Journal of Official Statistics
Vol. 7. No. 2, 1991, pp. 219–233
© Statistics Sweden

Model-Free Curve Estimation: Mutuality and Disparity of Approaches

Michael E. Tarter and Michael D. Lock[1]

Abstract: This paper is written to provide methodological background for researchers interested in applying curve estimation to fields such as environmental health. Basic approaches are introduced with special emphasis on shared features which may be of value in environmental and other investigations. Completeness and generality from the viewpoint of curve estimation are described as are new applications to nonparametric inference and mixture decomposition. Series, kernel, and penalized likelihood methodologies are compared as are different metrics and methods of counterbalancing representational complexity with data availability.

Curve estimation methodology is illustrated as a way of uncovering distributional bimodality. The danger inherent in relying on conventional parametric procedures is demonstrated by the case of an anomalous model, the log-Cauchy, which gives the false appearance of being a mixture model. The potential value of the new approach is illustrated by hybrid procedures which combine nonparametric estimation and rank-based inferential methodology.

Key words: Bump-hunting; Chi-square goodness-of-fit; decomposition; kernels; mean integrated square error; mixing parameters; multipliers; penalized likelihood; series.

1. Introduction

The distinction between the generalized, model-free or nonparametric, and model-based approaches to curve estimation can be illustrated as follows: Suppose a researcher were to base representations of the curves or functions encountered in environmental studies (for example, dose response regression curves), or clinical trials (for example, survival curves), or in other statistical studies (for example, estimated probability density functions (pdfs), cumulative distribution functions (cdfs) or quantile-quantile (Q-Q) plots) on the $\{\exp(2\pi ikx)\}, k = 0, \pm 1, \pm 2, \ldots$ orthogonal sequence. Suppose a second investigator were to base investigations either on the normal model $N(y|\mu, \sigma) = (2\pi\sigma)^{-1/2} \times \exp\{-1/2[(y - \mu)/\sigma]^2\}$ or a model like the lognormal $F(y) = \Phi[\{\log(y - \mu_1) - \mu_2\}/\sigma]$, where $y > \mu_1$ and $\Phi(y)$ represents the partial integral of $N(z|0, 1)$ from $-\infty$ to y. With sufficient data, the first researcher will eventually get the right answer to any properly posed statistical problem. The

[1] Department of Biomedical and Environmental Health Sciences, University of California, Berkeley, California, 94720, U.S.A.

Acknowledgements: The authors would like to thank W. Freeman for technical assistance, the referees for many useful comments, and C. Mellin for guiding the many stages of our project to completion. This work was supported by U.S. Environmental Protection Agency Grant R-813766 and National Institutes of Health, National Institute of Environmental Health Sciences Grant 1 R01 ES 05379-01.

researcher using the model-based approach could also arrive at the right answer. But no matter how much data were available, he/she could still be misled by an initial choice of the wrong model.

In today's curve estimation literature, there are almost as many alternatives to the use of the $\{\exp(2\pi ikx)\}$ system of orthogonal functions as there are alternatives to the choice of lognormal or normal models within the classical literature. For example, in Good and Gaskins (1980), *both* the $\{\exp(2\pi ikx)\}$ and the Hermite system of orthogonal functions were extensively applied to the important problem referred to by Good and Gaskins as series-based "bump-hunting." Anderson (1969), Diggle and Hall (1986), and Hall (1980, 1982) have considered various aspects of estimation based on many different orthogonal systems including the Laguerre, Legendre, and cosine functions. The fields of spline as well as kernel-based curve estimators also deal with a large spectrum of choices for the kernel function K or spline representation.

Given the diverse spectrum of choices, why is the selection of one particular curve estimation alternative different from the selection of a single model in the conventional model-based approach? The answer to this question can be stated in one word: "complete." For example, the $\{\exp(2\pi ikx)\}$ system of functions forms a complete orthogonal sequence for L^2. Generally speaking, L^2 includes almost all of the curves any practicing statistician is likely to encounter. The importance of completeness is established through the following theorem:

> Given a complete orthonormal sequence $\{\Psi_k(x)\}$ in L^2, every element $f(x)$ of the space L^2 admits an expansion, convergent in the mean (Sz.-Nagy 1965).

Thus, completeness is generality. However, this is not to say that all methods, all choices of orthogonal function, kernel, spline, or metric will yield identical estimators. Specifically, from a statistical point of view there can be differences in data-use-efficiency and bias. Nevertheless, the mathematical problem of representation is concerned with the question: Can a specific function or curve correspond to a given expression? Completeness implies that the function or curve can be adequately represented.

Lack of completeness leads to what Tapia and Thompson (1978) refer to as "problems of specification." Such problems involve the following two basic questions: (1) How likely is it that the curve a scientist is actually estimating can, like the normal density, be expressed as an elementary function? (Elementary functions are x^k, $\exp(kx)$, $\sin(kx)$, $\cos(kx)$, their inverses and composites.) (2) Given that the curve can be expressed as an elementary function, how likely is it that the scientist will know what particular elementary function represents the curve?

For example, a mathematical proof that the normal inverse cumulative, Φ^{-1}, cannot be expressed as an elementary function was presented by Rosenlicht (1975). In contrast to the normal inverse cumulative, completeness of an orthogonal system like $\{\exp(2\pi ikx)\}$ guarantees that a form of representation will be adequate for any given L^p curve, where $p > 1$ (Carleson 1966). (That a curve is L^p implies in practice that the area under the pth power of the curve is finite and well-defined.)

Besides most examples of what are now called series methods, the generality of the $\{\exp(2\pi ikx)\}$ system applies to a distinct and commonly used type of model-free approach. For most applied purposes, kernel nonparametric estimators (Silverman 1986) can be represented in terms of the Fourier coefficients of the kernel K. Thus the close connection between series representations based on $\{\exp(2\pi ikx)\}$ system on the one hand, and kernel representation on the other hand, allows one to conceptualize problems using whichever of two approaches is most suitable for a specific application. As data availability increases, either series or kernel curve estimators have the potential to approach (in some clearly defined sense) the estimated curve.

