Nonparametric Function Estimation

Nonparametric Regression

\[ f(x) \]

\[ \hat{f}(x) \]

Density Estimation

\[ \hat{\rho}(x) \]
Nonparametric Probability Density Estimation

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”).
Nonparametric Probability Density Estimation

We consider several nonparametric estimators of a density; that is, estimators of a general nonnegative function that integrates to 1 and for which we make no assumptions about a functional form other than, perhaps, smoothness.

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

- \( \hat{p}(y) \geq 0 \) for all \( y \);

- \( \int_{\mathbb{R}^d} \hat{p}(y) \, dy = 1 \).
Bona Fide Density Estimator

A probability density estimator that is nonnegative and integrates to 1 is called a \textit{bona fide} estimator.

Rosenblatt has shown that no unbiased bona fide estimator can exist for all continuous \( p \).

Rather than requiring an unbiased estimator that cannot be a bona fide estimator, we generally seek a bona fide estimator with small mean squared error or a sequence of bona fide estimators \( \hat{p}_n \) that are asymptotically unbiased; that is,

\[
E_p(\hat{p}_n(y)) \rightarrow p(y) \quad \text{for all } y \in \mathbb{R}^d \text{ as } n \rightarrow \infty.
\]
The Likelihood Function

Suppose that we have a random sample, \(y_1, \ldots, y_n\), from a population with density \(p\).

Treating the density \(p\) as a variable, we write the likelihood functional as

\[
L(p; y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i).
\]

The maximum likelihood method of estimation obviously cannot be used directly because this functional is unbounded in \(p\).
Modifications of the Objective Function To Be Optimized

We may, however, seek an estimator that maximizes some modification of the likelihood.

How?

Alternatively, we may seek an estimator within some restricted class that maximizes the likelihood.

How?
A Restricted Maximum Likelihood Estimator

Consider a restricted maximization problem, for the random variables in $\mathbb{R}^d$.

For $k = 1, \ldots, m$, define connected regions $T_k$ such that $T_k \cap T_{\tilde{k}} = \emptyset$ and $\bigcup_{k=1}^m T_k = \mathbb{R}^d$.

Then define the maximization problem:

$$\max_{p \ni \forall y \in T_k, p(y) = c_k} L(p; y_1, \ldots, y_n). \quad (1)$$
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: $v_{n,k}$. 
Histogram Estimators

For the sample $y_1, \ldots, y_n$, the histogram estimator of the probability density function is defined as

$$\hat{p}_H(y) = \sum_{k=1}^{m} \frac{1}{v_k} \frac{\sum_{i=1}^{n} I_{T_k}(y_i)}{n} I_{T_k}(y), \quad \text{for } y \in D,$$

$$= 0, \quad \text{otherwise.}$$

The histogram is the restricted maximum likelihood estimator (1).

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

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

we have the simpler expression for the histogram over $D$,

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

(2)
Histogram Estimators

This is a bona fide estimator:

\[ \hat{p}_H(y) \geq 0 \]

and

\[
\int_{\mathbb{R}^d} \hat{p}_H(y) dy = \sum_{k=1}^{m} \frac{n_k}{nv_k} v_k = 1.
\]
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

\[
a = t_{n,1} < t_{n,2} < \ldots < t_{n,m+1} = b.
\]

The univariate histogram is

\[
\hat{p}_H(y) = \sum_{k=1}^{m} \frac{n_k}{n(t_{n,k+1} - t_{n,k})} I_{T_k}(y).
\] (3)

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

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

This class of functions consists of polynomial splines of degree 0 with fixed knots, and the histogram is the maximum likelihood estimator over the class of step functions.
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.

It is sensitive both to the bin size and to the choice of the origin.

An important advantage of the histogram estimator is its simplicity, both for computations and for analysis.
Global Properties of the Histogram Estimator

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

\[ g(t) = \begin{cases} c_k, & \text{for } t \in T_k, \\ 0, & \text{otherwise,} \end{cases} \]

and where \( g(t) \geq 0 \) and \( \int_{\bigcup_k T_k} g(t) \, dt = 1. \)
Pointwise and Binwise Properties

Properties of the histogram vary from bin to bin.

From equation (2), the expectation of the histogram estimator at the point $y$ in bin $T_k$ is

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

(4)

where

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

(5)

is the probability content of the $k^{th}$ bin.
**Pointwise and Binwise Properties: Bias**

Let’s consider some pointwise properties of the histogram estimator.

