

Nonparametric Estimation of Probability Density Functions

Estimation of a probability density function is similar to the estimation of any function, and the properties of the function estimators that we have discussed are relevant for density function estimators. A density function $p(y)$ is characterized by two properties:

- it is nonnegative everywhere;
- it integrates to 1 (with the appropriate definition of “integrate”).

Bona Fide Density Estimator

It seems reasonable that we require the density estimate to have the characteristic properties of a density:

- $\hat{p}(y) \geq 0$ for all y ;
- $\int_{\mathbb{R}^d} \hat{p}(y) \, dy = 1$.

A probability density estimator that is nonnegative and integrates to 1 is called a *bona fide* estimator.

Maximum Likelihood Estimation

Given a random sample, y_1, \dots, y_n , from a population with density p . The likelihood functional is

$$L(p; y_1, \dots, y_n) = \prod_{i=1}^n p(y_i).$$

The density p itself is a variable.

The *maximum likelihood method* of estimation obviously cannot be used directly because this functional is unbounded in p .

Maximum Likelihood Estimation

We may seek an estimator that maximizes some modification of the likelihood. There are two reasonable ways to approach this problem.

One is to restrict the domain of the optimization problem. This is called *restricted maximum likelihood*.

The other is to *regularize* the estimator by adding a penalty term to the functional to be optimized. This is called *penalized maximum likelihood*.

Restricted Maximum Likelihood Estimation

We may seek to maximize the likelihood functional subject to the constraint that p be a bona fide density.

If we put no further restrictions on the function p , however, infinite Dirac spikes at each observation give an unbounded likelihood, so a maximum likelihood estimator cannot exist, subject only to the restriction to the bona fide class.

An additional restriction that p be Lebesgue-integrable over some domain D (that is, $p \in L^1(D)$) does not resolve the problem because we can construct sequences of finite spikes at each observation that grow without bound.

We therefore must restrict the class further.

Restricted Maximum Likelihood Estimation

Consider a finite dimensional class, such as the class of step functions that are bona fide density estimators. We assume that the sizes of the regions over which the step function is constant are greater than 0.

For a step function with m regions having constant values, c_1, \dots, c_m , the likelihood is

$$\begin{aligned} L(c_1, \dots, c_m; y_1, \dots, y_n) &= \prod_{i=1}^n p(y_i) \\ &= \prod_{k=1}^m c_k^{n_k}, \end{aligned}$$

where n_k is the number of data points in the k^{th} region. For the step function to be a bona fide estimator, all c_k must be nonnegative and finite.

A maximum therefore exists in the class of step functions that are bona fide estimators.

Restricted Maximum Likelihood Estimation

If v_k is the measure of the volume of the k^{th} region (that is, v_k is the length of an interval in the univariate case, the area in the bivariate case, and so on), we have

$$\sum_{k=1}^m c_k v_k = 1.$$

We incorporate this constraint to form the Lagrangian,

$$L(c_1, \dots, c_m) + \lambda \left(1 - \sum_{k=1}^m c_k v_k \right).$$

Differentiating the Lagrangian function and setting the derivative to zero, we have at the maximum point $c_k = c_k^*$, for any λ ,

$$\frac{\partial L}{\partial c_k} = \lambda v_k.$$

Using the derivative of L , we get

$$n_k L = \lambda c_k^* v_k.$$

Restricted Maximum Likelihood Estimation, Continued

Summing both sides of this equation over k , we have

$$nL = \lambda,$$

and then substituting, we have

$$n_k L = nL c_k^* v_k.$$

Therefore, the maximum of the likelihood occurs at

$$c_k^* = \frac{n_k}{n v_k}.$$

The restricted maximum likelihood estimator is therefore

$$\begin{aligned} \hat{p}(y) &= \frac{n_k}{n v_k}, \quad \text{for } y \in \text{region } k, \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

Restricted Maximum Likelihood Estimation

Instead of restricting the density estimate to step functions, we could consider other classes of functions, such as piecewise linear functions.

We may also seek other properties, such as smoothness, for the estimated density. One way of achieving other desirable properties for the estimator is to use a penalizing function to modify the function to be optimized. Instead of the likelihood function, we may use a penalized likelihood function of the form

$$L_p(p; y_1, \dots, y_n) = \prod_{i=1}^n p(y_i) e^{-\mathcal{T}(p)},$$

where $\mathcal{T}(p)$ is some transform that measures a property that we would like to minimize.

For example, to achieve smoothness, we may use the transform $\mathcal{R}(p)$ in the penalizing factor.

Histogram Estimators

Let us assume finite support D , and construct a fixed partition of D into a grid of m nonoverlapping bins T_k . (We can arbitrarily assign bin boundaries to one or the other bin.) Let v_k be the volume of the k^{th} bin (in one dimension, v_k is a length and in this simple case is often denoted h_k ; in two dimensions, v_k is an area, and so on).

The number of such bins we choose, and consequently their volumes, depends on the sample size n , so we sometimes indicate that dependence in the notation: $v_{n,k}$. For the sample y_1, \dots, y_n , the histogram estimator of the probability density function is defined as

$$\begin{aligned}\hat{p}_H(y) &= \sum_{k=1}^m \frac{1}{v_k} \frac{\sum_{i=1}^n \mathbf{I}_{T_k}(y_i)}{n} \mathbf{I}_{T_k}(y), \quad \text{for } y \in D, \\ &= 0, \quad \text{otherwise.}\end{aligned}$$

The histogram is the restricted maximum likelihood estimator discussed before.