2. A Methodological Summary of Model-Free Methods

Application of many model-free curve estimators involves the following three choices:

1. A mathematical representation for the curve is selected. Series and kernel representations are commonly used, as are sequences of splines joined so that a specified number of derivatives are equal at the intersection of consecutive splines.

2. A metric, error function or roughness criterion is chosen which defines the curve estimator's goodness-of-fit. An example of such a metric is the mean integrated square error (MISE):

$$J(H^*, H, w) = E\int \{H^*(x) - H(x)\}^2 \times w(x)dx$$

where H is some targeted curve, H^* represents a curve estimator of H, and w represents a weight function designed to emphasize or de-emphasize particular regions of the underlying random variate's support.

When $H^* \equiv f^*$ represents the observed proportions and $H \equiv f$ the expected proportions of data points grouped within a sequence of class intervals, the metric $J(f^*, f, f^{-1})$ can be interpreted as the property estimated by the usual Chi-square goodness-of-fit test statistic. Alternatively, the metric $J(f^*, f, (f^*)^{-1})$ corresponds to expression (77) of Neyman (1949), Neyman's BAN alternative Chi-square.

In much of the curve estimation literature, the weight function w is chosen to be one. This departure from the usual Chi-square contingency table and goodness-of-fit approach is motivated primarily by the ease with which the metric $J(f^*, f, 1)$ can be applied with both Fourier series and kernel representation. For example, when $w \equiv 1$ Parseval's Theorem (Sz.-Nagy 1965) can be used to show that for an extremely general class of functions f, J can be expressed in terms of the squared moduli of the Fourier coefficients of individual series terms.

3. A bandwidth, class interval size, degree of abridgement, multiplier sequence, method of tapering, series term stopping rule, term inclusion rule, or some other means is selected to counterbalance representational complexity or smoothness with data availability. The capability to systematically deal with counterbalancing is critical to the curve estimation approach.

The remainder of this section consists of a brief comparison of choices of estimators and forms of representation, for example, kernel and Fourier series, and was written to introduce some important concepts and terminology. Suppose that $X_1, \ldots, X_n$ are iid random variables with probability density f. One kernel estimator of the density f is defined as

$$\hat{f}(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)$$

where K is called the kernel function and h is called the smoothing parameter or bandwidth. The kernel function is commonly selected a priori from a class of nonnegative and symmetric elementary functions such that

$$\int K(x)dx = 1 \text{ and } \int xK(x)dx = 0$$

(Silverman 1986).

In many instances (particularly in the early curve estimation literature) K itself is a density function such as the standard normal, $N(x|0, 1)$. An estimate constructed with a kernel satisfying these conditions will itself be a density.

Assume f is an L^2 function with support on [0, 1]. (For densities with support on some other finite region, data can be rescaled to lie within [0, 1] by a simple linear transformation.) Then, under very general conditions f can be represented by a Fourier series expansion:

$$f(x) = \sum_{k=-\infty}^{\infty} B_k \Psi_k(x)$$

where $\Psi_k(x) = \exp(2\pi ikx)$, conjugate $\Psi_k^*(x) = \Psi_{-k}(x)$, and

$$B_k = \int_0^1 f(x)\Psi_{-k}(x)dx.$$

The Fourier series estimator of f can be defined as

$$\hat{f}(x) = \sum_{k\in M} b_k \hat{B}_k \Psi_k(x)$$

where the kth sample Fourier coefficient, $\hat{B}_k = n^{-1}\Sigma_{j=1}^{n}\Psi_{-k}(X_j)$, is an unbiased estimator of B_k, M is a set of indices (possibly an infinite set), and $\{b_k\}$ is a sequence of real-valued multipliers chosen to optimize the estimator in some respect. For example, a multiplier sequence might be chosen to minimize the MISE, $J(\hat{f}, f, w)$.

There are so many different uses of multiplier sequences that applications of generalized statistical methods can often be interpreted in terms of a single problem: the selection or estimation of a multiplier sequence. That multipliers should play such an important role in statistical practice is indicated by the crucial role this topic plays in both applied and theoretical mathematics. For example, Larsen (1970) contains an 18 page bibliography which lists more than 300 papers on the topic of multipliers.

Both kernel and Fourier series estimators belong to a general class of procedures known as general weight function estimators (Whittle 1958), which are of the form

$$\hat{f}(x) = n^{-1}\sum_{j=1}^{n} W(X_j, x)$$

where W satisfies $\int W(x)dx = 1$. For kernel estimators, $W(X_j, x) = h^{-1}K[(x - X_j)/h]$, and for Fourier series estimators $W(X_j, x) = \Sigma_{k\in M} b_k \Psi_k(x - X_j)$.

If $b_k = b_{-k}$ and $\Sigma_{k=-\infty}^{\infty} b_k < \infty$, then the general weight function form of the Fourier series estimator is

$$\hat{f}(x) = n^{-1}\sum_{j=1}^{n}\sum_{k\in M} b_k\Psi_k(x - X_j)$$
$$= n^{-1}\sum_{j=1}^{n} K(x - X_j).$$

Hence, a series estimator of f can be written as a kernel estimator whose multipliers $\{b_k\}$ are the Fourier coefficients of the kernel K. Also, by using expressions for the characteristic function of truncated densities given by Kronmal and Tarter (1968) many kernel estimators can be expressed as Fourier series estimators.

Due to the periodicity of Fourier expansions, it may seem surprising that so many kernel estimators can be effectively expressed as Fourier series with particular multiplier sequences $\{b_k\}$. For example, expansions based on the $\{\exp(2\pi ikx)\}$ system must take on equal values at zero and at one. The seeming contradiction between the cyclical nature and the generality of Fourier series was first posed by the mathematician Euler with regard to the problem of representing the instantaneous position of a vibrating string stretched within the interval AB. In the literature of the history of mathematics it is referred to as "the periodicity argument" (Grattan–Guinness 1970, p. 10).

