Chapter 1
Directional Data

1. Introduction. Kronmal and Tarter (1968) proposed Fourier series density estimates. These have since been widely studied and applied. This chapter will provide a superficial review, with no claim to novelty. Our purpose will be to illustrate, using this particularly elegant method, the most important properties of linear density estimates in general. Because our purpose is expository, we will show them in an unfamiliar context, and using slightly unconventional notation.

2. Directional and Periodic Data. Directional data are a random sample of measured directions from a reference point. We shall measure those directions in radians of the angle from some reference direction, say straight east, which will have a radian value of zero. The mathematical peculiarity of directions is that if you let their value increase, moving counterclockwise conventionally, you eventually turn all the way around until you are pointing the same way you started. One complete rotation is $2\pi$ radians; so we might state this mathematically as follows: any function $f(x)$ of a direction $x$ has the property $f(x) = f(x + 2\pi)$ for all $x$. Any such function is said to be periodic with period $2\pi$. By repeated application of this relationship, we see that $f(x + 2\pi j) = f(x)$ for any integer (positive or negative) $j$.

Notice that directions are not the only quantities that can be studied using periodic functions. If you were studying how traffic patterns or shopping behavior varied over a typical week, you might use periodic functions with period 7 days. For studying normal seasonal weather changes, functions with period 365.25 days are appropriate. Everything we say about direction data will apply to all periodic data, with small adjustments to which we will return. Yet another source of periodic data would be locations on a circle, with those locations parametrized by angles; for example, the longitudes at which tropical storms arose in a given year. A text on the various statistical methods for analyzing periodic data is Fisher (1993).

We will attempt to model a random sample of measured directions by assuming it is i.i.d. (independently and identically distributed) from an unknown density $f$. This density must be periodic with period $2\pi$; and because it is a density, $f(x) \geq 0$. Further it must integrate to one over a single period; as an arbitrary convention, let that period be $[-\pi, \pi]$. Then we require $\int_{-\pi}^{\pi} f(x) \, dx = 1$. The density function will carry graphically the information that certain ranges of directions are more probable than others.
Since we treat the density as unknown, we will as usual want to estimate it using functions that are as simple as possible; that should make the graph most informative. If we did not require periodicity, we might use polynomials as examples of particularly simple functions; but these are not periodic. The simplest periodic function is just a constant. The next most simple might be a sine wave: \( a \cos(x + x_0) \), where \( a \) is the maximum amplitude and \(-x_0\), the phase, is the angle at which the maximum amplitude is attained. Then an estimated density might be \( \frac{1}{2} + \frac{1}{2} \cos(x + x_0) \), which meets our positivity and integration conditions; notice it is a linear combination of a constant and a sine wave.

Applying one of those sum formulas from trigonometry to our general sine wave gets \( a \cos(x + x_0) = a \cos(x) \cos(x) - \sin(x) \sin(x) \). This shows that a general sine wave may be decomposed as a linear combination of \( \cos(x) \) and \( \sin(x) \). It will turn out to be very convenient to express such general sine waves somewhat differently. We will use deMoivre's famous formula \( e^{ix} = \cos(x) + i \sin(x) \); where \( i \) is one of the square roots of \(-1\). This allows us to write \( \cos(x) = \frac{e^{ix} + e^{-ix}}{2} \) and \( \sin(x) = \frac{e^{ix} - e^{-ix}}{2i} \). Therefore, any sine wave may be written as a linear combination of \( e^{ix} \) and \( e^{-ix} \).

Of course, we have now allowed values of our functions to be general complex numbers. This is not a very useful degree of generality, since probability densities are real-valued functions. Under what conditions is a function \( a e^{ix} + b e^{-ix} \) real-valued? Remember that the conjugate of a complex number is \( \overline{c + id} = c - id \). You may verify quickly that \( \overline{z + w} = \overline{z} + \overline{w} \) and \( \overline{zw} = \overline{z} \cdot \overline{w} \). Then deMoivre's formula gets that \( \overline{e^{ix}} = e^{-ix} \). Now a complex number is real just when \( \overline{z} = z \). This lets us check that \( \overline{a e^{ix} + b e^{-ix}} = a e^{-ix} + b e^{ix} \); and therefore our exponential representation of a periodic function is real valued just when \( b = \bar{a} \).