Histogram Estimators

Letting n_k be the number of sample values falling into T_k ,

$$n_k = \sum_{i=1}^n \mathbf{I}_{T_k}(y_i),$$

we have the simpler expression for the histogram over D ,

$$\hat{p}_H(y) = \sum_{k=1}^m \frac{n_k}{nv_k} \mathbf{I}_{T_k}(y).$$

This is a bona fide estimator:

$$\hat{p}_H(y) \geq 0$$

and

$$\begin{aligned} \int_{\mathbb{R}^d} \hat{p}_H(y) dy &= \sum_{k=1}^m \frac{n_k}{nv_k} v_k \\ &= 1. \end{aligned}$$

Histogram Estimators

In the univariate case, we assume that the support is the finite interval $[a, b]$. We partition $[a, b]$ into a grid of m nonoverlapping bins $T_k = [t_{n,k}, t_{n,k+1})$ where

$$a = t_{n,1} < t_{n,2} < \dots < t_{n,m+1} = b.$$

The univariate histogram is

$$\hat{p}_H(y) = \sum_{k=1}^m \frac{n_k}{n(t_{n,k+1} - t_{n,k})} \mathbf{I}_{T_k}(y).$$

If the bins are of equal width, say h (that is, $t_k = t_{k-1} + h$), the histogram is

$$\hat{p}_H(y) = \frac{n_k}{nh}, \quad \text{for } y \in T_k.$$

Some Properties of the Histogram Estimator

The histogram estimator, being a step function, is discontinuous at cell boundaries, and it is zero outside of a finite range.

An important advantage of the histogram estimator is its simplicity, both for computations and for analysis. In addition to its simplicity, as we have seen, it has two other desirable global properties:

- It is a bona fide density estimator.
- It is the unique maximum likelihood estimator confined to the subspace of functions of the form

$$\begin{aligned}g(t) &= c_k, \text{ for } t \in T_k, \\ &= 0, \text{ otherwise,}\end{aligned}$$

and where $g(t) \geq 0$ and $\int_{\cup_{k=1}^m T_k} g(t) dt = 1$.

Pointwise and Binwise Properties

Properties of the histogram vary from bin to bin.

We see that the number in the k^{th} bin, n_k , is a binomial random variable with parameters n and p_k , where

$$p_k = \int_{T_k} p(t) dt$$

is the probability content of the k^{th} bin.

The **expectation** of the histogram estimator at the point y in bin T_k is

$$E(\hat{p}_H(y)) = \frac{p_k}{v_k},$$

and the **variance** of the histogram at the point y within the k^{th} bin is

$$\begin{aligned} V(\hat{p}_H(y)) &= V(n_k)/(nv_k)^2 \\ &= \frac{p_k(1-p_k)}{nv_k^2}. \end{aligned}$$

Pointwise Bias

The **bias** of the histogram at the point y within the k^{th} bin is

$$\frac{p_k}{v_k} - p(y).$$

Note that the bias at y is a function of y , and it is different from bin to bin, even if the bins are of constant size.

The bias tends to decrease as the bin size decreases.

Pointwise Bias

We can bound the bias if we assume a regularity condition on p .

If there exists γ such that for any $y_1 \neq y_2$ in an interval

$$|p(y_1) - p(y_2)| < \gamma \|y_1 - y_2\|,$$

we say that p is Lipschitz-continuous on the interval, and for such a density, for any ξ_k in the k^{th} bin, we have the bound

$$\begin{aligned} |\text{Bias}(\hat{p}_H(y))| &= |p(\xi_k) - p(y)| \\ &\leq \gamma_k \|\xi_k - y\| \\ &\leq \gamma_k v_k. \end{aligned}$$

Pointwise Variance

The variance of the histogram at the point y within the k^{th} bin is

$$\mathbb{V}(\hat{p}_H(y)) = \frac{p_k(1 - p_k)}{nv_k^2}.$$

Notice that the variance decreases as the bin size increases. Note also that the variance is different from bin to bin. We can bound the variance:

$$\mathbb{V}(\hat{p}_H(y)) \leq \frac{p_k}{nv_k^2}.$$

By the mean-value theorem, we have $p_k = v_k p(\xi_k)$ for some $\xi_k \in T_k$, so we can write

$$\mathbb{V}(\hat{p}_H(y)) \leq \frac{p(\xi_k)}{nv_k}.$$

Tradeoffs

Notice the tradeoff between bias and variance:

as h increases,

the variance decreases, but

the bound on the bias increases.

Pointwise Mean Squared Error

The **mean squared error** of the histogram at the point y within the k^{th} bin is

$$\text{MSE}(\hat{p}_H(y)) = \frac{p_k(1 - p_k)}{nv_k^2} + \left(\frac{p_k}{v_k} - p(y) \right)^2.$$

For a Lipschitz-continuous density, within the k^{th} bin we have

$$\text{MSE}(\hat{p}_H(y)) \leq \frac{p(\xi_k)}{nv_k} + \gamma_k^2 v_k^2.$$

Pointwise Mean Squared Error

We easily see that the histogram estimator is L_2 pointwise consistent for a Lipschitz-continuous density if, as $n \rightarrow \infty$, for each k , $v_k \rightarrow 0$ and $nv_k \rightarrow \infty$.

