0, 1, 2, \dots \quad (14)$$

The first few Hermite polynomials are listed in Table 1 and $H_k(x)$, $k = 1, 2, 3, 4$, are graphed in Figure 1. The Hermite polynomials satisfy the symmetry condition $H_k(-x) = (-1)^k H_k(x)$. The tails of the Hermite functions $\psi_k(x)$ are

Table 1: *The first eight Hermite polynomials, $H_k(x)$, $k = 1, 2, \dots, 8$.*

k	$H_k(x)$
0	1
1	$2x$
2	$4x^2 - 2$
3	$8x^3 - 12x$
4	$16x^4 - 48x^2 + 12$
5	$32x^5 - 160x^3 + 120x$
6	$64x^6 - 480x^4 + 720x^2 - 120$
7	$128x^7 - 1344x^5 + 3360x^3 - 1680x$
8	$256x^8 - 3584x^6 + 13440x^4 - 13440x^2 + 1680$

heavily weighted by $e^{-x^2/2}$, which protects them against unusual behavior in the tails of X . The normalized Hermite functions,

$$\psi_k(x) = \frac{e^{-x^2/2}}{(2^k k! \pi^{1/2})^{1/2}} H_k(x), \quad k = 0, 1, 2, \dots, \quad (15)$$

satisfy the orthogonality conditions (12) and so form an orthonormal basis for an L_2 approach.

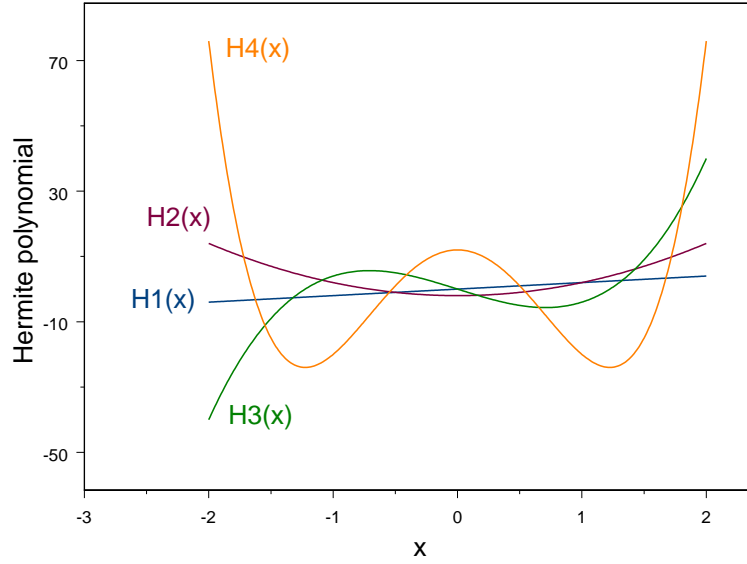


Figure 1: Graphs of the Hermite polynomials, $H_k(x)$, for $k = 1, 2, 3, 4$.

If p has compact support $[0, 1]$, say, the *Fourier series estimator* is formed from the system of discrete Fourier functions,

$$\psi_k(x) = e^{2\pi i k x}, \quad i = \sqrt{-1}, \quad k = 0, 1, 2, \dots, \quad (16)$$

and is the most popular choice of orthonormal basis to use in an estimator of the form (18) below.

Estimation Strategies

Given an iid sample, X_1, X_2, \dots, X_n , from p and a system $\{\psi_k\}$, the $\{a_k\}$ can be unbiasedly estimated by

$$\hat{a}_k = n^{-1} \sum_{i=1}^n \psi_k^*(X_i). \quad (17)$$

The obvious estimator of p , obtained by plugging (17) into (11) in place of a_k , may not be well-defined: it has infinite variance and is not consistent in the ISE sense.