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

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

(6)

Note that the bias 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 and Binwise Properties: Bias

We can bound the bias if we assume a regularity condition on $p$. If there exists $\gamma$ 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 $\xi_k$ in the $k^{th}$ bin, we have

$$|\text{Bias}(\hat{p}_H(y))| = |p(\xi_k) - p(y)| \\
\leq \gamma_k \|\xi_k - y\| \\
\leq \gamma_k v_k. \quad (7)$$
Pointwise and Binwise Properties: Variance

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

\[
V(\hat{p}_H(y)) = \frac{V(n_k)/(nv_k)^2}{n v_k^2} = \frac{p_k(1 - p_k)}{nv_k^2}.
\]  

(8)

This is easily seen by recognizing that \( n_k \) is a binomial random variable with parameters \( n \) and \( p_k \).

Notice that the variance decreases as the bin size increases. Note also that the variance is different from bin to bin.
Pointwise and Binwise Properties:

Variance

We can bound the variance:

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

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

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

Notice the tradeoff between bias and variance: as $h$ increases the variance, equation (8), decreases, but the bound on the bias, equation (7), increases.
Pointwise and Binwise Properties: MSE

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

$$\text{MSE}(\hat{p}_H(y)) = \frac{p_k(1-p_k)}{nv_k^2} + \left(\frac{p_k}{v_k} - p(y)\right)^2. \quad (9)$$

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

$$\text{MSE}(\hat{p}_H(y)) \leq \frac{p(\xi_k)}{nv_k} + \gamma_k^2v_k^2. \quad (10)$$

We easily see that the histogram estimator is $L_2$ pointwise consistent for a Lipschitz-continuous density if, as $n \to \infty$, for each $k$, $v_k \to 0$ and $nv_k \to \infty$. 
Optimal Bin Size and Order of MSE

Now, let’s consider the univariate case. To remind us of this, we’ll denote the bin width as \( h_k \), instead of the volume \( v_k \).

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

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

(11)

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 expressions for the integrated variance and the integrated squared bias are quite complicated because they 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.
AMISE of Histogram Estimators

First, consider the integrated variance,

\[
\text{IV}(\hat{p}_H) = \int_{\mathbb{R}^d} \text{V}(\hat{p}_H(t)) \, dt
\]

\[
= \sum_{k=1}^{m} \int_{T_k} \text{V}(\hat{p}_H(t)) \, dt
\]

\[
= \sum_{k=1}^{m} \frac{p_k - p_k^2}{n v_k}
\]

\[
= \sum_{k=1}^{m} \left( \frac{1}{n v_k} - \frac{\sum p(\xi_k)^2 v_k}{n} \right) + o(n^{-1})
\]

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

\[
S(g) = \int_{\mathbb{R}^d} (g(t))^2 \, dt.
\]
We have the integrated variance,

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

and the asymptotic integrated variance,

$$\text{AIV}(\hat{p}_H) = \sum_{k=1}^{m} \frac{1}{nv_k}.$$  (14)

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

Now, consider the other term in the integrated MSE, the integrated squared bias.

We will consider the case of rectangular bins, in which $h_k = (h_{k1}, \ldots, h_{kd})$ is the vector of lengths of sides in the $k^{th}$ bin, so $v_k = \prod_{j=1}^{d} h_{kj}$.
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

$$p(\bar{t}_k + t) = p(\bar{t}_k) + t^T \nabla p(\bar{t}_k) + \frac{1}{2} t^T H_p(\bar{t}_k) t + \cdots,$$  \hspace{1cm} (15)

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

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

$$p_k = \int_{\bar{t}_k + t \in T_k} p(\bar{t}_k + t) \, dt$$

$$= \int_{-h_{kd}/2}^{h_{kd}/2} \cdots \int_{-h_{k1}/2}^{h_{k1}/2} \left( p(\bar{t}_k) + t^T \nabla p(\bar{t}_k) + \cdots \right) \, dt_1 \cdots dt_d$$

$$= v_k p(\bar{t}_k) + \mathcal{O}(h_{k*}^{d+2}),$$  \hspace{1cm} (16)

where $h_{k*} = \min_j h_{kj}$. 

continued ...
The bias at a point $\bar{t}_k + t$ in the $k^{th}$ bin, after substituting equations (15) and (16) into equation (6), is

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

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

$$\text{ISB}_k(\hat{p}_H) = \int_{T_k} \left( (t^T \nabla p(\bar{t}_k))^2 - 2O(h_{k*}^2) t^T \nabla p(\bar{t}_k) + O(h_{k*}^4) \right) dt$$

$$= \int_{-h_{kd}/2}^{h_{kd}/2} \cdots \int_{-h_{kd}/2}^{h_{kd}/2} \sum_i \sum_j t_{ki}^j t_{kj} \nabla_i p(\bar{t}_k) \nabla_j p(\bar{t}_k) dt_1 \cdots dt_d + O(h_{k*}^{4+d}).$$

(17)

Many of the expressions above are simpler if we use a constant bin size, $v$, or $h_1, \ldots, h_d$.  