By differentiating, we see that the minimum of the bound on the MSE in the k^{th} bin occurs for

$$h^*(k) = \left(\frac{p(\xi_k)}{2\gamma_k^2 n} \right)^{1/3}.$$

Substituting this value back into MSE, we obtain the order of the optimal MSE at the point x ,

$$\text{MSE}^*(\hat{p}_H(y)) = O(n^{-2/3}).$$

Asymptotic MISE (or AMISE) of Histogram Estimators

Global properties of the histogram are obtained by summing the binwise properties over all of the bins.

The integrated variance and the integrated squared bias depend on the bin sizes and the probability content of the bins.

We will first write the general expressions, and then we will assume some degree of smoothness of the true density and write approximate expressions that result from mean values or Taylor approximations.

We will assume rectangular bins for additional simplification. Finally, we will then consider bins of equal size to simplify the expressions further.

Integrated Variance

First, consider the integrated variance,

$$\begin{aligned}\text{IV}(\hat{p}_H) &= \int_{\mathbb{R}^d} \mathbb{V}(\hat{p}_H(t)) dt \\ &= \sum_{k=1}^m \int_{T_k} \mathbb{V}(\hat{p}_H(t)) dt \\ &= \sum_{k=1}^m \frac{p_k - p_k^2}{nv_k} \\ &= \sum_{k=1}^m \left(\frac{1}{nv_k} - \frac{\sum p(\xi_k)^2 v_k}{n} \right) + o(n^{-1})\end{aligned}$$

for some $\xi_k \in T_k$, as before.

Integrated Variance, Continued

Now, taking $\sum p(\xi_k)^2 v_k$ as an approximation to the integral $\int (p(t))^2 dt$, and letting \mathcal{S} be the functional that measures the variation in a square-integrable function of d variables,

$$\mathcal{S}(g) = \int_{\mathbb{R}^d} (g(t))^2 dt,$$

we have the integrated variance,

$$\text{IV}(\hat{p}_H) \approx \sum_{k=1}^m \frac{1}{nv_k} - \frac{\mathcal{S}(p)}{n},$$

and the asymptotic integrated variance,

$$\text{AIV}(\hat{p}_H) = \sum_{k=1}^m \frac{1}{nv_k}.$$

The measure of the variation, $\mathcal{S}(p)$, is a measure of the roughness of the density because the density integrates to 1.

Integrated Squared Bias

Now, consider the other term in the integrated MSE, the integrated squared bias. We will consider the case of rectangular bins, in which $h_k = (h_{k_1}, \dots, h_{k_d})$ is the vector of lengths of sides in the k^{th} bin. In the case of rectangular bins, $v_k = \prod_{j=1}^d h_{k_j}$.

We assume that the density can be expanded in a Taylor series, and we expand the density in the k^{th} bin about \bar{t}_k , the midpoint of the rectangular bin. For $\bar{t}_k + t \in T_k$, we have

$$p(\bar{t}_k + t) = p(\bar{t}_k) + t^\top \nabla p(\bar{t}_k) + \frac{1}{2} t^\top H_p(\bar{t}_k) t + \dots,$$

where $H_p(\bar{t}_k)$ is the Hessian of p evaluated at \bar{t}_k .

Integrated Squared Bias, Continued

The probability content of the k^{th} bin, p_k can be expressed as an integral of the Taylor series expansion:

$$\begin{aligned} p_k &= \int_{\bar{t}_k + t \in T_k} p(\bar{t}_k + t) dt \\ &= \int_{-h_{kd}/2}^{h_{kd}/2} \cdots \int_{-h_{k1}/2}^{h_{k1}/2} \left(p(\bar{t}_k) + t^\top \nabla p(\bar{t}_k) + \dots \right) dt_1 \cdots dt_d \\ &= v_k p(\bar{t}_k) + \mathcal{O}(h_{k*}^{d+2}), \end{aligned}$$

where $h_{k*} = \min_j h_{kj}$. The bias at a point $\bar{t}_k + t$ in the k^{th} bin is

$$\frac{p_k}{v_k} - p(\bar{t}_k + t) = -t^\top \nabla p(\bar{t}_k) + \mathcal{O}(h_{k*}^2).$$

Integrated Squared Bias, Continued

For the k^{th} bin the integrated squared bias is

$$\begin{aligned} \text{ISB}_k(\hat{p}_H) &= \int_{T_k} \left((t^\top \nabla p(\bar{t}_k))^2 - 2\mathcal{O}(h_{k*}^2) t^\top \nabla p(\bar{t}_k) + \mathcal{O}(h_{k*}^4) \right) dt \\ &= \int_{-h_{kd}/2}^{h_{kd}/2} \cdots \int_{-h_{k1}/2}^{h_{k1}/2} \sum_i \sum_j t_{ki} t_{kj} \nabla_i p(\bar{t}_k) \nabla_j p(\bar{t}_k) dt_1 \cdots dt_d + \mathcal{O}(h_{k*}^{4+d}). \end{aligned}$$

Integrated Variance and Squared Bias for Constant Bin Sizes

Many of the expressions above are simpler if we use a constant bin size, v , or h_1, \dots, h_d . In the case of constant bin size, the asymptotic integrated variance becomes

$$\text{AIV}(\hat{p}_H) = \frac{m}{nv}.$$

In this case, the integral simplifies as the integration is performed term by term because the cross-product terms cancel, and the integral is

$$\frac{1}{12}(h_1 \cdots h_d) \sum_{j=1}^d h_j^2 (\nabla_j p(\bar{t}_k))^2.$$

This is the asymptotic squared bias integrated over the k^{th} bin.