3. Fourier Series. The density expressed in terms of a constant and a sine wave is of limited interest, of course, because it is too simple. It has a maximum density at a single point; and a minimum density at a distance of \( \pi \) radians, in the opposite direction. Interesting densities would likely have more structure, such as multiple local maxima. Some more interesting but simple periodic functions would then be \( \cos(x + x_j) \), where \( j \) is a positive integer. These have period \( 2\pi \); and therefore are also of period \( 2\pi \), and apply to directional data. They may be written as linear combinations \( a e^{jix} + b e^{-jix} \); when these are real, we again must have \( b = \bar{a} \).
We shall therefore investigate (finite) trigonometric series that look like
\[ g(x) = \frac{1}{2\pi} \sum_{j=-m}^{m} \phi_j e^{-ijx}, \]
and are real-valued whenever \( \phi_{-j} = \phi_j \), as possible density estimates. The larger we allow \( m \) to be, the more wiggly complexity is possible to \( g \). This will be the first example of a smoothing parameter; the less wiggly estimates involving small values of \( m \) we will think of as smoother.

Naturally, we will use the method of least-squares to find \( g \) as close as possible to the true density \( f \); that is \( \min_{\phi_j} \int_{-\pi}^{\pi} [f(x) - g(x)]^2 \, dx \). That squared difference must be rewritten using the complex absolute value \( |z|^2 = zz \). The constants \( \phi_j \) must be chosen in order to determine \( g \); so our problem becomes

\[
\min_{\phi_j} \int_{-\pi}^{\pi} [f(x) - g(x)]^2 \, dx = \min_{\phi_j} \int_{-\pi}^{\pi} \left( f(x) - \frac{1}{2\pi} \sum_{j=-m}^{m} \phi_j e^{-ijx} \right) \left( f(x) - \frac{1}{2\pi} \sum_{j=-m}^{m} \phi_j e^{-ijx} \right) \, dx,
\]

or

\[
\min_{\phi_j} \left[ \int_{-\pi}^{\pi} f(x) \, dx - \frac{1}{2\pi} \sum_{j=-m}^{m} \phi_j \int_{-\pi}^{\pi} f(x) e^{-ijx} \, dx - \frac{1}{2\pi} \sum_{j=-m}^{m} \phi_j \int_{-\pi}^{\pi} f(x) e^{ijx} \, dx + \frac{1}{4\pi^2} \sum_{j=-m}^{m} \sum_{k=-m}^{m} \overline{\phi_j} \phi_k \int_{-\pi}^{\pi} e^{-i(j-k)x} \, dx \right].
\]

Fortunately, this horrifying expression simplifies; the first term is irrelevant to the minimization. And, by the reality condition \( \phi_{-j} = \overline{\phi}_j \), the third term turns out to be identical to the second. In the final term, there are two relevant cases. If \( j \neq k \),

\[
\int_{-\pi}^{\pi} e^{(k-j)x} \, dx = \frac{1}{i(k-j)} e^{i(k-j)x} \bigg|_{-\pi}^{\pi} = 0
\]

by periodicity; if \( j = k \),

\[
\int_{-\pi}^{\pi} e^{(k-k)x} \, dx = \int_{-\pi}^{\pi} e^{0x} \, dx = 2\pi.
\]
After we simplify, and reverse the sign, our problem has become:
\[
\max_{\phi_j} \left[ \frac{1}{\pi} \sum_{j=-m}^{m} \phi_j \int_{-\pi}^{\pi} \frac{f(x)e^{-ixj}}{2\pi} \right].
\]

Note that this corresponds to the form of least-squares approximation $\max g \rightarrow S f g$ discussed in the preface. Now we can solve the problem by ordinary calculus: differentiate the expression with respect to each $\phi_j$ in turn and set equal to zero (notice that each non-zero $j$ appears twice in the second sum), to get $
abla \phi_j = \frac{1}{\pi} \sum_{j=-m}^{m} \phi_j e^{-ixj}$. These are called the Fourier coefficients of $f$; the collection of all the $\phi_j = \phi(0)$, for any integer value of $j$, is called the Fourier transform of $f$. When $f$ is the density of a random variable $X$, notice that $\phi(0) = \int_{-\pi}^{\pi} f(x)e^{ixj}dx = E(e^{ix})$; so in this case it is called the characteristic function of $X$ (notice that, therefore, $\phi(0) = 1$).

