Preface

1. Introduction. In this era of machine data recording and huge social data bases, we often find ourselves with big collections of measurements under similar conditions, for which we have no simple probability model. Of course, we wish to do something exploratory, and usually graphical, to help us guess the model, or at least important model features. For example, perhaps there were several sub-populations represented that we did not know about? Techniques for seeking such goals are called nonparametric estimation of the probability law underlying a sample of observations.

Of course, classical estimation gives us a fine solution to that problem: the empirical cumulative distribution function is known to be the uniformly minimum variance unbiased estimate of the c.d.f. (pointwise, it is just a binomial proportion). So what is our problem? If the law has a density, we would like to know what it is, if only because of the common experience that the graph of a density is easier to interpret visually than the graph of a c.d.f. But the empirical c.d.f. has no associated density, because it is a step function and cannot be differentiated in any space of ordinary functions. So how are we to do nonparametric density estimation?

Of course, statisticians have done this informally for several centuries. The familiar block chart, or histogram, may be thought of as a density estimate. David Scott (1979) has made this formal. But statisticians had long intuited something of the nature and limitations of the histogram as a density estimate. The modern era of attempts to improve on it perhaps should begin with Rosenblatt(1956), who both invented new nonparametric density estimates and established the most important constraints under which they must operate. In the ensuing decades, a cornucopia of new density methods came to be proposed, with names like kernel methods, orthogonal series methods, penalized likelihood methods, and spline methods. A good survey of these techniques may be found in Scott (1992).

Grace Wahba (1981) has argued that enough is enough. We have a practical plethora of methods; it was time to concentrate on issues of implementation. In particular, all nonparametric density estimates have, explicitly or implicitly, a smoothing parameter that dominates their properties. The archetype for these is the bin width of a histogram. If this width is small, the heights of the blocks vary wildly, as by chance observations concentrate in certain neighborhoods. The resulting graph looks like the skyline of a great city. By consensus, it is too noisy to be as useful as it should be. If the width is too large, there are only a handful of blocks. The graph is not capable of displaying potentially interesting local features in the density. The choice of smoothing parameter is therefore critical to any practical application of nonparametric density estimation.
In recent years, the problem of smoothing parameter selection has dominated the density estimation research literature. For a survey, see Sheather (1992). It is a matter of considerable controversy how much useful progress has been made by this program (see my discussion of Sheather’s paper, Terrell [1992]).

This monograph takes a very different point of view. That enormous variety of density estimation methods is not an occasion for complacency; rather, it is of itself an important obstruction to the usability of the technology. Potential users are intimidated by the choice, and they understandable fear that any selection they make will tread on the toes of several competing schools. Long ago it was realized that objective measures of superiority do not solve this problem. Wildly different methods are known to be very similar in asymptotic performance; and close to the theoretical performance limits for some classes of densities.

We will attempt to show here that much of the variety in nonparametric density estimation methods is illusory. Most of the most commonly-used methods are in fact the method of least-squares, in disguise. The differences between methods arise when certain natural choices are made, in the support of the estimated density, in the computational approximations made for convenience, and in the sort of smoothness that one expects of the estimated density. The user can decide which density estimation technique to use, from the nature of the application and the uses to which the answer will be put.

Consider the familiar least-squares criterion for regression: given a set of independent variable values $x_i$, and a corresponding set of dependent variable values to be predicted $y_i$, we look for a function connecting them $\hat{y} = g(x)$. We choose that function $g$ from among some class (such as polynomials) or by its degree of smoothness, by solving

$$\min_g \sum_{i=1}^n (y_i - g(x_i))^2 \text{ (or some weighted variation).}$$

In density estimation, we want an estimated density function $\hat{f} = \hat{f}(x)$ where $f$ is the true density. But to approximate the whole function, rather than just at a set of design points, we are inspired to use an integral criterion: $\min_{\hat{f}} \int [f(x) - \hat{f}(x)]^2 dx$, where the integral is over the support of the random variable. This is not helpful, though, because $f$ is of course unknown. All we have is a random sample $x_i$.

Matts Rudemo (1982) and Adrian Bowman (1984) overcame this difficulty in the following ingenious way: expand that integral criterion into three terms

$$\min_{\hat{f}} \left\{ \int f(x)^2 dx - 2 \int f(x) \hat{f}(x) dx + \int \hat{f}(x)^2 dx \right\}.$$ 

Now notice that the first term, though unknown, is irrelevant to the process of finding a minimizer, since it does not involve $\hat{f}$. The third term may be computed at any time. The integral in the second may be written
This is not known; but we see that there is a canonical estimate of it from our random sample, the empirical expectation $E_n[\hat{X}] = \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i$. Ignoring the irrelevant term, we have a natural estimate of our least-squares approximation if we minimize $-\frac{2}{n} \sum_{i=1}^{n} \hat{X}_i + \int \hat{X}^2 dx$ over some suitable class of functions $g$. So that the quantities involved shall generally be positive, we shall let the least-squares criterion for a nonparametric density estimate be

$$\max_g \left[ \frac{2}{n} \sum_{i=1}^{n} \hat{X}_i - \int \hat{X}^2 dx \right].$$

Much of the rest of the monograph will be dedicated to solving this criterion under various constraints, penalties, and classes of estimates $g$; and to exploring the properties of these solutions. These will be called linear density estimates, for reasons that will become clear.