Integrated Variance and Squared Bias for Constant Bin Sizes

When we sum over all bins, the $(\nabla_j p(\bar{t}_k))^2$ become $\mathcal{S}(\nabla_j p)$, and we have the asymptotic integrated squared bias,

$$\text{AISB}(\hat{p}_H) = \frac{1}{12} \sum_{j=1}^d h_j^2 \mathcal{S}(\nabla_j p).$$

Combining the asymptotic integrated variance and squared bias for the histogram with rectangular bins of constant size, we have

$$\text{AMISE}(\hat{p}_H) = \frac{1}{n(h_1 \cdots h_d)} + \frac{1}{12} \sum_{j=1}^d h_j^2 \mathcal{S}(\nabla_j p).$$

As we have seen before, smaller bin sizes increase the variance but decrease the squared bias.

Bin Sizes

The histogram is very sensitive to the bin sizes, both in appearance and in other properties.

The AMISE assuming constant rectangular bin size is often used as a guide for determining the bin size to use when constructing a histogram.

This expression involves $\mathcal{S}(\nabla_j p)$ and so, of course, cannot be used directly. Nevertheless, differentiating the expression for AMISE with respect to h_j and setting the result equal to zero, we have the bin width that is optimal with respect to the AMISE,

$$h_{j*} = \mathcal{S}(\nabla_j p)^{-1/2} \left(6 \prod_{i=1}^d \mathcal{S}(\nabla_i p)^{1/2} \right)^{\frac{1}{2+d}} n^{-\frac{1}{2+d}}.$$

Bin Sizes

Plugging the optimal bin width back into the expression for the AMISE, we get the optimal value of the AMISE

$$\frac{1}{4} \left(36 \prod_{i=1}^d \mathcal{S}(\nabla_i p)^{1/2} \right)^{\frac{1}{2+d}} n^{-\frac{2}{2+d}}.$$

Notice that the optimal rate of decrease of AMISE for histogram estimators is $O(n^{-\frac{2}{2+d}})$.

Although histograms have several desirable properties, this order of convergence is not good compared to that of some other bona fide density estimators.

Bin Sizes

The expression for the optimal bin width involves $\mathcal{S}(\nabla_j p)$, where p is the unknown density.

An approach is to choose a value for $\mathcal{S}(\nabla_j p)$ that corresponds to some good general distribution.

Bin Sizes

A “good general distribution”, of course, is the normal with a diagonal variance-covariance matrix. For the d -variate normal with variance-covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$,

$$\mathcal{S}(\nabla_j p) = \frac{1}{2^{d+1} \pi^{d/2} \sigma_j^2 |\Sigma|^{1/2}}.$$

For a univariate normal density with variance σ^2 ,

$$\mathcal{S}(p') = 1/(4\sqrt{\pi}\sigma^3),$$

so the optimal constant one-dimensional bin width under the AMISE criterion is

$$3.49\sigma n^{-1/3}.$$

Bin Sizes

The more robust estimate of the scale based on the sample interquartile range, r leads to a bin width of $2rn^{-1/3}$.

The AMISE is essentially an L_2 measure. The L_∞ criterion—that is, the sup absolute error (SAE)—also leads to an asymptotically optimal bin width that is proportional to $n^{-1/3}$. Based on that criterion, we have the rule

$$1.66s \left(\frac{\log n}{n} \right)^{1/3},$$

where s is an estimate of the scale.

Bin Sizes By Choice of the Number of Equal-Width Bins

One of the most commonly used rules is for the number of bins rather than the width. Assume a symmetric binomial model for the bin counts, that is, the bin count is just the binomial coefficient. The total sample size n is

$$\sum_{k=0}^{m-1} \binom{m-1}{k} = 2^{m-1},$$

and so the number of bins is

$$m = 1 + \log_2 n.$$

Bin Shapes

In the univariate case, histogram bins may vary in size, but each bin is an interval. For the multivariate case, there are various possibilities for the shapes of the bins.

The simplest shape is the direct extension of an interval, that is a hyperrectangle. The volume of a hyperrectangle is just $v_k = \prod h_{kj}$.

There are, of course, other possibilities; any tessellation of the space would work.

Bin Shapes

The bins as geometric objects may or may not be regular, and they may or may not be of equal size.

Regular, equal-sized geometric figures such as hypercubes have the advantages of simplicity, both computationally and analytically. In two dimensions, there are three possible regular tessellations: triangles, squares, and hexagons.

For two dimensions hexagons are slightly better than squares and triangles with respect to the AMISE.

Various other tessellations may also work well.