A better strategy is to use a tapered estimator of the form

$$\hat{p}(x) = \sum_{k=-\infty}^{\infty} b_k \hat{a}_k \psi_k(x), \quad x \in \Omega, \quad (18)$$

where $0 < b_k < 1$ is a symmetric weight ($b_{-k} = b_k$) which shrinks \hat{a}_k towards the origin, and $\sum_k |b_k| < \infty$ is needed for pointwise convergence of (4.84). Substituting (17) into the tapered estimator (18) yields

$$\hat{p}(x) = n^{-1} \sum_{i=1}^n \sum_{k=-\infty}^{\infty} b_k \psi_k^*(X_i) \psi_k(x). \quad (19)$$

For example, if we substitute the Fourier series (16) into (19), we have

$$\hat{p}(x) = n^{-1} \sum_{i=1}^n \left[\sum_{k=-\infty}^{\infty} b_k e^{2\pi i k(x-X_i)} \right] = n^{-1} \sum_{i=1}^n K(x - X_i), \quad (20)$$

where

$$K(x) = \sum_{k=-\infty}^{\infty} b_k e^{2\pi i k x}, \quad (21)$$

which has the form of a convolution of a kernel estimator with the filter $\{b_k\}$. If we set $b_k = 1$ for $-r \leq k \leq r$, and 0 otherwise (usually known as the *boxcar filter*), then the real part of (20) reduces to

$$\hat{p}_r(x) = n^{-1} \sum_{i=1}^n \frac{\sin\{\pi(2r+1)(x - X_i)\}}{\sin(\pi(x - X_i))}. \quad (22)$$

The summand in (22) is known as the *sinc function* or *Dirichlet kernel*. Use of the boxcar filter in (18) leads to the partial sums of (11) being approximated by

$$\hat{p}_r(x) = \sum_{k=-r}^r \hat{a}_k \psi_k(x), \quad x \in \Omega, \quad (23)$$

where the $\{\hat{a}_k\}$ are given by (17). More sophisticated weighting systems have been considered, including a two-parameter system of weights,

$$b_k = b_k(\lambda, m) = \frac{1}{1 + \lambda(2\pi k)^{2m}}, \quad -r \leq k \leq r, \quad (24)$$

where $\lambda > 0$ is a smoothing parameter and $m > 1/2$ is a shape parameter (Wahba, 1977). To estimate the $\{b_k\}$, one can use likelihood cross-validation (Wahba, 1981) or least-squares cross-validation (Hall, 1987b).

Asymptotics

If p has unbounded support and if $r = r_n$ satisfies $r_n/n \rightarrow 0$ as $r_n \rightarrow \infty$, then $\text{MISE} \rightarrow 0$ as $n \rightarrow \infty$; moreover, if $r_n = O(n^{1/q})$ for $q \geq 2$, then $\text{MISE} = O(n^{-(1-1/q)})$ (Schwartz, 1967). Note that the MISE convergence rate is independent of the dimension of the data, which gives the Hermite series estimator an advantage over the kernel estimator for multivariate density estimation.

The Hermite system does not provide a basis for the L_1 approach, however, and the Hermite series estimator is neither translation invariant nor consistent in the L_1 sense. These properties led Devroye and Györfi (1985, p. 314) to conclude that “the Hermite series estimate seems ill-suited as a general-purpose density estimate on the real line.”

Studies of Fourier series density estimators have revealed effects of periodicity and the Gibbs phenomenon. For the Fourier series estimator, under suitable conditions on p , if $r_n/n \rightarrow 0$ as $r_n \rightarrow \infty$, then $\text{MIAE} \rightarrow 0$ as $n \rightarrow \infty$ (Devroye and Györfi, 1985, Section 12.4).

Choice of Number of Terms

The performance and smoothness of the orthogonal series density estimator (4.92) depend upon r , the number of terms in the expansion. The most well-known criterion for choosing r involves a term-by-term optimal stopping rule which minimizes an estimated MISE criterion (Kronmal and Tarter, 1968); however, there are problems with such a strategy: the rule may not yield the optimal r , it tends to stop too soon, thus yielding oversmoothed density estimates, and it gives poor performance in multimodal situations.