Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data

I. J. GOOD and R. A. GASKINS*

The (maximum) penalized-likelihood method of probability density estimation and bump-hunting is improved and exemplified by applications to scattering and chondrite data. We show how the hyperparameter in the method can be satisfactorily estimated by using statistics of goodness of fit. A Fourier expansion is found to be usually more expeditious than a Hermite expansion but a compromise is useful. The best fit to the scattering data has 13 bumps, all of which are evaluated by the Bayesian interpretation of the method. Eight bumps are well supported. The result for the chondrite data suggests that it is trimodal and confirms that there are (at least) three kinds of chondrite.

KEY WORDS: Nonparametric density estimation; Penalized likelihood; High-energy scattering data; Chondrites; Clustering by bump-hunting; Bump evaluation.

1. INTRODUCTION

This article is about the estimation of probability densities and the location and probabilistic evaluation of bumps, especially those the estimated probabilities of which exceed one-half. Statisticians who are more concerned with exploratory data analysis than with statistical inference should also refer to Boneva, Kendall, and Stefanov (1971) and to Tukey (1977, Ch. 7).

A bump in a probability density curve without straight parts is the part lying between two points of inflection and that is concave when viewed from below. If convex we call this part a dip. Bumps and dips almost always alternate in univariate problems. In n-variate problems a bump (concave from below) or dimple (convex) is the part of the density n-dimensional "surface" (without flat parts) lying within a closed $(n-1)$-dimensional "contour" on which the n-dimensional curvature vanishes, but in this article we are mainly concerned with univariate problems. A bump, even if it is not a local maximum, indicates some feature of a random variable requiring an explanation. For example, bumps arise when unimodal distributions are mixed and each bump can be interpreted as a cluster.

Bump-hunting is not quite the same problem as density estimation although the two are closely related. The problem of whether a bump is present is one of significance testing more than of estimation. Orear and Cassel (1971) described the problem as "one of the major current activities of high-energy physicists" (p. 281). In high-energy physics the problem is to detect "real" bumps (and dips) in mass spectra in scattering experiments, that is, bumps in the population density curve when the evidence is in the form of a sampled histogram. (The bumps and dips give evidence, but not the only evidence, concerning the "partial-wave scattering amplitudes" (Omnés and Froissart 1963, p. 19) and sometimes concerning new elementary particles.) In the present article we shall extend the theory and practice of bump-hunting while illustrating the methods with two applications, one to scattering data and one to meteorite data.

A method that may be called the (maximum) penalized-likelihood (MPL) method for estimating probability densities and for bump-hunting was described earlier (Good and Gaskins 1972); this article will be referred to as G & G′ hereafter. This article was an invited elaboration of Good and Gaskins (1971). (See also Good 1971a,b.) The articles published in 1971 did not discuss histogram (grouped) data but only raw data (continuous data) in the form of N independent and identically distributed random real variables with a continuous density. In some high-energy experiments the expression *raw data* is misleading because the data are reported in the form of a histogram in which the cells are called bins by physicists.

In 1971 Malcolm Mac Gregor, who had obtained it from Protopopescu, kindly supplied a mass-spectrum histogram that was gathered at the Lawrence Radiation Laboratory in Berkeley. The histogram consists of

* I.J. Good is a University Distinguished Professor, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. R.A. Gaskins is an Associate Professor, Department of Mathematics and Computer Science, P.O. Box 311, Hampden-Sydney College, Hampden-Sydney, VA 23943. The research was supported by the U.S. Department of Health, Education and Welfare, National Institutes of Health Grant R01–GM18870. The authors are grateful to the referees and to an associate editor for their painstaking work and useful suggestions, and to V. Teplitz for information concerning the interpretation of bumps in scattering data. This article was presented at a special *JASA* Invited Paper Session at the ASA Annual Meeting on Tuesday, August 14, 1979, in Washington, D.C. The paper was selected from among those under review by the Applications Section during 1978, and galley proofs were made available several months before the meetings. Discussion that followed the presentation of the paper by Good included comments by Emanuel Parzen, William Eddy, David Scott, Michael Tarter, and Bernard Silverman. These and two other Comments are included here, followed by a Rejoinder.