Higher Dimensions

For hyperrectangles of constant size, the univariate theory generally extends fairly easily to the multivariate case. The histogram density estimator is

$$\hat{p}_H(y) = \frac{n_k}{nh_1h_2 \cdots h_d}, \quad \text{for } y \in T_k,$$

where the h 's are the lengths of the sides of the rectangles. The variance within the k^{th} bin is

$$V(\hat{p}_H(y)) = \frac{np_k(1 - p_k)}{(nh_1h_2 \cdots h_d)^2}, \quad \text{for } y \in T_k,$$

and the integrated variance is

$$IV(\hat{p}_H) \approx \frac{1}{nh_1h_2 \cdots h_d} - \frac{\mathcal{S}(f)}{n}.$$

Other Density Estimators Related to the Histogram

There are several variations of the histogram that are useful as probability density estimators. The most common modification is to connect points on the histogram by a continuous curve. A simple way of doing this in the univariate case leads to the *frequency polygon*.

This is the piecewise linear curve that connects the midpoints of the bins of the histogram. The endpoints are usually zero values at the midpoints of two appended bins, one on either side.

The *histospline* is constructed by interpolating knots of the empirical CDF with a cubic spline and then differentiating it. More general methods use splines or orthogonal series to fit the histogram.

The Average Shifted Histogram

The histogram is somewhat sensitive in appearance to the location of the bins, even for a fixed width of the bins.

To overcome the problem of location of the bins, a density estimator that is the average of several histograms with equal bin widths but different bin locations can be used. This is called the *average shifted histogram*, or ASH.

It also has desirable statistical properties, and it is computationally efficient in the multivariate case.

Kernel Estimators

Kernel estimators can be introduced using the example of a shifted histogram suggested by Rosenblatt.

For the one-dimensional case, we can form a histogram that is shifted to be centered on the point at which the density is to be estimated.

Given the sample y_1, \dots, y_n , Rosenblatt's histogram estimator at the point y is

$$\hat{p}_R(y) = \frac{\#\{y_i \text{ s.t. } y_i \in (y - h/2, y + h/2]\}}{nh}.$$

This histogram estimator avoids the ordinary histogram's constant-slope contribution to the bias. This estimator is a step function with variable lengths of the intervals that have constant value.

Kernel Estimators

Rosenblatt's centered histogram can also be written in terms of the ECDF:

$$\hat{p}_R(y) = \frac{P_n(y + h/2) - P_n(y - h/2)}{h},$$

where, as usual, P_n denotes the ECDF.

As seen in this expression, Rosenblatt's estimator is a centered finite-difference approximation to the derivative of the empirical cumulative distribution function (which, of course, is not differentiable at the data points).

We could use the same idea, and form other density estimators using other finite-difference approximations to the derivative of P_n .

Kernel Estimators

Another way to write Rosenblatt's shifted histogram estimator over bins of length h is

$$\hat{p}_R(y) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{y - y_i}{h}\right),$$

where $K(t) = 1$ if $|t| < 1/2$ and $= 0$ otherwise.

The function K is a kernel or filter.

In Rosenblatt's estimator, it is a "boxcar" function, but other kernel functions could be used.

Kernel Estimators

The estimator extends easily to the multivariate case.

In the general kernel estimator, we usually use a more general scaling of $y - y_i$,

$$V^{-1}(y - y_i),$$

for some positive-definite matrix V .

The determinant of V^{-1} scales the estimator to account for the scaling within the kernel function.

The general kernel estimator is given by

$$\hat{p}_K(y) = \frac{1}{n|V|} \sum_{i=1}^n K\left(V^{-1}(y - y_i)\right),$$

where the function K is called the *kernel*, and V is the *smoothing matrix*.

The Smoothing Matrix

The determinant of the smoothing matrix is exactly analogous to the bin volume in a histogram estimator.

The univariate version of the kernel estimator is the same as Rosenblatt's estimator, but in which a more general function K is allowed.

In practice, V is usually taken to be constant for a given sample size, but, of course, there is no reason for this to be the case, and indeed it may be better to vary V depending on the number of observations near the point y .

The dependency of the smoothing matrix on the sample size n and on y is often indicated by the notation $V_n(y)$.

Properties of Kernel Estimators

The appearance of the kernel density estimator depends to some extent on the support and shape of the kernel.

Unlike the histogram estimator, the kernel density estimator may be continuous and even smooth.

It is easy to see that if the kernel satisfies

$$K(t) \geq 0,$$

and

$$\int_{\mathbb{R}^d} K(t) dt = 1$$

(that is, if K is a density), then $\hat{p}_K(y)$ is a bona fide density estimator.

Properties of Kernel Estimators

There are other requirements that we may impose on the kernel either for the theoretical properties that result or just for their intuitive appeal.

It also seems reasonable that in estimating the density at the point y , we would want to emphasize the sample points near y . This could be done in various ways, but one simple way is to require

$$\int_{\mathbb{R}^d} tK(t) dt = 0.$$

In addition, we may require the kernel to be symmetric about 0.

Properties of Kernel Estimators

For multivariate density estimation, the kernels are usually chosen as a radially symmetric generalization of a univariate kernel. Such a kernel can be formed as a product of the univariate kernels.

For a product kernel, we have for some constant σ_K^2 ,

$$\int_{\mathbb{R}^d} tt^{\top} K(t) dt = \sigma_K^2 I_d,$$

where I_d is the identity matrix of order d . (We could also impose this as a requirement on any kernel, whether it is a product kernel or not.)

This makes the expressions for bias and variance of the estimators simpler. The spread of the kernel can always be controlled by the smoothing matrix V , so sometimes, for convenience, we sometimes require $\sigma_K^2 = 1$.