In his basic text on the history of analysis, Grattan–Guinness points out that the current trend among mathematicians is to dismiss Euler's periodicity argument. Only representation inside the interval AB is deemed to be critically important, so that the representation repeats itself outside AB is irrelevant. The same reasoning applies to the problem of nonparametically estimating a curve outside the data range. While it is not without some practical importance, the statistical problem of model-free extrapolation is speculative.

A frequently cited distinction between kernel and Fourier series estimates is that kernel estimates are always nonnegative while series estimates can take on negative values. However, kernel estimates are guaranteed to be nonnegative only if the kernel function is restricted to be nonnegative. Müller (1984) investigated density estimates with this restriction relaxed and showed that they can take on negative values. The condition that an estimator be everywhere nonnegative has rarely been emphasized by series researchers, possibly because there may be support subregions where one simply does not have sufficient information to trust any general estimation procedure. One could, however, assure estimator nonnegativity if one requires the multiplier sequence $\{b_k\}$ to satisfy $\Sigma_{k,j} b_{k-j} B_k B_j^* \geqslant 0$, for the square summable sequence $\{B_k\}$ (Anderson and de Figueiredo 1980). Thus, for both series and kernel methodology, estimator nonnegativity is within the control of the investigator and is not a condition intrinsic to either representation.

There is a subtle, yet crucial, distinction between the kernel and the series estimator approaches. In the kernel density estimation process the smoothing parameter h is explicitly defined to be a multiplicative scaling factor of the kernel K. Thus, K and h have distinct roles; K controls the shape of the kernel and h determines its spread. It is this duality which provides the rationale for investigators who choose a kernel function a priori, and then rely on a data-based procedure to choose the single bandwidth or smoothing parameter, h, subject to an optimality criterion (Rudemo 1982).

In comparison to h-based kernel methodology, the goodness-of-fit of the Fourier series estimator need not be determined by a single parameter but can instead be governed by the entire multiplier sequence $\{b_k\}$. This lack of dependence upon a single explicitly specified smoothing parameter has allowed researchers to experiment with a wide variety of strategies for controlling the smoothness of an estimate. For example, Hart (1985) as well as Diggle and Hall (1986) have suggested data-dependent methods for choosing an optimal number of terms, m, to include in the Fourier series estimate (in the sense of minimizing MISE). This corresponds to choosing the set of multipliers defined by $b_k = 1$, $|k| \leqslant m$; $b_k = 0$, $|k| > m$, and results in what Wahba (1981) refers to as the "raw" Fourier series estimator

$$\hat{f}(x) = \sum_{k=-m}^{m} \hat{B}_k \Psi_k(x).$$

Here the truncation point m controls the amount of smoothing, performing a role analogous to that of the smoothing parameter h of the kernel estimator.

An alternative multiplier sequence of the form

$$b_k = \frac{1}{(1 + \lambda(2\pi k)^{2p})},$$

$$\lambda \geqslant 0, p > 1/2, |k| \leqslant n/2;$$

$$b_k = 0, |k| > n/2$$

has been suggested by Wahba (1981). The Wahba procedure fixes the number of terms used in the curve representation and then minimizes an estimator of the MISE weight with respect to the dual parameters λ and p. Brunk (1978) proposes the multiplier $b_k = n/(n - 1 + \pi_k)$ where π_k is a measure of precision based on prior information. He notes that the kernel resulting from the use of this multiplier sequence depends on the specification of a prior distribution.

Rather than specifying smoothness by one, two or any fixed number of parameters, Watson (1969) as well as Fellner and Tarter (1971) have considered each term of the multiplier sequence individually. Since a multiplier sequence also defines a particular kernel, this method simultaneously estimates the shape, spread, and all other properties of the kernel. How to shape or smooth is of fundamental importance to the field of curve estimation. This question has been the focus of much research over the last twenty years and has resulted in a myriad of competing and interconnected criteria and methods, several of which are discussed in Silverman (1986). Comparisons by means of computer simulation are made in Bowman (1985) and Scott and Factor (1981).

Many strategies and criteria for smoothing can be categorized into the following two basic classes: Class 1 is based on the minimization of a predetermined measure of error, such as MISE. Examples of such methods include least squares cross validation (Bowman 1984; Hall 1983; Rudemo 1982) and the closely related generalized cross validation procedures (Craven and Wahba 1979; Wahba 1977), as well as an iterative approach which seeks to estimate $\int f''(x)^2 dx$, a quantity which appears in the expression for the smoothing parameter which theoretically minimizes the MISE of the kernel estimator (Scott, Tapia, and Thompson 1977). Class 2 is based on variants of maximum likelihood. Two such methods are likelihood cross validation (Duin 1976; Habbema, Hermans, and van der Broek 1974), and the penalized approach (Good and Gaskins 1971, 1980; Tapia and Thompson 1978).

Although most smoothing procedures are applicable to both kernel and Fourier series representation the $\{\exp(2\pi ikx)\}$ Fourier series estimator leads naturally to methods based on minimizing MISE. For example, for the case of an m-term truncated series estimator the MISE can be simply expressed in terms of the Fourier coefficients of the unknown density:

$$J(\hat{f}, f, 1) = n^{-1} \sum_{|k| \leqslant m} b_k^2(1 - |B_k|^2) + \sum_{|k| > m} (1 - b_k)^2 |B_k|^2.$$

Since each individual Fourier coefficient B_k is easy to estimate, a natural estimator of $J(\hat{f}, f, 1)$ can be readily constructed. Smoothing procedures based on such estimates have been described by Davis (1977), Diggle and Hall (1986), Fellner and Tarter (1971), Hart (1985), Kronmal and Tarter (1968), Tarter (1979a), and Wahba (1981).