© Journal of the American Statistical Association
March 1980, Volume 75, Number 369
Applications Section

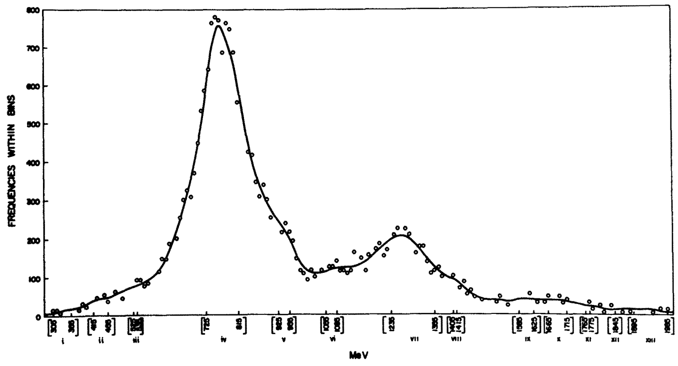
1. *The* LRL *Data of* $N = 25{,}752$ *"Events," and the Penalized Likelihood Fit With* $\beta^* = .225$*. Each Bin Is of Width 10* MeV

			(i)										
(a)	285	295	[305	315	325	335	345	355]	365	375	385	395	405
(b)	5	11	17	21	15	17	23	25	30	22	36	29	33
(c)	5	8	12	15	18	20	23	25	27	29	32	35	38
	(ii)											(iii)	
	[415	425	435	445	455]	465	475	485	495	505	515	[525	535]
	43	54	55	59	44	58	66	59	55	67	75	82	98
	42	46	50	53	55	57	60	62	65	69	74	79	84
	545	555	565	575	585	595	605	615	625	635	645	655	665
	94	85	92	102	113	122	153	155	193	197	207	258	305
	88	93	99	106	116	128	143	160	179	200	225	253	285
						(iv)							
	675	685	695	705	715	[725	735	745	755	765	775	785	795
	332	318	378	457	540	592	646	773	787	783	695	774	759
	321	364	413	471	533	598	658	709	744	761	759	742	709
													(v)
	805	815]	825	835	845	855	865	875	885	895	905	915	[925
	692	559	557	499	431	421	353	315	343	306	262	265	254
	664	611	557	504	455	412	376	346	321	300	282	266	251
													(vi)
	935	945	955]	965	975	985	995	1,005	1,015	1,025	1,035	1,045	[1,055
	225	246	225	196	150	118	114	99	121	106	112	122	120
	235	218	199	178	158	140	126	118	113	112	113	115	118
	1,065	1,075	1,085]	1,095	1,105	1,115	1,125	1,135	1,145	1,155	1,165	1,175	1,185
	126	126	141	122	122	115	119	166	135	154	120	162	156
	121	124	125	126	127	129	132	136	140	144	148	154	161
					(vii)								
	1,195	1,205	1,215	1,225	[1,235	1,245	1,255	1,265	1,275	1,285	1,295	1,305	1,315
	175	193	162	178	201	214	230	216	229	214	197	170	181
	169	176	184	193	202	210	215	217	214	208	199	188	175
									(viii)				
	1,325]	1,335	1,345	1,355	1,365	1,375	1,385	1,395	[1,405	1,415]	1,425	1,435	1,445
	183	144	114	120	132	109	108	97	102	89	71	92	58
	163	150	138	128	120	113	107	100	94	88	81	74	68
	1,455	1,465	1,475	1,485	1,495	1,505	1,515	1,525	1,535	1,545	1,555	1,565	1,575
	65	55	53	40	42	46	47	37	49	38	29	34	42
	62	56	52	48	45	43	42	41	40	39	38	39	40
	(ix)								(x)				
	[1,585	1,595	1,605	1,615	1,625]	1,635	1,645	1,655	[1,665	1,675	1,685	1,695	1,705
	45	42	40	59	42	35	41	35	48	41	47	49	37
	41	42	43	44	43	43	43	43	43	43	43	42	41
						(xi)							
	1,715]	1,725	1,735	1,745	1,755	[1,765	1,775]	1,785	1,795	1,805	1,815	1,825	1,835
	40	33	33	37	29	26	38	22	27	27	13	18	25
	39	37	35	33	32	30	28	27	25	23	22	21	21
	(xii)					(xiii)							
	[1,845]	1,855	1,865	1,875	1,885	[1,895	1,905	1,915	1,925	1,935	1,945	1,955	1,965
	24	21	16	24	14	23	21	17	17	21	10	14	18
	21	20	20	20	20	20	19	19	18	18	17	16	15
	1,975	1,985]	1,995										
	16	21	6										
	13	9	6										