Pointwise Properties of Kernel Estimators

The pointwise properties of the kernel estimator are relatively simple to determine because the estimator at a point is merely the sample mean of n independent and identically distributed random variables.

The expectation of the kernel estimator at the point y is the convolution of the kernel function and the probability density function,

$$\begin{aligned} E(\hat{p}_K(y)) &= \frac{1}{|V|} \int_{\mathbb{R}^d} K(V^{-1}(y-t)) p(t) dt \\ &= \int_{\mathbb{R}^d} K(u) p(y-Vu) du, \end{aligned}$$

where $u = V^{-1}(y-t)$ (and, hence, $du = |V|^{-1}dt$).

Pointwise Expectation, Continued

If we approximate $p(y - Vu)$ about y with a three-term Taylor series, using the properties of the kernel and using properties of the trace, we have

$$\begin{aligned} \mathbb{E}(\hat{p}_K(y)) &\approx \int_{\mathbb{R}^d} K(u) \left(p(y) - (Vu)^\top \nabla p(y) + \frac{1}{2} (Vu)^\top \mathbf{H}_p(y) Vu \right) du \\ &= p(y) - 0 + \frac{1}{2} \text{trace} \left(V^\top \mathbf{H}_p(y) V \right). \end{aligned}$$

To second order in the elements of V (that is, $O(|V|^2)$), the bias at the point y is therefore

$$\frac{1}{2} \text{trace} \left(V V^\top \mathbf{H}_p(y) \right).$$

Pointwise Variance

Using the same kinds of expansions and approximations to evaluate $E\left((\hat{p}_K(y))^2\right)$ to get an expression of order $O(|V|/n)$, and subtracting the square of the expectation in equation, we get the approximate variance at y as

$$V(\hat{p}_K(y)) \approx \frac{p(y)}{n|V|} \int_{\mathbb{R}^d} (K(u))^2 du,$$

or

$$V(\hat{p}_K(y)) \approx \frac{p(y)}{n|V|} \mathcal{S}(K).$$

Integrating this, because p is a density, we have

$$\text{AIV}(\hat{p}_K) = \frac{\mathcal{S}(K)}{n|V|},$$

and integrating the square of the asymptotic bias in expression, we have

$$\text{AISB}(\hat{p}_K) = \frac{1}{4} \int_{\mathbb{R}^d} \left(\text{trace} \left(V^T H_p(y) V \right) \right)^2 dy.$$

Pointwise Variance

These expressions are much simpler in the univariate case, where the smoothing matrix V is the smoothing parameter or window width h . We have a simpler approximation for $E(\hat{p}_K(y))$:

$$E(\hat{p}_K(y)) \approx p(y) + \frac{1}{2}h^2 p''(y) \int_{\mathbb{R}} u^2 K(u) du,$$

and from this we get a simpler expression for the AISB.

After likewise simplifying the AIV, we have

$$\text{AMISE}(\hat{p}_K) = \frac{\mathcal{S}(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 \mathcal{R}(p),$$

where we have left the kernel unscaled (that is, $\int u^2 K(u) du = \sigma_K^2$).

Minimizing this with respect to h , we have the optimal value of the smoothing parameter

$$\left(\frac{\mathcal{S}(K)}{n\sigma_K^4 \mathcal{R}(p)} \right)^{1/5}.$$

Pointwise Variance

Substituting this back into the expression for the AMISE, we find that its optimal value in this univariate case is

$$\frac{5}{4} \mathcal{R}(p) (\sigma_K \mathcal{S}(K))^{4/5} n^{-4/5}.$$

The AMISE for the univariate kernel density estimator is thus $O(n^{-4/5})$.

Recall that the AMISE for the univariate histogram density estimator is $O(n^{-2/3})$.

Kernel Estimators

We see that the bias and variance of kernel density estimators have similar relationships to the smoothing matrix that the bias and variance of histogram estimators have.

As the determinant of the smoothing matrix gets smaller (that is, as the window of influence around the point at which the estimator is to be evaluated gets smaller), the bias becomes smaller and the variance becomes larger.

This agrees with what we would expect intuitively.

Choice of Kernels

Standard normal densities have these properties described above, so the kernel is often chosen to be the standard normal density.

As it turns out, the kernel density estimator is not very sensitive to the form of the kernel.

Although the kernel may be from a parametric family of distributions, in kernel density estimation, we do not estimate those parameters; hence, the kernel method is a nonparametric method.

Choice of Kernels

Sometimes, a kernel with finite support is easier to work with. In the univariate case, a useful general form of a compact kernel is

$$K(t) = \kappa_{rs}(1 - |t|^r)^s \mathbf{I}_{[-1,1]}(t),$$

where

$$\kappa_{rs} = \frac{r}{2\mathbf{B}(1/r, s + 1)}, \quad \text{for } r > 0, s \geq 0,$$

and $\mathbf{B}(a, b)$ is the complete beta function.

Choice of Kernels

This general form leads to several simple specific cases:

- for $r = 1$ and $s = 0$, it is the rectangular kernel;
- for $r = 1$ and $s = 1$, it is the triangular kernel;
- for $r = 2$ and $s = 1$ ($\kappa_{rs} = 3/4$), it is the “Epanechnikov” kernel, which yields the optimal rate of convergence of the MISE;
- for $r = 2$ and $s = 2$ ($\kappa_{rs} = 15/16$), it is the “biweight” kernel.