3. Bimodality and the Mixture Model

There is as close a relationship between generalized statistical representation and mixture decomposition. As described by Särndal (1971), it was investigations conducted by Lexis (1877) and other pioneer statisticians involving mixtures that first shook the foundations of purely normal density-based statistics. Therefore, it is not surprising that one of the most interesting and useful applications of model-free curve estimation methodology is the problem of bump-hunting (Good and Gaskins 1980). This problem can be considered a special case of mixture decomposition where the mixing parameter p of some distributional component is of moderate size. Bump-hunting, cluster analysis, or mixture decomposition methods are often used to investigate the possible existence of dichotomous or polychotomous "hidden" variates. Once the existence of such a variate or variates is suspected, one means of corroboration involves the adding of this variate to one's model, and then a subsequent check upon whether or not the bump disappears.

A weakness of this approach is the possibility that one has chosen the correct variate but, by using an overly simple covariate or partialing model, one has failed to remove a bump. This is a multivariate regression problem. Yet even with univariate data, curve estimation investigations have uncovered subtle, but important properties of bump-hunting, cluster analysis and mixture decomposition processes. As indicated below, these features can be used to construct improved curve estimators.

Consider the lack of conceptual identity between bimodality and the two-component mixture model. The former dates from Pearson's analysis of crab frontal breadth measurement data, whose distribution was characterized by this statistical pioneer as having a "double hump" (Kevles 1985, p. 28). To analyze this type of data, Pearson proposed the mixture model, which is primarily used in the normal case (Henna 1985).

A special case of Pearson's model is the "parameter one-half mixture"

$$\begin{aligned}\gamma(x) &= \tfrac{1}{2}[N(x\,|-1/2, \sigma) + N(x\,|\,1/2, \sigma)]\\ &= K[e^{x/2\sigma^2} + e^{-x/2\sigma^2}]e^{-(x^2+1/4)/2\sigma^2}\end{aligned}$$

where the constant K assures that $\int_{-\infty}^{\infty}\gamma(x)dx = 1$. The derivative $\gamma'(x)$ equals

$$K\left[\frac{(e^{x/2\sigma^2} - e^{-x/2\sigma^2})e^{-(x^2+1/4)/2\sigma^2}}{2\sigma^2}\right] - Kx\left[\frac{(e^{x/2\sigma^2} + e^{-x/2\sigma^2})e^{-(x^2+1/4)/2\sigma^2}}{\sigma^2}\right].$$

For the one-half parameter mixture, by symmetry, the point at which the mode or the local minimum between two modes occurs must be $x = 0$. Thus, by evaluating the second derivative $\gamma''(x)$ at zero one finds that for any value of the parameter σ which is greater than one-half, the density $\gamma(x)$ will be unimodal, while for any value of the parameter σ which is less than one-half, the density will be bimodal.

The relationship between the shape of $\gamma(x)$ and σ illustrates the distinction between multimodality and the mixture model. That a mixture model like $\gamma(x)$ need not be multimodal was probably so obvious to Karl Pearson that he did not bother to make a point of this lack of identity. However, the following simple example demonstrates a slightly less obvious, but equally important, point: Bimodality need not be associated with some mixture model.

Let the function $g(x)$ be defined as

$$g(x) = \sigma p(x - \mu_1) \times \left\{1 + \left[\frac{\log(x - \mu_1) - \mu_2}{\sigma}\right]^2\right\},$$

$\sigma > 0.$

For $x > \mu_1$, the log-Cauchy model, which is analogous tu the lognormal model, can be defined as $f(x) = [g(x)]^{-1}$. By taking the first and second derivatives of g it can be shown that whenever σ is less than one, there will be a trough between *two* modes of the log-Cauchy density.

As discussed in detail by Särndal (1971), the work of Lexis (1877), and as discussed by Kevles (1985), the investigations of Pearson (1894) set the following pattern for many statistical investigations: If a simple elementary function model such as the binomial or normal was deemed to be inadequate for the representation of a statistical density then a mixture of such models seemed called for. Put concisely, mixtures were viewed as a means with which to generalize statistical methodology. Such a generalization seemed particularly relevant when bimodal, or what Pearson called "double-humped", densities were estimated.

The log-Cauchy illustrates that it will not necessarily be the case that positing a mixture model or even dividing the support region of a density into two contiguous subintervals will improve representation capability. Unlike the danger of an approach towards generalization and completeness which relies on the mixture model, curve estimation methodology can be used to represent either separate parts, or the composite whole, of any L^2 density.

Notice that the above claim uses the word *representation* rather than the word *estimation*. Separated parts of a mixture can be individually estimated by curve estimation procedures, and then the whole composite density estimated by using a composition of these modularly estimated parts (Tarter 1979b, sec. 4). A similar modular approach was used by Tarter, Freeman, and Polissar (1990) to greatly enhance the long-term survival capabilities of life-table and Kaplan–Meier procedures.

For statistics, the problem of mixtures has played a role very similar to the role played by the problem of discontinuous function representation in mathematics. Except for the point zero, in the case of variates such as survival time, and some artificially induced cut point, such as the grade-point average required for admission to a university, most densities are unlikely to be discontinuous at a single fixed point. However, many densities are likely to be mixtures.

The mixture decomposition or bump-hunting problem provides an excellent example of the different applications of three major classes of curve estimation methodology. Assume that the normal kernel $K(x) = N(x|0, h)$ has been chosen for use with the conventional kernel approach to curve estimation. The characteristic function of the normal $N(x|0, h)$ which has mean zero and scale parameter h, is proportionate to a normal with mean zero and scale parameter $1/h$. Consider the series curve estimator based on the orthogonal system $\{\exp(2\pi ikx)\}$ with multipliers b_k, $k = 0, \pm 1, \pm 2, \ldots$, which are approximately proportional to equally spaced evaluations of $N(x|0, 1/h)$; in other words, Fourier coefficients of the expansion of $N(x|0, h)$. For all applied purposes this estimator is identical to normal-Kernel curve estimator

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{j=1}^{n} \phi\left(\frac{x - X_j}{h}\right)$$

where ϕ represents the standard normal density $N(x|0, 1)$ (Anderson and de Figueiredo 1980; Tarter and Raman 1971).