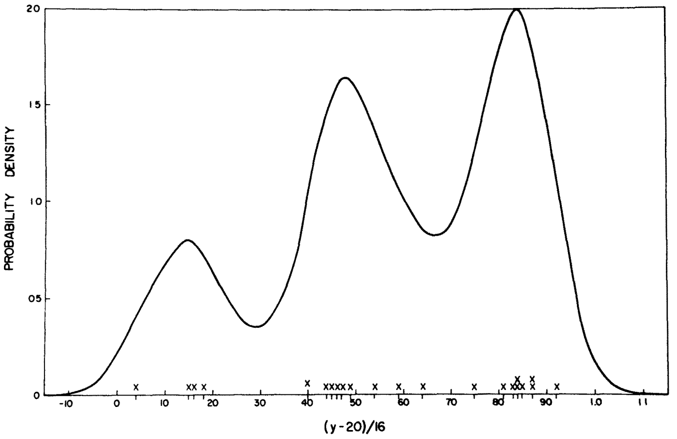
NOTE: No observations were made outside the 172 bins shown. Row (a) gives the centers of the bins in MeV; row (b) the observed frequencies; and row (c) the fitted frequencies to the nearest integer. The bumps (i) to (xiii) are indicated by bracketed intervals.

$N = 25{,}752$ events from a scattering reaction, but this account should be intelligible to a statistician who knows no physics. The data had been used as part of the basis of an article by Alston-Garnjost et al. (1971), but the complete histogram was not published there. (We give the histogram in Table 1; also see Figure A.) We call these data the LRL data. The midpoints of the bins are labeled in MeV and each bin is of width 10 MeV. Mac Gregor was interested in whether there was appreciable evidence for a bump or dip in the vicinity of 655 MeV because he has a theory that, among many other things, predicts either a bump or a dip in that vicinity, corresponding to "$\frac{2}{3}$ of a nucleon," of width between about 7 and 15 MeV, and of unknown height. By drawing a freehand graph we clearly see that, if the evidence is there, it will favor a bump rather than a dip. Bumps are generally more interesting than dips, and in the following paragraphs we shall consider mainly the evidence in favor of bumps but we shall also evaluate one dip of importance in the LRL data. There seems to be no gen-

A. LRL Data, Fitted Density of f(x) if β = .225, and 13 Bumps in f(x). (The observed bin frequencies are represented by small circles, but, to avoid cluttering the diagram, some of the circles have been omitted when they would be indistinguishable by eye from the fitted curve. Corresponding to each bump is a pair of brackets that lie close to and between the corresponding pair of points of inflection. Each bracket is within 10 MeV of a point of inflection.)*



B. Best Fit to the Chondrite Data (β = .030). (The scale of the x axis is that used by Leonard (1978) and is (y − 20)/16, where y is the percentage of silica given in Table 2. Thus, −1.25 < x < 5 by definition. The 22 observations are marked by crosses above the x axis. Also see Table 5.)*



erally accepted way to subtract the background corresponding to the removal of trend in a time series, and our location of bumps will be done by the penalized-likelihood method, while the evaluation of the bumps will use the Bayesian interpretation of the method.