continued ...
In the case of constant bin size, the asymptotic integrated variance in equation (14) becomes

\[ \text{AIV}(\hat{p}_H) = \frac{m}{nv}. \tag{18} \]

In this case, the integral in equation (17) simplifies as the integration is performed term by term because the cross-product terms cancel, and the integral is

\[ \frac{1}{12}(h_1 \cdots h_d) \sum_{j=1}^{d} h_j^2 (\nabla_j p(\bar{t}_k))^2. \tag{19} \]

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

When we sum the expression (19) over all bins, the \((\nabla_j p(\bar{t}_k))^2\) becomes \(S(\nabla_j p)\), and we have the asymptotic integrated squared bias,

\[ \text{AISB}(\hat{p}_H) = \frac{1}{12} \sum_{j=1}^{d} h_j^2 S(\nabla_j p). \tag{20} \]
Combining the asymptotic integrated variance, equation (18), and squared bias, equation (20), for the histogram with rectangular bins of constant size, we have

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

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

As we have mentioned and have seen by example, the histogram is very sensitive to the bin sizes, both in appearance and in other properties.

Equation (21) for the AMISE assuming constant rectangular bin size can be used as a guide for determining the bin size to use when constructing a histogram.

This expression involves $S(\nabla_j p)$ and so, of course, cannot be used directly.
Nevertheless, differentiating the expression in Equation (21) 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,

$$h_{j^*} = S(\nabla_j p)^{-1/2} \left( 6 \prod_{i=1}^{d} S(\nabla_i p)^{1/2} \right)^{\frac{1}{2+d}} n^{-\frac{1}{2+d}}. \quad (22)$$

Substituting this into equation (21), we have the optimal value of the AMISE

$$\frac{1}{4} \left( 36 \prod_{i=1}^{d} S(\nabla_i p)^{1/2} \right)^{\frac{1}{2+d}} n^{-\frac{2}{2+d}}. \quad (23)$$
Bin Sizes, continued ...

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, continued ...

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

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

A “good general distribution”, of course, is the normal with a diagonal variance-covariance matrix.

For the $d$-variate normal with variance-covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$,

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

For a univariate normal density with variance $\sigma^2$,

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

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

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

In practice, of course, an estimate of $\sigma$ must be used. The sample standard deviation the sample interquartile range are obvious choices.
Bin Sizes, continued ...

The AMISE is essentially an $L_2$ measure. The $L_\infty$ criterion—that is, the sup absolute error (SAE) of equation (??)—also leads to an asymptotically optimal bin width that is proportional to $n^{-1/3}$.

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 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.
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.
Other Density Estimators Related to the Histogram

As we have mentioned and have seen by example, the histogram is somewhat sensitive in appearance to the location 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 methods are probably the most widely used technique for building nonparametric probability density estimators.

They are best understood by developing them as a special type of histogram.

The difference is that the bins in kernel estimators are centered at the points at which the estimator is to be computed.

The problem of the choice of location of the bins in histogram estimators does not arise.
Rosenblatt’s Histogram Estimator; Kernels

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

Given the sample $y_1, \ldots, y_n$, Rosenblatt’s histogram estimator at the point $y$ is

$$\hat{p}_R(y) = \frac{\#\{y_i \text{ s.t. } y_i \in (y - h/2, y + h/2]\}}{nh}.$$  \hspace{1cm} (24)

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.
Rosenblatt’s Histogram Estimator; Kernels

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, of course, use the same idea and form other density estimators using other finite-difference approximations to the derivative of \( P_n \).
Rosenblatt’s Histogram Estimator; Kernels

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

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

(25)

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.
Rosenblatt’s Histogram Estimator; Kernels

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

$$\hat{p}_K(y) = \frac{1}{n|V|} \sum_{i=1}^{n} K \left( V^{-1}(y - y_i) \right), \quad (26)$$

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

The determinant of the smoothing matrix is exactly analogous to the bin volume in a histogram estimator.
Rosenblatt’s Histogram Estimator; Kernels

The univariate version of the kernel estimator is the same as Rosenblatt’s estimator (25), 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.

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

\[ K(t) \geq 0, \quad (27) \]

and

\[ \int_{\mathbb{R}^d} K(t) \, dt = 1 \quad (28) \]

(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

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

(29)

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 $\sigma_K^2$,

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

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.
Properties of Kernel Estimators

The spread of the kernel can always be controlled by the smoothing matrix $V$, so sometimes, for convenience, we require $\sigma_{K}^2 = 1$.

We usually require that the kernel satisfies the properties in equations (27) through (30).