If h is any positive value, then for a sample of size n, the second cumulant of the curve estimator $\hat{f}_h$, based on the normal kernel, will be $(n - 1)/n$ times the sample variance s^2 plus the quantity h^2. Like the penalized-likehood approach which emphasizes the smoothness of the estimator through a roughness controlling constant β, for the above kernel method, the inflation of the second cumulant by the constant h^2 *increases* the smoothness of the estimator $\hat{f}_h$.

However, suppose that one substitutes the value ih, where $i = (-1)^{1/2}$, for h in the truncated series-multiplier form of the estimator $\hat{f}_h$. One then obtains the procedure developed by Doetsch (1936) and Kronmal (1964), and then applied to biomedical data by Gregor (1969), Tarter and Silvers (1975) and Tarter (1979b). By means of the Doetsch–Kronmal (DK) method, under very general conditions one can estimate a hypothetical distribution, which, if one's original density is one of a large class of mixture distributions, will also be a mixture, but with the second cumulants of all components reduced by the value $h = \lambda$.

A major distinction can be made between the Good and Gaskins (1980) series and the above DK λ-multiplier estimator. While the former emphasizes the smoothness or penalizes the roughness of the true underlying density, the latter purposefully increases roughness as a way of examining the placement and verifying the existence of hidden distributional components. For example, the latter, in the case of the half-parameter normal mixture model where $\sigma > 1/2$, will accentuate or heighten bumps or bimodal structure.

The penalized likelihood and the DK λ-multiplier methods illustrate opposite trends of generalized or nonparametric application. For exploratory data analysis, the DK method is a valuable way of searching for clues concerning the nature of one's data, since it allows one to discern otherwise hidden bumps. On the other hand, the penalized likelihood method, because of the penalty placed on roughness, is of value for confirmatory applications. If a bump persists as one increases the roughness penalty setting β, the existence of the bump becomes increasingly believable, in other words, one obtains evidence for the claim that the bump actually exists.

Consider the usual kernel approach to density estimation which uses the choice of bandwidth h as a means of assuring goodness of fit of the estimator to the estimated density. Like the roughness penalty approach, the kernel method provides confirmatory support for the existence of bumps. However conversely, unlike the DK method, it tends to obscure difficult-to-discern mixture subcomponents.

The kernel method has been generalized by allowing h to be functionally related either to the data point X_j or the running variate x. For example, Breiman, Meisel, and Purcell (1977) consider h to be proportionate to $\alpha_k d_{j,k}$ where $d_{j,k}$ is the distance from the data point X_j to its kth nearest neighbor and α_k is a constant multiplicative factor. Like the choice of h of the conventional kernel estimator or the choice of roughness penalty β, the choice of α_k is made in terms of the estimator's goodness of fit to the estimated density. To date, it is only the user-selected constant λ of the DK method and not h, α_k or β which can be chosen to purposely increase the roughness of the fitted function, to accentuate bumps or hidden mixture components.

4. Nonparametric Inference and Curve Estimation

There are many applications of the new generalized forms of statistical representation besides mixture decomposition. However, as its name implies, the field of nonparametric inference merits special attention with regard to its connections with curve estimation. It will now be shown that the $\{\exp(2\pi ikx)\}$ form of representation can be applied to rank-based "nonparametric" inference. By employing one common form of expression, it is possible to create a wide spectrum of nonparametric inferential and curve estimation hybrids. The power of the particular hybrid corresponding to the two-sample rank test described below can match and even slightly surpass that of the conventional form of implementation. That the improvement in statistical efficiency is small is attributable to the already highly advanced state of nonparametric inferential methodology, rather than to any limitation to curve estimation's potential to contribute to statistical inference.

To explain how hybrid curve estimation-nonparametric inferential procedures can be devised, note that the steps used to derive current model-based tests are different from those used for rank-based inferential methods. Model-based inferential procedures are initially formulated in terms of a population's distributional model and a hypothesis concerning one or more parameters of the model. Only as a secondary consideration is the problem of estimation dealt with. On the other hand, rank-based procedures are initially formulated in terms of properties of a ranked sample (Hajek and Sidak 1967; Kendall 1962; and Lehmann and D'Abrera 1975).

From a computational point of view, ranking is the equivalent of the estimation of the population cumulative by means of the sample cumulative F^*. Specifically, if one has access to a computer algorithm by which the sample cumulative of a given sample, $\{X_j\}$, $j = 1, \ldots, n$, can be calculated, one can proceed to determine the ranks of one's sample elements by computing $nF^*(X_j)$ for each of the unranked observations within $\{X_j\}$. Therefore, when one bases a statistical investigation on a rank test, one has in effect proceeded from the assumption that the estimator $nF^*(X_j)$ is in some way superior to any alternative estimator.

In reality F^* is only one of many estimators of the population cumulative F available today, all of which can yield alternatives to F^*. For example, one such alternative can be constructed from the partial integral of the Fourier series density estimator $\hat{f}(x) = \Sigma_k b_k \hat{B}_k \Psi_k(x)$ where $\{b_k\}$ is an estimated multiplier sequence described in Section 2. (In Kronmal and Tarter (1968, sec. 3), the raw estimator special case of $\hat{f}(x)$ was shown to approach F^* as $m \rightarrow \infty$.) The availability of alternatives to F^* makes it possible to greatly generalize rank-based test calculation. One can as a first step formulate functionals of the population cumulative which correspond to particular rank tests. Then, as a second step, one can investigate alternative estimators of these functionals.