Since our expression for the $\phi_j$ does not depend on the smoothing parameter $m$, we can imagine letting $m$ go to infinity and writing $\phi(x) = \frac{1}{\pi} \sum_{j=-\infty}^{\infty} \phi_j e^{-ixj}$. This called the Fourier series of $f$. There is a wonderful, classical result that says that $g$ actually converges to $f$, and is thus no longer just an approximation, at points where $f$ is continuous (see e.g. Jones(1993)); this is an important example of a Fourier inversion theorem. It says that, if we take the Fourier transform of $f$, we can essentially reconstruct $f$ using this series. We shall see much more of these concepts, later in this monograph.

4. Orthogonal Series. Our technique for actually calculating that smoothed approximate function $g$ by least-squares is only the most important example of a class of related techniques. If we have a set of functions indexed by integers $P(x)$ with the property that, for some weight function $w(x) \geq 0$, $\int P(x)P(x)w(x)dx = 0$ for all $j \neq k$ but $\int P(x)P(x)w(x)dx = k_j > 0$, we call this an orthogonal system. The trigonometric functions are an example: let $w = 1$ on $[-\pi, \pi)$ and zero elsewhere, and let $P(x) = e^{-ix}$; then $k_j = 2\pi$.

Now build functions from series of these: $q(x) = \sum_{j \neq k} a_j P(x)$. We can use these series to approximate a function $f$ by solving $\min \int [f(x) - q(x)]^2 w(x)dx$; this proceeds exactly as above. We get for the unknown constants $a_j = \int f(x)P(x)w(x)dx$. If there are a rich enough variety of functions in the orthogonal system that that integrated squared error may be made arbitrarily small for any $f$ with $\int [f(x)]^2 w(x)dx$ finite (as in the trigonometric system), we call the $P(x)$ a complete orthogonal system. We shall make extensive use of
such series approximations, later in the monograph. In many of our applications, everything is real, and so all those irritating complex conjugates disappear.

If we have such a complete orthogonal system, let us compute \( \int |\phi|^2 w(x) \, dx = \int |\hat{\phi}|^2 w(x) \, dx \) by the same technique of expansion used earlier. Generally, let \( g(x) \) have expansion coefficients \( a_j \), and a second function \( h(x) \) have expansion coefficients \( b_j \). Then using the orthogonality conditions above, you should check that \( \int g(x) h(x) \, dx = \sum a_j b_j \). In the periodic case, let \( h(x) \) have Fourier coefficients \( \gamma_j \); then \( \int g(x) h(x) \, dx = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \phi_j \gamma_j \), which is called Parseval's theorem. In our application, this becomes \( \int \phi(x)^2 \, dx = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} |\phi_j|^2 \).

We can extract important information from this immediately. If \( \phi(x) = \frac{1}{2\pi} \sum_{j=-m}^{m} \phi_j e^{-ijx} \) is a finite least squares approximation to \( f \), then the difference in the two is almost everywhere \( \phi(x) - \phi(x) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \phi_j e^{-ijx} \). The integrated squared error in this approximation is then \( \int \phi(x)^2 \, dx = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} |\phi_j|^2 = \frac{1}{\pi} \sum_{j=m+1}^{\infty} |\phi_j|^2 \); and there is an obvious generalization to any other complete orthonormal system.

5. Nonparametric Density Estimates. Of course, for the nonparametric density estimation problem, we do not know that density \( f \); rather, we have an i.i.d. random sample \( x_k \) for \( k = 1, \ldots, n \). We noted, however, that \( \phi_j = E(e^{ijX}) \). Following the Preface, we have a canonical estimate for this parameter, \( \hat{\phi}_j = E(e^{ijX}) = \frac{1}{n} \sum_{k=1}^{n} e^{ijx_k} \). We have used a sample average to estimate a population average. This leads to a Fourier series density estimate \( \phi(x) = \frac{1}{2\pi} \sum_{j=-m}^{m} \hat{\phi}_j e^{-ijx} \). This is real-valued, because the reality condition \( \hat{\phi}_{-j} = \overline{\hat{\phi}_j} \) is obviously met. Small values of \( m \) lead to a featureless density estimate; while larger values soon lead to a complicated, 'noisy' one. This quantity is therefore a smoothing parameter; and we hope to find an optimal value between the two extremes.
Directions preferred by 76 turtles (Stephens 1969) (m = 4)