As we shall see, the penalized-likelihood method, which is nonparametric, did not support the presence of a real bump near 655 MeV, but the method also does not give appreciable evidence against Mac Gregor's theory.

Our best smoothing of the LRL data had 13 bumps, all of which we evaluated statistically. Our method implies that any bump appearing in the best smoothing, and not predicted in advance, is odds on, that is, has estimated probability exceeding one-half of being present in the population density curve, assuming the value of β (as defined later) that is associated with the best smoothing. (A more hierarchical Bayesian method would allow for a distribution of β.) The method can also estimate a Bayes factor in favor of any fully specified bump that is suggested for scientific reasons. Note that two adjacent bumps might be better explained by a dip between them, at least for scattering data.

In addition to the LRL data we have reanalyzed the data for the distribution of silica in 22 chondrite meteors reported by Ahrens (1965) and Burch and Parsons (1976, p. 290). These data have been analyzed by T. Leonard (1978) using a different method of density estimation, and he suggested that we apply our method to these data for a comparison of the results.

These chondrite data differ from the LRL data in two important respects: The number of observations is small (22 instead of 25,752), and the observations are continuous, that is, sampled from a continuous distribution without grouping (see Table 2 and Figure B). (We first considered the chondrite data both ways: as grouped with each observation located at the center of a bin of width $h = .01$, and as continuous with repeated observations and $h = 0$. Because the results were almost identical and the grouping was artificial, we chose the continuous approach.) Our best fit for the chondrite data was trimodal, which is not surprising because there are several kinds of chondrite.

Continuous data are often conceived as sampled on the whole real line but are usually truncated because observing all values may be impracticable. (E.g., when scintillations, due to electrons impinging on a screen, are observed, the screen is of finite size.) We have written our computer program to cope with nontruncated continuous data, with truncated continuous data, and with grouped data (which are usually truncated).

The aims of the present article are mainly

1. To describe how we use tests of goodness of fit, specifically X^2 and D (Kolmogorov-Smirnov), in conjunction, to fix the hyperparameter β of G & G′, namely, the coefficient that determines the magnitude of the roughness penalty (see (2.1)). For the LRL data a new sign test was used as a confirmation. Our method could also

2. Percentages y of Silica in 22 Chondrites (Ahrens 1965)

y	20.77	22.56	22.71	22.99	26.39	27.08	27.32	27.33	22.57	27.81	28.69
Scaled	.04	.15	.16	.18	.40	.44	.45	.46	.47	.49	.54
y	29.36	30.25	31.89	32.88	33.23	33.28	33.40	33.52	33.83	33.95	34.82
Scaled	.59	.64	.75	.81	.83	.84	.84	.85	.87	.87	.92

NOTE: Row (ii) is the scaling used by Leonard (1978) and, to be consistent with him, we have used it in our analysis. It denotes (y − 20)/16 with rounding. Here $x^* = .57227$, $s = .27226$. Also see Figure B.

be applied to other density estimation procedures. The method, apart from the sign test, was first given publicity in the General Methodology lecture by Good at the August 1974 meetings of the American Statistical Association in St. Louis, Missouri.
2. To exemplify the penalized-likelihood method by using real data.
3. To show the relevance to one-dimensional cluster analysis.
4. To locate potential bumps in the LRL data and to evaluate them.
5. To evaluate Mac Gregor's bump.
6. To discuss methods of avoiding only local maxima in the maximization.
7. To make the methods available for other bump-hunters and density estimators.
8. To report a large decrease in running time achieved by the use of Fourier series in the place of Hermite series. Thus, two-dimensional problems should now be tractable.
9. To point out an advantage of using both Fourier and Hermite series in conjunction (for one-dimensional problems). In any case the costs are negligible in comparison with those of collecting the scattering data.
10. To describe how to choose the number of terms in these series.

See Appendix J regarding the computer programs, but the appendices can be omitted on a first reading.