For example, consider that the usual rank correlation statistic r_s computed from a sample of size n (Dixon and Massey 1983, p. 402) is an estimator of the quantity

$$K_1 - K_2 E_{f(x,y)}[F_x(X)F_y(Y)]$$

where

$$K_1 = 1 - 2(n + 1)(2n + 1)/(n - 1),$$

$$K_2 = 12n/(n - 1),$$

f is the joint density of the variate (X, Y) while F_x and F_y are the cumulatives of the variates X and Y, respectively. Here

$$E_{f(x,y)}[F_x(X)F_y(Y)] = \{1 - \pi^{-2} \sum_u \sum_v [(B_{u,0} - 1) \times (B_{0,-v} - 1)B_{u,v}/uv]\}/4$$

for any variates X and Y whose support is on the unit interval and whose means equal one half, where the symbol $B_{u,v}$ represents the u,vth Fourier coefficient of f, and the indices of summation range over all nonzero integer values. If one uses an $\{\exp(2\pi ikx)\}$-series penalized likelihood or MISE procedure, one can estimate $E_{f(x,y)}[F_x(X)F_y(Y)]$ by substituting an appropriate series estimator coefficient for each $B_{u,v}$. Provided that the kernel bivariate density estimator satisfies the conditions for series and kernel identity described in Tarter and Raman (1971) one can also base the above hybrid statistic on any kernel bivariate density estimator and, in this way, devise hybrid kernel-based nonparametric tests.

As a second example of a hybrid test, let $\bar{X}_1$ and $\bar{X}_2$ be the sample means of two random samples of size n_1 and n_2 whose elements have support on the unit interval.

Define the statistic

$$T_m = n_1 n_2 \left\{ \bar{X}_1 - \bar{X}_2 + \sum_{k=-m}^{m} \frac{B_{-k}^{(1)} B_k^{(2)}}{2\pi i k} \right\} + \frac{n_1(n_1 + n_2 + 1)}{2}$$

where each $B_k^{(j)}$, $j = 1, 2$, is the kth sample Fourier coefficient calculated from the jth sample, $j = 1, 2$. As $m \to \infty$, T_m approaches the two sample rank sum test statistic T defined in Dixon and Massey (1983, p. 394).

In summary, by using the above procedures one can find hybrids between any of the curve estimators which can be represented in series form and either the two sample rank test, T, or rank correlation, r_s.

The following simulation study was performed to provide an example of a hybrid nonparametric inferential procedure's performance. In all trials which compared the rank sum statistic T with T_m, m was chosen by applying the series term stopping rule described by Tarter and Kronmal (1976) to the series estimator described in Section 2.

Two studies were conducted. Study 1 was designed to simply compare the numerical values of the two alternatives, T and T_m, when both test statistics were computed from the same sample. In Study 1, sixty trials were conducted. Each trial involved two samples which consisted of $n_1 = 300$ and $n_2 = 100$ random normal variates, respectively. In all trials the variance of the

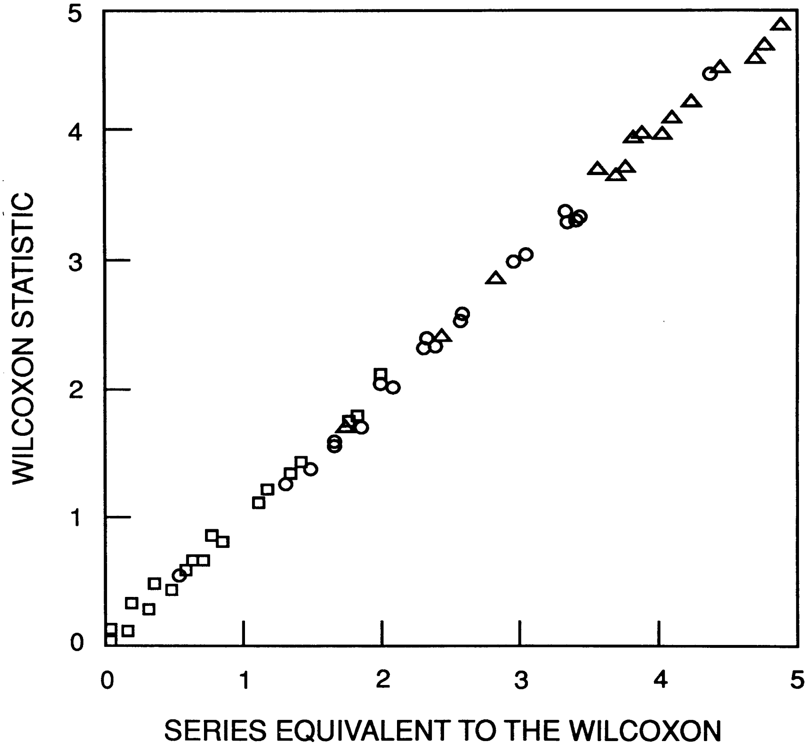

Fig. 1. Comparison of Series and Wilcoxon Test Statistics. □ population means from which the two samples were drawn are 0.0 and 0.0, respectively. ○ population means are 0.0, 0.2. △ population means are 0.0, 0.4. Each point represents one trial where the samples of 300 and 100 points are drawn from normal populations of unit variance and the indicated difference in population means.

normal distribution was chosen to equal 1. In Fig. 1, values of the test statistics which correspond to the mean pair (μ_1, μ_2) = (0.0, 0.0) are depicted by the symbol "□." Values of the test statistic which correspond to the mean pair (μ_1, μ_2) = (0.0, 0.2) are depicted by the symbol "○"; values corresponding to (μ_1, μ_2) = (0.0, 0.4) are depicted by "△." The x-coordinate of each point shown is the value of the series-equivalent of the rank sum based on T_m; the y-coordinate is the conventional rank sum statistic T.