We may write our density estimate in a slightly different form in order to garner some insights into how it works. Include the formula for the estimated coefficient to get

\[ q(x) = \frac{1}{2\pi} \sum_{m=-n}^{m=n} \left( \frac{1}{n} \sum_{k=1}^{m} e^{i k x} \right) e^{-i j x} = \frac{1}{2\pi} \sum_{m=-n}^{m=n} \sum_{k=1}^{m} e^{i (k-m) x} \cdot \]

Reverse the order of summation, to get

\[ q(x) = \frac{1}{2\pi} \sum_{i,j=\pm m}^{m=n} \sum_{k=1}^{m} e^{i (k-m) x} \cdot \]

We will give a name to that inner sum: \( K_m(y) = \frac{1}{2\pi} \sum_{j=-m}^{j=m} e^{i j y} \). That finite geometric series may be summed to get

\[ K_m(y) = \frac{1}{2\pi} e^{i j y} - e^{-i j y} = \frac{1}{2\pi} \frac{\sin[(m+2)j y]}{\sin(j y)} , \text{ which is called a Poisson kernel. It is symmetric about zero, and } \int_{-\pi}^{\pi} K_m(y) dy = 1. \] The kernel becomes more concentrated near zero as m grows:
The kernel with \( m = 2 \) is the solid line; \( m = 5 \), dashed; and \( m = 10 \), dotted.

We may now write the density estimate using our new notation: 
\[
\hat{f}(x) = \frac{1}{n} \sum_{k=1}^{n} K_m(x-x_k).
\]

We now see formally how Fourier series density estimates perform their smoothing action: observations \( x_k \) are replaced by kernels centered at those points. When \( m \) is large, the result looks ‘spiky’, with spikes for each datum; as \( m \) gets smaller, the influence of each point melds with neighboring points to achieve a smooth effect. Nonparametric estimates that may be written this way, with other kernel functions possible, are called kernel density estimates. We will see much more of these later in the book.

6. Error Estimates. To investigate how well our method works, we shall as promised use the mean integrated squared error \( \text{MISE} = \mathbb{E}\left(\int [\hat{f}(x) - f(x)]^2 \, dx\right) \). Subtract and add \( \mathbb{E}[g(x)] \) inside the square and simplify to get 
\[
\text{MISE} = \int \left\{ \left[ \hat{f}(x) - \mathbb{E}[\hat{f}(x)] \right] - \left[ f(x) - \mathbb{E}[f(x)] \right] \right\}^2 \, dx + \int \left[ \mathbb{E}[\hat{f}(x)] - f(x) \right]^2 \, dx.
\]
These are familiarly written \( \text{MISE} = \int \text{Bias}^2 [\hat{f}(x)] \, dx + \int \text{Var} [\hat{f}(x)] \, dx \). We call the terms the integrated squared bias and the integrated variance. This is of course the usual decomposition of the mean-squared error of any estimate.
We shall first study the bias. Notice that, because it is an empirical expectation, \( E(\phi) = \phi \). Thus, the parameter estimates are unbiased: and all the bias in the density estimate comes from the missing Fourier series terms: \( \{\hat{f} - E[f]\} = \frac{1}{2\pi i} \sum_{m} \phi_{m} e^{-imx} \). Then our earlier work gives us the integrated squared bias \( \int_{-\pi}^{\pi} (\text{Bias}(\{f\}))^2 \ dx = \frac{1}{\pi} \sum_{m} |\phi|^2 \).

This term decreases to zero as \( m \) increases.