Perhaps it will eventually be more convenient if our methods for bump-hunting in univariate problems are carried out by human and machine synergy with the help of a graphic data tablet. With this device, a human could vary the putative density curve by eye, the computer returning the score ω and R_0 (as defined later). A graphic data tablet was not available to us. In writing this article we have partially integrated the two applications into the general discussion for the sake of concreteness and, we hope, greater interest.

(A longer version of this article is available from the authors.)

2. STATISTICAL BACKGROUND

The MPL method for estimating either a density function f for continuous data or the discrete probabilities corresponding to grouped (histogram) data involves the maximization of a score $\omega = \omega(f)$. The score is defined as the log-likelihood minus a (roughness) penalty, $\Phi = \Phi(f)$, where the penalty depends on only the putative density function f (or on the putative probabilities in the case of categorized data) apart from a proportionality parameter β. We shall explain how β depends on the observations, but just for a moment let us imagine that it does not depend on them at all once f is assumed. Then, although the method can be regarded as non-Bayesian, it also has a simple Bayesian interpretation. For we can regard $e^{-\Phi}$ as proportional to a prior density ("improper") in function space (the space of functions f) when estimating a density function, or in the space of the discrete probabilities when estimating the physical probabilities for categorized data. Then the MPL method maximizes the posterior density in the function space. (This density should not be confused with the density function f itself.) For the evaluation of small bumps, as described later, we have to do more than just maximize the penalized likelihood, and for this purpose we use the Bayesian interpretation.

2.1 Relation to Other Problems

For categorized data (contingency tables), the MPL method, although not named, was to some extent proposed by Good (1963, p. 931; 1965, p. 76), but with the penalty proportional to negative entropy, when the values taken by each facet (or "dimension") have no natural order. (Also see Pelz 1977.) When the values of each facet have a natural order, smoothness is relevant and the derivatives used in the continuous problem, as described in the following paragraphs, can be replaced by finite differences in one or more dimensions. We understand from Thompson that this or a similar method has been successfully tried by David Scott at Rice University.[1]

The MPL method could also be applied to the estimation of parameterized density functions, for example, to multivariate normal densities and to their mixtures. In this manner one would induce a prior distribution of the parameters. We hope to experiment with this method later.

2.2 Specific Details

For continuous densities we use a penalty depending on the derivatives of the putative density. Because smoothness is related intuitively to curvature we emphasize the second derivative. Specifically we write $\gamma(x) = \surd\{f(x)\}$ (positive or negative), we call $\int[\gamma''(x)]^2dx$ the roughness of f, where the primes denote differentiation, and we choose the roughness penalty as β times the roughness for some positive β to be determined. Thus, the MPL method, in the specific form used here and in G & G′, uses the maximization of the score ω where

$$\omega = \omega(f) = L - \Phi(f) = L - \beta \int [\gamma''(x)]^2 dx\,, \tag{2.1}$$

where L denotes the log-likelihood. (Also see Appendix A.) For a histogram of J bins we have

$$L = \sum_{j=1}^{J} n_j \log_e \int_{B_j} f(x)dx\ , \tag{2.2}$$

n_j being the sample frequency in the jth bin B_j. Because the MPL method has a Bayesian as well as a non-Bayesian interpretation, β is called the hyperparameter,

[1] James R. Thompson, Personal communication, January 1976.

being a parameter in a prior. But β can be called a procedure parameter if one wishes to emphasize the non-Bayesian interpretation.

We have found that estimating β from the data is necessary. It would be possible to assign to β a hyperprior distribution (or Type III in the hierarchical terminology of Good 1952, 1965, 1967; Good and Crook 1974), but we have preferred to use other methods for estimating β.

When maximizing ω there is a constraint $\int [\gamma(x)]^2 dx = 1$, and in the maximization we use Lagrange's method of undetermined multipliers, the multiplier being denoted by λ (see Appendix B). If we can assume a value for the roughness, then the MPL method can be interpreted as the maximization of the likelihood with two constraints where β is then also a Lagrange undetermined multiplier.