Figure 2 summarizes the findings of Study 2 which compared the power characteristics of T and T_m. Here 800 trials were conducted, where for each trial independent $n_1 = 300$ and $n_2 = 100$ samples were generated. The near identity of the estimated power curves shown in Fig. 2 indicates that the power characteristics of tests based on the T_m and those based on T are comparable. For certain alternatives, the hybrid Wilcoxon statistic T_m may slightly improve the power of this test. This finding is in accord with studies which show that the Kaplan–Meier (1958) estimator, as well as the conventional life-table (Shryrock and Siegel 1973), may be substantially improved by using series estimators based on the metric $J(\hat{f}, f, w)$ (Tarter, Freeman, and Polissar 1990).

The brief study presented above illustrates only one potential use of inferential hybrid procedures. Where previously only a single estimator like F^* was used for almost all applications, there are today a wide

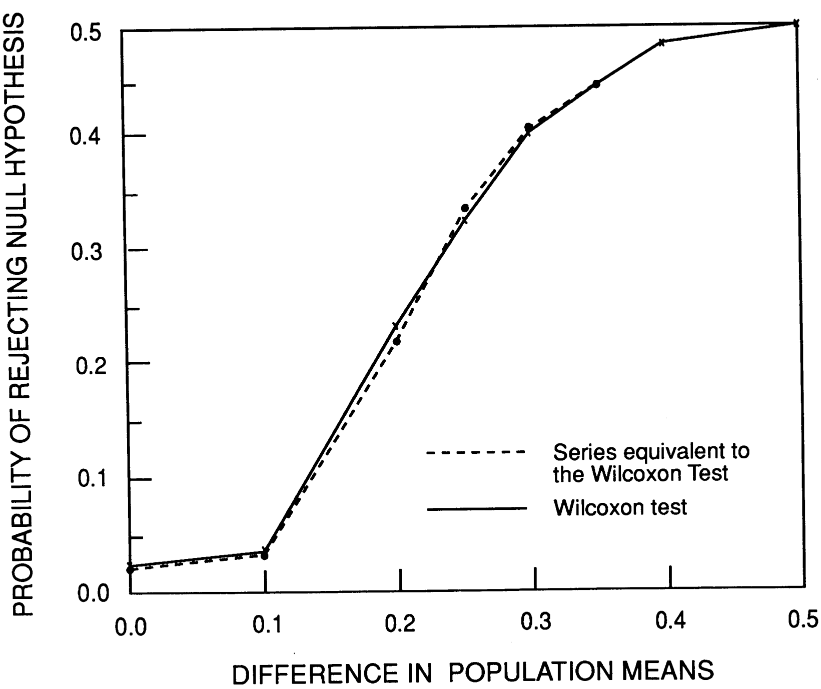

Fig. 2. Power Curves for the Wilcoxon and its series Equivalent. Each point marked on the power curves was derived from 100 trials where the samples were drawn from normal populations with the given difference in population means and unit variance. (The sample sizes were 300 and 100 where the mean of the population from which the first sample was drawn was always zero.)

variety of alternatives. Because most of the new *cdf* estimation methods do not require the ranking of data as a computational preliminary, the field of rank statistics may need to be reconsidered in the light of the computational convenience and the power characteristics of the new hybrid procedures.

5. References

Anderson, G.D. (1969). A Comparison of Methods of Estimating a Probability Density Function. Unpublished doctoral dissertation, University of Washington.

Anderson, G.L. and de Figueiredo, R.J.P. (1980). An Adaptive Orthogonal-Series Estimator for Probability Density Functions. Annals of Statistics, 11, 25–38.

Bowman, A.W. (1984). An Alternative Method of Cross-Validation for the Smoothing of Density Estimates. Biometrika, 71, 353–360.

Bowman, A.W. (1985). A Comparative Study of some Kernel-based Nonparametric Density Estimators. Journal of Statistical Computation and Simulation, 21, 313–327.

Breiman, L., Meisel, W., and Purcell, E. (1977). Variable Kernel Estimates of Multivariate Densities. Technometrics, 19, 135–144.

Brunk, H.D. (1978). Univariate Density Estimation by Orthogonal Series. Biometrika, 65, 521–528.

Carleson, L. (1966). On Convergence and Growth of Partial Sums of Fourier Series. Acta Math, 116, 135–137.

Craven, P. and Wahba, G. (1979). Smoothing Noisy Data with Spline Functions, Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. Numerical Mathematics, 31, 377–403.

Davis, K.B. (1977). Mean Integrated Square Error Properties of Density Estimates. Annals of Statistics, 5, 530–535.

Diggle, P.J. and Hall, P. (1986). The Selection of Terms in an Orthogonal Series Density Estimator. Journal of the American Statistical Association, 81, 230–233.

Dixon, W.J. and Massey, F.J. (1983). Introduction to Statistical Analysis, Fourth edition. New York: McGraw–Hill.

Doetsch, G. (1936). Zerlegung einer Funktion in Gauss'sche Fehlerkurven. Mathematische Zeitschrift, 41, 283–318.

Duin, R.P.W. (1976). On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions. IEEE Transactions on Computers, C-25, 1175–1179.

Fellner, W.H. and Tarter, M.E. (1971). Some New Results Concerning Density Estimates Based Upon Fourier Series. Proceedings of the Fifth Interface Symposium between Statistics and Computer Science, 54–64.

Good, I.J. and Gaskins, R.A. (1971). Nonparametric Roughness Penalties for Probability Densities. Biometrika, 58, 255–277.

Good, I.J. and Gaskins, R.A. (1980). Density Estimation and Bump-hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data. Journal of the American Statistical Association, 75, 42–73.

Grattan-Guinness, I. (1970). The Developments of the Foundations of Mathematical Analysis from Euler to Riemann. Cambridge, Mass: The MIT Press.