This proposal is anything but out of the blue. The traditional study of properties of density estimators has assumed that good ones should minimize the mean integrated squared error $\text{MISE} = E \left\{ \int (f - \hat{f})^2 dx \right\}$, (or convenient asymptotic approximations $A \text{MISE}$) where the expectation is taken over the sample space of possible random samples for an assumed fixed true density $f$. This has allowed us to study questions of the, on average, best values of smoothing parameters and best shapes of kernel functions. These criteria have usually been justified by mathematical convenience; but we can see that they are in some sense natural ways to study least-squares density estimates.

Furthermore, Rudemo and Bowman invented their trick in the course of developing least-squares cross-validation, a method for letting data determine the smoothing parameter for such methods as histograms and kernel estimates. This has been the starting point for most recent research in the area. This monograph will spend little time on the issue of choosing a good smoothing parameter; but the intimate connection with cross-validation suggests that there are important implications here for future research.

By no means are all important density estimators linear in our sense. For example, penalized maximum likelihood estimates (Good and Gaskins 1971) and log-spline estimates (Stone, Buja, and Hastie 1994) are not. The later chapters of this monograph will extend the applicability of our techniques by showing that certain estimates fitted by the criterion of maximum likelihood are similar to least-squares estimates. These are the almost-linear methods of the title.
2. Outline. The book is organized as follows:

Chapter 1: Directional Data. In this overture to the rest of the book, we give an elementary introduction to a classical method of density estimation, that of Fourier series. In the process, most of the important themes of the rest of the book are introduced, in a familiar context.

Chapter 2: B-Splines. Histograms are only the simplest of a family of methods, linear B-spline methods, that may be estimated by least-squares; these provide remarkably compact representations of density estimates on the entire real line. We show them to have asymptotic properties competitive with more familiar estimators.

Chapter 3: Roughness Penalties. Least-squares methods whose solution is penalized for large values of the integrated square of a derivative lead to tapered Fourier-series estimates for directional data and to certain regular kernel methods on the real line. Linear combinations of roughness penalties, which we call Sobolev penalties, greatly generalize this approach. Then we show that a number of properties of familiar density estimators follow immediately from their characterization as solutions of least-squares problems.

Chapter 4: Regular Kernels. We establish that Klonias-type roughness penalties are essentially coextensive with kernel density estimates. This makes the asymptotic properties of linear methods transparent.

Chapter 5: Restricted Support. One obvious advantage of B-spline methods is that they apply with no difficulty to estimation when there are known boundaries for the random variable. Then we note that roughness penalties lead to natural solutions of many boundary-kernel problems.

Chapter 6: Variable Smoothing. The important and difficult problem of varying the degree of smoothing over the support has a straightforward solution for B-spline estimates. Furthermore, variably weighting their roughness penalties leads to natural solutions for kernel-type estimators.

Chapter 7: Discretization. Restricting data to a fine, regular mesh (which used to be called grouping) leads to a variety of computationally easy methods for smoothing. Essentially, we generalize and dualize ASH and FFT methods. Fourier analysis aids in the understanding and application of these finite difference methods.
Chapter 8: Advanced Discretization. Finite difference methods lead to fast solutions of restricted support and variable smoothing problems. Finite element methods, which are just roughness-penalized B-spline methods on finer meshes, also lead to tractable and attractive approximate solutions of many of our linear density estimation problems. (in preparation)

Chapter 9: Multivariate Methods. All the techniques of earlier chapters generalize readily to data in several dimensions. Fourier series and B-spline methods confront the curse of dimensionality. Penalty methods turn out to require higher-derivative terms in order to have finite solutions. (in preparation)

Chapter 10: (Donggeon Kim) Least-squares Mixture Decomposition. Professor Kim shows that the method of least-squares leads to a finite mixture decomposition algorithm that, unlike traditional methods, always converges. Furthermore, the data tend to choose the number of components in a mixture.

Chapter 11: Least-squares estimation of a square-root of a density is asymptotically equivalent to maximum likelihood. (in preparation)

Chapter 12: Therefore, many of the linear methods above may, suitably generalized, be applied to nonlinear maximum likelihood problems. (in preparation)

3. Purpose of the Book. This is intended to be a research monograph. I hope here to show the community interested in nonparametric density estimation the power and promise of one particular point of view, that of least-squares.

This is not a survey of the status of research in this area. Other books, e.g. Scott (1992) do that far better, and have far more useful bibliographies. The focus here is on recent work by the author.

It is not a textbook. However, it was written to be elementary, approachable, and self-contained. Early chapters have much that is not new (and perhaps better explained elsewhere). My purpose was to make it readable by not-very-sophisticated graduate students with a minimum of reliance on other resources.

It is not a handbook for people who wish to apply density-estimation technology. There are a number of books on the market better for that purpose; e.g. Silverman
(1986). It is, however, written at a level low enough and with enough examples and comments on application that it could conceivably be used for that purpose.

It by no means “cleans out” the areas it discusses. It is full of partial solutions of important problems; and of suggestions for future research.

It is not a work of mathematics. The reader will find many informal derivations, and few theorems. The mathematically inclined may perhaps treat it as a series of hints of theorems to be carefully stated and proven.

The author looks forward to hearing from readers who have indeed found entertaining and useful directions to take the work that is only begun here.