The method used for maximizing the score ω is described in G & G′. (Also see Appendix C where an objection is answered.) The method makes use of the expansion

$$\gamma(x) = \sum_{m=0}^{r-1} \gamma_m \phi_m(x) \ , \tag{2.3}$$

where $\phi_0(x)$, $\phi_1(x)$, ... are normal orthogonal functions and where r is finite in practice but infinite in theory. Before 1976 we used the Hermite normal orthogonal system. We have now tried substituting the Fourier for the Hermite system, as well as a mixed system, with good results. In the Fourier system, if $f(x)$ is treated as if it were periodic with period $2\pi b_0$, we take

$$\begin{aligned} \phi_0(x) &= (2\pi b_0)^{-\frac{1}{2}} \ , \\ \phi_{2n-1}(x) &= (\pi b_0)^{-\frac{1}{2}} \sin\,(nx/b_0) \ , \\ \phi_{2n}(x) &= (\pi b_0)^{-\frac{1}{2}} \cos\,(nx/b_0)(n > 0) \ . \end{aligned} \tag{2.4}$$

When we wish to distinguish between the Fourier and Hermite systems we write $\phi_m{}^F$ and $\phi_m{}^H$, with similar notations $\gamma_m{}^F$ and $\gamma_m{}^H$ for the coefficients in (2.3). We denote the number of terms by r for both systems or sometimes by r^F and r^H, and we shall soon discuss the choice of r.

Continuous data are sometimes defined on a circle (Mardia 1972), and then b_0 will have a physical meaning. Otherwise the choice of b_0 involves a judgment that the density is negligible outside the interval selected. (The judgment could be changed after f is estimated, and the calculations then repeated, but we believe doing so would not often be necessary.) A similar choice must be made for grouped data by adding artificial bins to the left and right of the genuine bins. (This must be done even when the Hermite system is to be used, because these artificial bins will contain expected frequencies, based on the putative probability density function. These frequencies are relevant for the calculation of the X^2 test for goodness of fit.)

When the Fourier system is to be used, let the interval outside of which the density function $\gamma^2(x)$ is assumed to vanish, or to be negligible, be denoted for the moment by (a_1, a_2). We want this interval to be wider than the interval containing all the observations. (Because there were 28 naturally occurring empty bins in the left tail of the LRL data, we thought we may as well append an equal number of empty bins in the right tail. Doing so made a total of 228 bins with $a_1 = 0$ MeV and $a_2 = 2{,}280$ MeV.)

We found that moving the origin to the mean $\bar{x}$ of the observations was important for both the Fourier system and the Hermite system. In addition, we scaled the x axis so as to make the sample variance $s_z{}^2$, of the new variable z, equal to $\frac{1}{2}$, because $\phi_0{}^H(\cdot)$ is proportional to a Gaussian curve with this variance. From now on our notations for the coefficients γ_m, $\gamma_m{}^F$, and $\gamma_m{}^H$ refer to the scaled or standardized data. One reason for the scaling is implicit in Appendix E. The standardized variable z, for the Fourier system, is now in an interval $(-\pi b, \pi b)$ in which $\pi b s\sqrt{2} = \max(\bar{x} - a_1,\ a_2 - \bar{x})$, s being the standard deviation of the observations. (For the LRL data, $\bar{x} = 940$ MeV, and $\pi b = 3.09$.)

For the effect of scaling on β see Appendix F and note that β^* denotes the value of the hyperparameter in the scaled coordinate system.

The Fourier system required many fewer terms than the Hermite system. One reason why the latter requires many terms is that there is a strong correlation between $\gamma_m{}^H$ and $\gamma_{m+4}{}^H$. Indeed, as we knew in 1971, the signs of the coefficients $\gamma_m{}^H(m = 0, 1, 2, 3, \ldots)$ contain very long runs with the pattern

$$+ + - - + + - - + + - - + + - - + + - - + + - - + + - - .$$

We took five years to find a reason for this pattern (see Appendix G).