In a similar way, we will tackle the variances and covariances of the coefficient estimates. By independence of the observations

\[
\var(\phi) = \var(\frac{1}{n} \sum_{k=1}^{n} e^{ikx}) = \frac{1}{n} \var(e^{ikx}) = \frac{1}{n} \left[ E(e^{ikx} e^{-ikx}) - |E(e^{ikx})|^2 \right] = \frac{1}{n} (1 - |\phi|^2);
\]

and similarly, \( \text{Cov}(\phi_i, \phi_j) = \frac{1}{n} (\phi_i - \phi_j \phi \phi_i - \phi_j \phi_i) \). Now we may write

\[
\var(\{f\}) = \frac{1}{4\pi^2 n} \sum_{m} \left( |\phi|^2 \right) + \frac{1}{4\pi^2 n} \sum_{m} \sum_{m} (\phi_{m+1} - \phi_{-m}) e^{-i2mx};
\]

Then we get \( \int_{-\pi}^{\pi} \var(\{f\}) \ dx = \frac{n}{2\pi} \sum_{m} \left( 1 - |\phi|^2 \right) \). Notice that the second part of that sum converges to the finite quantity \( \int_{-\pi}^{\pi} \var(\{f\}) \ dx \) as \( m \) gets large. Therefore, the integrated variance is dominated by \( \pi \var \); and is asymptotic to it for large \( m \).

Our integrated squared bias and our integrated variance tell us a crucial fact. Fourier series density estimates are consistent in MISE if \( m \to \infty \); but sufficiently slowly that \( n \var \to \infty \).

We can get better qualitative information about how to keep MISE to a minimum, if we are willing to make some assumptions about how smooth our underlying density is. Densities are presumably smooth if their curvature, as measured perhaps by \( f' \), is usually small in size. In the spirit of least-squares, we might then measure the overall smoothness by \( \int_{-\pi}^{\pi} |f''|^2 \ dx \); which we will assume is finite. Then \( f'' \) is almost everywhere the limit of its Fourier series, which you should check is equal to \( \frac{1}{2\pi} \sum_{m=1}^{m} \phi_{m} e^{-imx} \); and so

\[
\int_{-\pi}^{\pi} |f''|^2 \ dx = \frac{1}{\pi} \sum_{m=1}^{m} |\phi|^2 \text{ converges. Then } \lim_{m \to \infty} \frac{1}{\pi} \sum_{m=1}^{m} |\phi|^2 = 0. \text{ Now replace each term in the sum by a smaller one to get } \lim_{m \to \infty} m^{1/2} \sum_{m+1}^{\pi} |\phi|^2 = 0. \text{ Comparing this to our squared bias expression for a Fourier series density estimate } \{f\}, \text{ we conclude that for sufficiently large } m, \text{ we may choose a constant } a \text{ so that } \int_{-\pi}^{\pi} (\text{Bias}(\{f\}))^2 \ dx \leq am^{-4}. \text{ Then in-}
\]
cluding our asymptotic estimate for the integrated variance, we find that 
\[ \text{MISE} \leq m \pi n \]
\[ \text{MISE} \leq m \pi n \cdot a m^{-4}. \]

We can see how to choose an \( m \) to make this upper bound as small as possible. By ordinary calculus, when \( m = (4 an)^{4/n^{4}} \), the upper bound becomes 
\[ \frac{5}{4} \pi n^{-\frac{4+1}{4}}. \]
Though this information is of limited practical value (since \( f \) is unknown, we are at a bit of a loss to find good values for \( a \)), it does suggest the rate at which \( m \) should grow compared to the sample size \( n \). We notice also that our bound does not attain the usual rate of decrease of estimate variance \( n^{-1} \) from parametric statistics.

Generally, if we are willing to assume that the measure of density smoothness 
\[ \int_0^{\pi} \left[ f^0(x) \right]^2 dx \]
is finite, then if we let \( m \) increase at a rate \( \Theta \left( \frac{n^\beta}{\pi T} \right) \), we can bound the MISE by a quantity \( \Theta \left( \frac{1}{n^{2\beta}} \right) \).

7. Other Applications of Fourier Series. Our density estimation technique applies with very little change to periodic, as opposed to directional data. Imagine that our measurements are periodic on an interval \((0, T]\) (remember our example of seasonal data in days starting January 1, so that \( T = 365.25 \)). We may estimate its density by linearly transforming the data to \((-\pi, \pi]\), estimating the density there, then transforming the density back to the original scale (this is how the author's computer programs work). But it is of some interest to see what Fourier series density estimators look like in such a case.