If $r = 2$ and $s \rightarrow \infty$, we have the Gaussian kernel (with some rescaling).

Multivariate Kernels

For multivariate density estimation, the kernels are often chosen as a product of the univariate kernels. The product Epanechnikov kernel, for example, is

$$K(t) = \frac{d+2}{2c_d} (1 - t^\top t) \mathbf{I}_{(t^\top t \leq 1)},$$

where

$$c_d = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)}.$$

We have seen that the AMISE of a kernel estimator depends on $\mathcal{S}(K)$ and the smoothing matrix V . As we mentioned above, the amount of smoothing (that is, the window of influence) can be made to depend on σ_K . We can establish an approximate equivalence between two kernels, K_1 and K_2 , by choosing the smoothing matrix to offset the differences in $\mathcal{S}(K_1)$ and $\mathcal{S}(K_2)$ and in σ_{K_1} and σ_{K_2} .

Computation of Kernel Density Estimators

If the estimate is required at one point only, it is simplest just to compute it directly. If the estimate is required at several points, it is often more efficient to compute the estimates in some regular fashion.

If the estimate is required over a grid of points, a fast Fourier transform (FFT) can be used to speed up the computations. Using a Gaussian kernel, first take the discrete Fourier transform of the data (using a histogram on 2^k cells) and then invert the product of that and the Fourier transform of the Gaussian kernel, $\exp(-h^2 s^2 / 2)$.

Choice of Window Widths

An important problem in nonparametric density estimation is to determine the smoothing parameter, such as the bin volume, the smoothing matrix, the number of nearest neighbors, or other measures of locality. In kernel density estimation, the window width has a much greater effect on the estimator than the kernel itself does.

An objective is to choose the smoothing parameter that minimizes the MISE. We often can do this for the AMISE.

It is not as easy for the MISE. The first problem, of course, is just to estimate the MISE.

In practice, we use cross validation with varying smoothing parameters and alternate computations between the MISE and AMISE.

Choice of Window Widths

In univariate density estimation, the MISE has terms such as $h^\alpha \mathcal{S}(p')$ (for histograms) or $h^\alpha \mathcal{S}(p'')$ (for kernels). We need to estimate the roughness of a derivative of the density.

Using a histogram, a reasonable estimate of the integral $\mathcal{S}(p')$ is a Riemann approximation,

$$\begin{aligned}\hat{\mathcal{S}}(p') &= h \sum (\hat{p}'(t_k))^2 \\ &= \frac{1}{n^2 h^3} \sum (n_{k+1} - n_k)^2,\end{aligned}$$

where $\hat{p}'(t_k)$ is the finite difference at the midpoints of the k^{th} and $(k+1)^{\text{th}}$ bins; that is,

$$\hat{p}'(t_k) = \frac{n_{k+1}/(nh) - n_k/(nh)}{h}.$$

This estimator is biased.

Choice of Window Widths

For the histogram,

$$E(\hat{S}(p')) = S(p') + 2/(nh^3) + \dots$$

A standard estimation scheme is to correct for the $2/(nh^3)$ term in the bias and plug this back into the formula for the AMISE (which is $1/(nh) + h^2 S(r')/12$ for the histogram).

We compute the estimated values of the AMISE for various values of h and choose the one that minimizes the AMISE. This is called *biased cross validation* because of the use of the AMISE rather than the MISE.

Orthogonal Series Estimators

A continuous real function $p(x)$, integrable over a domain D , can be represented over that domain as an infinite series in terms of a complete spanning set of real orthogonal functions $\{f_k\}$ over D :

$$p(x) = \sum_k c_k f_k(x).$$

The orthogonality property allows us to determine the coefficients c_k in the expansion:

$$c_k = \langle f_k, p \rangle.$$

Orthogonal Series Estimators

Approximation using a truncated orthogonal series can be particularly useful in estimation of a probability density function because the orthogonality relationship provides an equivalence between the coefficient and an expected value.

Expected values can be estimated using observed values of the random variable and the approximation of the probability density function.

Assume that the probability density function p is approximated by an orthogonal series $\{q_k\}$ with weight function $w(y)$:

$$p(y) = \sum_k c_k q_k(y).$$

We have

$$\begin{aligned} c_k &= \langle q_k, p \rangle \\ &= \int_D q_k(y) p(y) w(y) dy \\ &= E(q_k(Y) w(Y)). \end{aligned}$$

Orthogonal Series Estimators

The c_k can therefore be unbiasedly estimated by

$$\hat{c}_k = \frac{1}{n} \sum_{i=1}^n q_k(y_i) w(y_i).$$

The orthogonal series estimator is therefore

$$\hat{p}_S(y) = \frac{1}{n} \sum_{k=0}^j \sum_{i=1}^n q_k(y_i) w(y_i) q_k(y)$$

for some truncation point j .

Without some modifications, this generally is not a good estimator of the probability density function. It may not be smooth, and it may have infinite variance.

The estimator may be improved by shrinking the \hat{c}_k toward the origin.

Orthogonal Series Estimators

The number of terms in the finite series approximation also has a major effect on the statistical properties of the estimator.

Having more terms is not necessarily better.

Orthogonal Series Estimators

One useful property of orthogonal series estimators is that the convergence rate is independent of the dimension. This may make orthogonal series methods more desirable for higher-dimensional problems.