Apart from requiring many fewer terms than the Hermite system, the Fourier system has the advantage of involving simpler equations (see Appendix B), and the approximate number of terms can be determined a priori if we assume a width for the thinnest bump of interest, which we take as a bin width for histogram data. For the determination of r see Appendix D. For the LRL data, although we needed about 3,000 Hermite terms, only 271 Fourier terms are sufficient. Thus, the Fourier method is much faster (even without using a fast Fourier transform) and requires much less storage when run on a computer.

The Fourier method, however, has the disadvantage that bad initial choices for the $\gamma_m{}^F$ can cause the iterative process to converge to a mere local maximum. In every instance these merely local maxima have led to absurd density curves (e.g., highly rejectable by an X^2 test) so that there seems to be little danger of being deceived.

In the Hermite system, we have found that initializing at $\gamma_0 = 1$, and all other $\gamma_m = 0$, apparently leads to the global maximum in all (some dozen) examples tried. In this sense the Hermite method is more robust than the Fourier method, a matter discussed in more detail in Appendix E. The convergence criterion for either method is given in Appendix H. Our original main reason for choosing the Hermite system rather than the Fourier system was that Hermite functions tend to zero at infinity, which might be the basic reason why the Hermite system has some advantages.

By switching back and forth between the Hermite and Fourier systems, we combined into a single computer program the robustness of the Hermite system with the speed of the Fourier system. Doing so was made possible by the numerical procedure described in Appendix I. The reader should note that a small value of r^H can be used to get the density curve in the right ball park.

It would be possible to start the convergence by drawing a freehand curve to fit the data and computing its Fourier coefficients by numerical methods. These coefficients could then be used directly to initialize the Fourier system or could be converted to Hermite coefficients to initialize the Hermite system. But we prefer our more automatic methods.

In Good and Gaskins (1971, Appendix D), we chose a value for β based on whether various density curves looked smooth and also on a weak intuitive argument (see also G & G′, Appendix D). But in G & G′ we reported the results of some simulation experiments that forced us to recognize the need to choose β (or β^*) in accordance with the data. We have used two methods for choosing β (or β^*), both of which involve some trial and error and both of which we believe to be reliable. (Starting with $\beta^* = \frac{1}{2}$ is adequate.) One of these methods, which we call the method of synthetic populations, is described in the longer version of this article (mentioned in Section 1). The other method, which is less time consuming, is the one we have used since mid-1974 and that depends on tests of goodness of fit such as X^2.

2.3 Method for Choosing β (or β^*), Including a Sign Test

Apparently all procedures for estimating probability densities involve procedure parameters analogous to the hyperparameter β, and our methods for choosing β apply to all such procedures. A cross-validation method is given by Wahba (1976) in a technical report that requires the reader to understand much more advanced mathematical concepts than we need in the present work, but cross-validation could also be applied for estimating our hyperparameter.

In a nutshell, for each putative value of β we can increase r until the criterion of Appendix D is satisfied, taking into account that when β is made larger there is no need to increase r (for given data of course). We then compute an X^2 value. Having done this for a variety of values of β, we select one value of β on the basis of the values of X^2 and D (the one-sample Kolmogorov-Smirnov statistic). We now give more details.

Let S denote a goodness-of-fit statistic (taking small values for close fits) and, given that f is the true density function, let the probability that S is less than the observed value be P_S.

If β is taken too large, the density function f will be oversmooth and will in general not "resemble" the observations enough so that P_S will be close to 1. If β is too small, f will usually be too rough and will resemble the observations too much so that P_S will be close to 0. The fit will then be too good to be true. We consider that the best value for P_S is $\frac{1}{2}$, because this value of P_S treats smoothness and roughness symmetrically and because, for this value of P_S, most statisticians would have the least tendency to reject the "hypothesis" f. ($P_S = \frac{1}{2}$ means that both tail-area probabilities of S are $\frac{1}{2}$.)

If several dependent statistics S_1, S_2, ... are used, the corresponding tail-area probabilities can be combined by the harmonic-mean rule of thumb (Good 1958), but our problem is special in that left and right tails correspond to conflicting phenomena, roughness and smoothness. We adopt the following procedure, which is symmetrical with respect to the two tails and which we regard as intuitively appealing.