First notice that the interval may no longer be centered at 0 (i.e., \((-T/2, T/2]\)). This is entirely irrelevant—our estimating functions treat observations one period apart as completely equivalent. Therefore, the linear transformation to take periodic to directional data may be the pure change of scale \( Y = X 2\pi/T \). Then let a Fourier coefficient be
\[ \rho_j = \int_0^T e^{(2\pi j T/2\pi T)} f(X) dX, \]
and its canonical estimate be \( \hat{\beta}_j = \frac{1}{n} \sum_{k=1}^n e^{(2\pi j T/2\pi T) X_k} \). The Fourier series density estimate is then
\[ \hat{f}(X) = \frac{1}{n} \sum_{m=-m}^m \beta_j e^{-(2\pi j T/2\pi T) X}. \]

Once again, we may interpret the coefficient formula as a Fourier transform
\[ \rho_j = E \left[ e^{(2\pi j X/T)} \right] = \phi(2\pi j T) \]
for any integer \( j \). Now the transform is defined, not on the integers, but on a lattice of spacing \( 2\pi/T \) (that is, the array of values \((-\infty, -2\pi/T, 0, 2\pi/T, 4\pi/T, \ldots)\)). For ever larger intervals of periodicity, the Fourier transform comes to be defined on an ever-finer lattice. The inversion formula may be written
\( f(\lambda) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \rho_j e^{i(2\pi j \lambda)} (2\pi T) \). This emphasizes a sophisticated interpretation; we have an integral over the lattice with measure for each lattice point equal to its spacing \( 2\pi T \). This point of view will come in handy later.

Another interpretation of Fourier series density estimates is of considerable historical importance. Directional and periodic data are less common in practice than simple numerical measurements with no a priori periodicity. But, because Fourier series estimates are so easy to compute and work with, we are tempted to use them anyway. The usual procedure is something like this: choose an interval \((a, b]\) that generously includes all the observations. Now pretend that the observations are periodic with this base period, and construct a density estimate. The result will start over below \( a \) and above \( b \); but presumably we do not care about density estimates so far from the observed data. We simply ignore the estimates outside that range.

As we observed earlier, the only thing important about the choice of interval \((a, b]\) is its length \( T = b - a \). Therefore, we compute the density estimate by 

\[
\hat{\rho}_j = \frac{1}{T} \sum_{k=1}^{\infty} e^{i(2\pi j (b-a))} x_k
\]

and then 

\[
f(\lambda) = \frac{1}{b-a} \sum_{j=-m}^{m} \hat{\rho}_j e^{-i(2\pi j (b-a)) x}. \]

Of course, this is based on a false assumption about the data; and we could presumably do better if we did not make that assumption. We will see how, in a later chapter.

%Silica in \( n = 22 \) chondrites, \( m = 4 \), interval = 18—36

This example comes from Ahrens (1965). The silica percentage in the 22 meteorites ranged from 20.77 to 33.4. The artificial interval of periodicity was set to 18. Notice the steep negative dip at 19%, which has no statistical interpretation. In general, Fourier se-
ries density estimates have sinusoidal wiggles in any wide interval devoid of data (look at
the extremes of the Poisson kernel to see why). Good and Gaskins (1980) call this the
“Nessie effect” (after the purportedly serpentine Loch Ness monster). When periodicity is
artificial, this may be reduced by narrowing the interval to barely include the data. The
cost is that very large and very small observations begin to inappropriately influence each
other’s estimates (because the estimate wraps around). In signal processing this is called
aliasing. At the other extreme, very wide intervals require the computational effort of
larger ms, and show the Nessie effect. Wide intervals are in fact better, but paradoxically
look worse. We shall see these problems again, and shall see other density estimation
methods that avoid them.