Gregor, J. (1969). An Algorithm for the Decomposition of a Distribution into Gaussian Components. Biometrics, 25, 79–93.

Habbema, J.D.F., Hermans, J., and van der Broek, K. (1974). A Stepwise Discrimination Program Using Density Estimation. In Bruckman, G. (ed.), Compstat 1974. Vienna: Physica Verlag, 100–110.

Hall, P. (1980). Estimating a Density on the

Positive Half Line by the Method of Orthogonal Series. Annals of the Institute for Statistics and Mathematics, 32(A), 351–362.

Hall, P. (1982). Comparison of Two Orthogonal Series Methods of Estimating a Density and its Derivatives on an Interval. Journal of Multivariate Analysis, 12, 432–449.

Hall, P. (1983). Large Sample Optimality of Least Squares Cross-Validation in Density Estimation. Annals of Statistics, 11, 1156–1174.

Hart, J.D. (1985). On the Choice of a Truncation Point in Fourier Series Density Estimation. Journal of Statistical Computation and Simulation, 21, 95–116.

Henna, J. (1985). On Estimating the Number of Constituents of a Finite Mixture of Continuous Distributions. Annals of the Institute for Statistics and Mathematics, 37(A), 235–240.

Hajek, J. and Sidak, Z. (1967). Theory of Rank Tests. New York: Academic Press.

Kaplan, E.L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53, 457–481.

Kevles, D.J. (1985). In the Name of Eugenics, Berkeley, CA: University of California Press.

Kendall, M.G. (1962). Rank Correlation Methods. New York: Hafner Publishing Company.

Kronmal, R.A. (1964). The Estimation of Probability Densities. Unpublished doctoral dissertation, Division of Biostatistics, University of California, Los Angeles.

Kronmal, R.A. and Tarter, M.E. (1968). The Estimation of Probability Densities and Cumulatives by Fourier Series Methods. Journal of the American Statistical Association, 63, 925–952.

Larsen, R. (1970). An Introduction to the Theory of Multipliers. New York: Springer-Verlag.

Lehman, E.L. and D'Abrera, H.J.M. (1975). Nonparametric Methods Based on Ranks. Oakland: Holden-Day.

Lexis, W.H.R.A. (1877). Zur Theorie der Massenerscheinigungen in der menschlichen Gesellschaft. Freiburg.

Müller, H.G. (1984). Smooth Optimum Kernel Estimators of Densities, Regression Curves and Modes. Annals of Statistics, 12, 766–774.

Neyman, J. (1949). Contribution to the Theory of the χ^2 Test. Proceedings of the Berkeley Symposium. Edited by J. Neyman, Berkeley, CA: University of California Press.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution. Philosophical Transactions of the Royal Society. 185A, 71–110.

Rosenlicht, M. (1975). Differential Extension Fields of Exponential Type. Pacific Journal of Mathematics, 57, 289–300.

Rudemo, M. (1982). Empirical Choice of Histograms and Kernel Density Estimators. Scandinavian Journal of Statistics, 9, 65–78.

Särndal, C-E. (1971). Studies in the History of Probability and Statistics. XXVII – The Hypothesis of Elementary Errors and the Scandinavian School in Statistical Theory. Biometrika, 58, 375–391.

Scott, D.W. and Factor, L.E. (1981). Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators. Journal of the American Statistical Association, 76, 9–15.

Scott, D.W., Tapia, R.A., and Thompson, J.R. (1977). Kernel Density Estimation Revisited. Nonlinear Analysis, 1, 339–372.

Shryrock, H.S., Siegel, J.S., Bayo, F., Davidson, M., Demeny, P., Glick, P.C., Grabill, W.H., Grove, R.D., Israel, R.A., Jaffe, A.J., Kindermann, C.R., Larmon, E.A., and Nam, C.B. (1973). The Methods and Materials of Demography. U.S. Department of Commerce, Washington D.C.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. London and New York: Chapman and Hall.

Sz.-Nagy, B. (1965). Introduction to Real Functions and Orthogonal Expansions. New York: University Texts in Mathematical Sciences.

Tapia, R.A. and Thompson, J.R. (1978). Nonparametric Probability Density Estimation. Baltimore, MD: Johns Hopkins University Press.

Tarter, M.E. (1979a). Trigonometric Maximum Likelihood Estimation and Applications to the Analysis of Incomplete Survival Information. Journal of the American Statistical Association, 74, 132–139.

Tarter, M.E. (1979b). Biocomputational Methodology: An Adjunct to Theory and Applications. Biometrics, 35, 9–24.

Tarter, M.E., Freeman, W.R., and Polissar, L. (1990). Modular Nonparametric Subsurvival Estimation. Journal of the American Statistical Association, 85, 29–37.

Tarter, M.E. and Kronmal, R.A. (1976). An Introduction to the Implementation and Theory of Nonparametric Density Estimation. American Statistician, 30, 105–112.

Tarter, M.E. and Raman, S. (1971). A Systematic Approach to Graphical Methods in Biometry. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, volume 4. Berkeley, CA: University of California Press.

Tarter, M.E. and Silvers, A. (1975). Implementation and Applications of Bivariate Gaussian Mixture Decomposition. Journal of the American Statistical Association, 70, 47–55.

Wahba, G. (1977). Optimal Smoothing of Density Estimates. In Van Ryzin, J. (ed.), Classification and Clustering, New York: Academic Press, 423–458.

Wahba, G. (1981). Data-Based Optimal Smoothing of Orthogonal Series Density Estimates. Annals of Statistics, 9, 146–156.

Watson, G.S. (1969). Density Estimation by Orthogonal Series. Annals of Mathematical Statistics, 40, 1496–1498.

Whittle, P. (1958). On the Smoothing of Probability Density Functions. Journal of the Royal Statistical Society, ser. B, 20, 334-343.

Received December 1989
Revised October 1990