There are several standard orthogonal series that could be used. These two most commonly used series are the Fourier and the Hermite. There is no consensus on which is preferable.

The Fourier series is commonly used for distributions with bounded support. It yields estimators with better properties in the L_1 sense.

For distributions with unbounded support, the Hermite polynomials are most commonly used.

Other orthogonal systems can be used in density estimation.

Other Methods of Density Estimation

There are several other methods of probability density estimation. Most of them are modifications of the ones we have discussed. In some cases, combinations of methods can be used effectively. Some methods work only in the univariate case, whereas others can be applied in multivariate density estimation.

All of the nonparametric methods of density estimation involve decisions such as window width, number of mixtures, number of terms in an expansion, and so on. All of these quantities can be thought of as smoothing parameters.

There are various rules for making these decisions in an asymptotically optimal fashion for some known distributions.

Absent assumptions about the nature of the true distribution, it is difficult to decide on the extent of smoothing. Much of the current work is on developing adaptive methods in which these choices are made based on the data.

Filtered Kernel Methods

As we mentioned earlier, it may be reasonable to vary the smoothing parameter in the kernel density estimator in such a way that the locality of influence is smaller in areas of high density and larger in sparser regions. T

In the filtered kernel density estimator a set of functions f_1, f_2, \dots, f_m , such that $\sum f_j(x) = 1$, and associated window widths h_j , are used to weight the standard univariate kernel estimator. Analogous to the kernel estimator, the univariate filtered kernel density estimator is

$$\hat{p}_F(y) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \frac{f_j(y_i)}{h_j} K\left(\frac{y - y_i}{h_j}\right).$$

Filtered Kernel Methods

If the kernel function is the standard normal density $\phi(t)$ and the filtering functions are weighted normal densities $\pi_j \phi(t | \mu_j, \sigma_j^2)$, we can express the filtered kernel density estimator as

$$\hat{p}_F(y) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \frac{\pi_j \phi(t | \mu_j, \sigma_j^2)}{h \sigma_k f_{\bullet}(y_i)} \phi\left(\frac{y - y_i}{h \sigma_j}\right),$$

where $f_{\bullet}(t) = \sum_{j=1}^m \pi_j \phi(t | \mu_j, \sigma_j^2)$. We now have the choices in the estimator as m , π_j , μ_j , and σ_j^2 .

This approach is similar to a mixture approach in that the number of different filter functions must be chosen.

Alternating Kernel and Mixture Methods

A density estimator can be formed by alternating between non-parametric filtered kernel estimators and parametric mixture estimators composed of the same number of terms (filter functions or component densities). The estimator is computed iteratively by beginning with a mixture estimator $\hat{p}_M^{(1)}(y)$ and a filtered kernel estimator $\hat{p}_F^{(1)}(y)$.

A new mixture estimator $\hat{p}_M^{(2)}(y)$ is chosen as the mixture estimator closest to $\hat{p}_F^{(1)}(y)$ (in some norm). The new mixture estimator is used to determine the choices for a new filtered kernel estimator.

The process is continued until

$$\|\hat{p}_M^{(k+1)}(y) - \hat{p}_M^{(k)}(y)\|$$

is small.

Methods of Comparisons of Methods

Nonparametric probability density estimation involves the fundamental tradeoff between a spike at each observation (that is, no smoothing) and a constant function over the range of the observations (that is, complete smoothing) ignoring differences in relative frequencies of observations in various intervals.

It is therefore not surprising that the comparison of methods is not a trivial exercise.

One approach for comparing methods of estimation is to define broad classes of densities and to evaluate the performance of various estimators within those classes.

Methods of Comparisons of Methods

Ten useful densities for study:

- standard normal, $\phi(y | 0, 1)$;
- mixed normals, $p(y) = 0.7\phi(y | -2, 1.5) + 0.3\phi(y | 2, 0.5)$;
- standard Cauchy, $p(y) = 1/(\pi(1 + y^2))$;
- standard extreme value distribution, $p(y) = \exp(-e^{-y} - y)$;
- standard logistic, $p(y) = e^{-y}/(1 + e^{-y})^2$;
- standard Laplace, $p(y) = e^{-|y|/2}$;
- claw density

$$p(y) = \frac{1}{10}(5\phi(y | 0, 1) + \phi(y | -1, 0.1) + \phi(y | -0.5, 0.1) + \phi(y | 0, 0.1) + \phi(y | 0.5, 0.1) + \phi(y | 1, 0.1));$$

- smooth comb density

$$p(y) = \frac{32}{63}\phi\left(y \mid -\frac{31}{21}, \frac{32}{63}\right) + \frac{16}{63}\phi\left(y \mid \frac{17}{21}, \frac{16}{63}\right) + \frac{8}{63}\phi\left(y \mid \frac{41}{21}, \frac{8}{63}\right) + \frac{4}{63}\phi\left(y \mid \frac{53}{21}, \frac{4}{63}\right) + \frac{2}{63}\phi\left(y \mid \frac{59}{21}, \frac{2}{63}\right) + \frac{1}{63}\phi\left(y \mid \frac{62}{21}, \frac{1}{63}\right);$$

- triangular density, $p(y) = (1 - |x|)_+$;
- saw-tooth density

$$g(y) = p(y + 9) + p(y + 7) + \cdots + p(y - 7) + p(y - 9),$$

where p is the triangular density.

These densities cover a wide range of shapes.