8. Empirical Densities. When we approximated a known function \( f \), we noted that
under mild conditions the integrated squared error in a Fourier series approximation went
to zero when we let \( m \) go to infinity. So what is the meaning of the density estimate
\[
\hat{f}(\chi) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \hat{f}_j e^{ij\chi},
\]
in which no smoothing takes place? We will call this expression the
empirical density. From our error analysis, it has no bias but infinite variance; and so
has no practical application. Rewrite it in kernel form
\[
\hat{f}(\chi) = \frac{1}{n} \sum_{k=1}^{n} K_{\infty}(\chi-x_k)
\]
where
\[
K_{\infty}(\gamma) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} e^{ij\gamma}.
\]
By extension of the case with finite \( m \), this is an arbitrarily thin and
high spike centered at 0, whose value goes to zero everywhere else. Yet, its integral must
be one. This is not in the usual sense a function; but formally it is a (periodic) Dirac func-
tional at zero. Such objects are called generalized functions. Therefore, the empirical density
is a linear combination of these Dirac spikes, each centered at a data point.

The essential property of the Dirac function is that for any periodic, square integrable
function \( g \) that is continuous at 0,
\[
\int_{\pi}^{\pi} q(\chi)K_{\infty}(\chi-x)dx = \int_{\pi}^{\pi} q(\chi) \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} e^{ij\chi}dx = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \int_{\pi}^{\pi} q(\chi) e^{ij\chi}dx = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \phi_j.
\]
But, staring at that last sum and then at the inversion theorem, \( \int_{\pi}^{\pi} q(\chi)K_{\infty}(\chi-x)dx = q(0) \).
Then by a linear change of variables \( \int_{\pi}^{\pi} q(\chi)K_{\infty}(x-y)dx = q(y) \) at any point of continuity \( y \).
For this reason, the Dirac functional is called an evaluation functional.

This suggests the following formal manipulation: express the general least-squares
density estimation problem as
\[
\min_{g} \int \left( \hat{f} - g \right)^2 = \int \hat{f}^2 - 2\int \hat{f} g + \int g^2 \text{ where } \hat{f} \text{ is our empirical density. As in the Preface, } \int \hat{f}^2 \text{ contains no } g \text{ and so is irrelevant to the minimization.}
\]
Reverse signs to get the equivalent problem
\[
\max_{g} \int 2\hat{f} g - \int g^2. \text{ But expressing } \hat{f} \text{ in terms}
of a linear combination of Dirac functionals, and using the results from the last paragraph; we get \( g = \frac{1}{n} \sum_{k=1}^{n} g(x) \). Then our least-squares criterion becomes \( \max_{g} \frac{1}{n} \sum_{k=1}^{n} d(x) - \int g^2 \), exactly as proposed in the Preface and applied earlier in this chapter.

Note, however, that this derivation is mathematically preposterous: \( \int g^2 \) does not exist, and should be thought of as infinite! We are minimizing a distance measure that is always infinite. Nevertheless, this is such a convenient alternative derivation that we shall often pretend to take it seriously. David Scott once commented to me that it sounded like the formal process of renormalizing by canceling infinite quantities that is important in quantum electrodynamics.

We find ourselves with two ways of looking at least-squares density estimation: (1) we are minimizing the estimated distance to the true density, or (2) we are directly minimizing the (infinite) distance to the empirical distribution of a random sample. The mathematics involved is identical.

9. Chapter Summary. As promised, no new ground has been broken in this chapter. However, we have, in a familiar context, introduced a number of the themes that will be most important in the rest of the book.

— The a priori support of a density (the circle, the real line, etc.) plays a fundamental role in its estimation.
— A version of the method of least squares is a powerful tool for estimating densities.
— The Fourier transform often aids in the solution of these least-squares problems. (In time series theory, this corresponds to studying problems in the frequency domain.)
— Orthogonal series are sometimes useful in solving our problems. (This has been called the Rayleigh-Ritz method.)
— An alternative way of looking at density estimates, the kernel representation, helps solve problems and study their solutions. (Time series theorists call this point of view the time domain.)
— MISE is the fundamental measure of how good a least-squares density estimate is. It decomposes into an integrated bias-squared term and an integrated variance term; and design decisions (the smoothing parameter) usually force trade-offs of these two quantities.
— Measures of smoothness of the underlying density like \( \int [f'(x)]^2 \) constrain the quality of least-squares density estimates, and suggest the choice of good ones.
— A somewhat ethereal object, the empirical density \( \hat{f} \), aids in the formulation of density estimation problems.