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.
Properties of Kernel Estimators

The expectation of the kernel estimator (26) at the point $y$ is the convolution of the kernel function and the probability density function,

$$E(\hat{p}_K(y)) = \frac{1}{|V|} \int_{\mathbb{R}^d} K \left(V^{-1}(y-t)\right) p(t) \, dt$$

$$= \int_{\mathbb{R}^d} K(u) p(y-Vu) \, du,$$

where $u = V^{-1}(y-t)$ (and, hence, $du = |V|^{-1} dt$).
Properties of Kernel Estimators: Bias

If we approximate $p(y - Vu)$ about $y$ with a three-term Taylor series, using the properties of the kernel in equations (27) through (30) and using properties of the trace, we have

$$\mathbb{E}(\hat{p}_K(y)) \approx \int_{\mathbb{R}^d} K(u) \left( p(y) - (Vu)^\top \nabla p(y) + \frac{1}{2}(Vu)^\top H_p(y)Vu \right) \, du$$

$$= p(y) - 0 + \frac{1}{2} \text{trace} \left( V^\top H_p(y)V \right). \quad (32)$$

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

$$\frac{1}{2} \text{trace} \left( VV^\top H_p(y) \right). \quad (33)$$
Properties of Kernel Estimators

Using the same kinds of expansions and approximations as in equations (31) and (32) to evaluate \( \mathbb{E}\left( (\hat{p}_K(y))^2 \right) \) to get an expression of order \( O(|V|/n) \), and subtracting the square of the expectation in equation (32), 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

\[
V(\hat{p}_K(y)) \approx \frac{p(y)}{n|V|} S(K). \tag{34}
\]

Integrating this, because \( p \) is a density, we have

\[
AIV(\hat{p}_K) = \frac{S(K)}{n|V|}, \tag{35}
\]

and integrating the square of the asymptotic bias in expression (33), we have

\[
AISB(\hat{p}_K) = \frac{1}{4} \int_{\mathbb{R}^d} \left( \text{trace} \left( V^T H_p(y) V \right) \right)^2 \, dy. \tag{36}
\]
Properties of Kernel Estimators

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 \( E(\hat{p}_K(y)) \) than that given in equation (32),

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

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

After likewise simplifying the AIV, we have

\[
\text{AMISE}(\hat{p}_K) = \frac{S(K)}{nh} + \frac{1}{4} \sigma^4_K h^4 \mathcal{R}(p), \tag{37}
\]

where we have left the kernel unscaled (that is, \( \int u^2 K(u) \, du = \sigma^2_K \)).
Optimal Smoothing Parameters

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

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

(38)

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

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

(39)

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 $O(n^{-2/3})$. 


Optimal Smoothing Parameters

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.

Sometimes, a kernel with finite support is easier to work with.
Choice of Kernels

In the univariate case, a useful general form of a compact kernel is

\[ K(t) = \kappa_{rs}(1 - |t|^r)^sI_{[-1,1]}(t), \]

where

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

and \( B(a, b) \) 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 (see Epanechnikov, 1969);

- for $r = 2$ and $s = 2$ ($\kappa_{rs} = 15/16$), it is the “biweight” kernel.

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

As mentioned above, for multivariate density estimation, the kernels are often chosen as a product of the univariate kernels.

The product Epanechnikov kernel, for example, is

\[ K(t) = \frac{d + 2}{2c_d} (1 - t^T t) I(t^T t \leq 1), \]

where

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

We have seen that the AMISE of a kernel estimator (that is, the sum of equations (35) and (36)) depends on \( S(K) \) and the smoothing matrix \( V \).
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, as in equation (22) on page 0. It is not as easy for the MISE.

The first problem, of course, is just to estimate the MISE.
Choice of Window Widths

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

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

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

\[
\hat{S}(p') = h \sum (\hat{p}'(t_k))^2 = \frac{1}{n^2 h^3} \sum (n_{k+1} - n_k)^2,
\]

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

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

This estimator is biased.

For the histogram, for example,

\[
E(\hat{S}(p')) = S(p') + 2/(nh^3) + \ldots
\]

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).
Choice of Window Widths

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.

These same techniques can be used for other density estimators and for multivariate estimators, although at the expense of considerably more complexity.
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 $D$:

$$p(x) = \sum_k c_k f_k(x). \quad (40)$$

The orthogonality property allows us to determine the coefficients $c_k$ in the expansion (40):

$$c_k = \langle f_k, p \rangle. \quad (41)$$
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.
Orthogonal Series Estimators

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

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

From equation (41), we have

$$c_k = \langle q_k, p \rangle = \int_{D} q_k(y) p(y) w(y) dy = \mathbb{E}(q_k(Y)w(Y)),$$

(42)

where $Y$ is a random variable whose probability density function is $p$.

How to estimate this?
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 $j$. (43)
Orthogonal Series Estimators

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.

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. The two most commonly used series are the Fourier and the Hermite.

Which is preferable depends on the situation.

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.