Suppose we have m tail-area probabilities less than $\frac{1}{2}$, say $P_1, P_2, \ldots, P_m$ and n that are at least $\frac{1}{2}$, say $Q_1, Q_2, \ldots, Q_n$. We compute the harmonic mean h_P of the P's and k_Q of the $(1 - Q)$'s. Convert these to odds and take the weighted geometric mean

$$O = (h_P/(1 - h_P))^{m/(m+n)}((1 - k_Q)/k_Q)^{n/(m+n)}$$

and convert O to a resultant probability

$$R = R(P_1, \ldots, P_m, Q_1, \ldots, Q_n) = O/(1 + O) \ .$$

For example, if $n = 0$, then R is the harmonic mean of the P's.

We applied this method to several X^2 values, as described in the following paragraphs, to obtain a resultant tail R_1. We then formed $R(R_1, R_2)$ where R_2 is the tail-area probability of a Kolomogorov-Smirnov statistic. Finally we formed $2|R(R_1, R_2) - \frac{1}{2}| = R_0$. Of course $0 \leq R_0 \leq 1$. The smaller R_0 is, the better we regard the estimate β^*. But, for example, a value of .001 for R_0 is not much better than .05.

We chose our definition of X^2 to cover both continuous and grouped data. (For more details see the longer version of this article.)

When X^2 is used for testing goodness of fit, the value of X^2 can be unluckily affected because a few observations happen to lie just on one side of a class boundary (Good 1978a). The difficulty occurs for the chondrite data, having only 22 observations. One proposal to meet this difficulty was suggested in the reference: to compute X^2 using various numbers of equiprobable class intervals, relatively prime in pairs, and then to combine the tail-area probabilities by the harmonic-mean rule of thumb. We applied this method for the chondrite data, and the results are reported in Section 4. For the LRL data a single X^2 value is adequate.

For histogram data with many bins, there is the useful possibility of examining the signs of the differences expected minus observed in each bin (Good 1978c). If the smoothing corresponds exactly to the underlying population, this sequence of signs will be almost a coin-tossing or Bernoulli sequence with parameter $\frac{1}{2}$. This test has the advantage that it is capable of detecting local

3. *Statistical Measures of Fit for Selected Values of β^* for the* LRL *Data*

β^*	$X_{[171]}^2$	P_{X^2}	D[a]	P_D	R_0	Iterations
.80	255.2	.9997	.007726	.908	.999	22
.40[b]	215.0	.987	.005254	.474	.784	10
.30[b]	202.2	.947	.004470	.319	.486	9
.275[b]	198.5	.926	.004255	.262	.356	9
.25	194.8	.898	.004035	.206	.204	23
.225	190.8	.857	.003804	.151	.143	23
.20	186.6	.804	.003552	.100	.194	22
.175[b]	182.1	.734	.003317	.061	.405	7
.15[b]	177.2	.643	.003059	.031	.357	9
.10	165.2	.389	.002508	.003	.988	23
.05	146.9	.091	.001876	.000	1.000	24

[a] In accordance with Good (1978d), $D = .00541 \pm .00162$. We mention this to "help the reader's eye," but of course P_D was calculated properly.

[b] Failed to converge to the global maximum when initialized by the standard normal density; but convergence to the global maximum was achieved when γ's from the run with $\beta^* = .20$ were used to initialize the iterative process.

NOTE: We selected the value $\beta^* = .225$ ($\beta = .0080$) as being the best. Convergence was achieved by the Fourier method with $r^F = 271$. For the notation R_0 see Section 2. The value of D was attained at 810 MeV for all values of β^* in this table.

anomalies (without accurately locating them), and so it differs greatly from the X^2 test. There are various statistics for testing randomness of coin-tossing sequences (Gončarov 1943, 1962; Good 1953). We did not build any of these tests into our program, but we examined the sign. sequence for the LRL data after selecting β (or β^*) by means of X^2 and D. The results of examining the sign sequence for the LRL data are given in Section 3. We happened to use only the probability of no run of length 6 or more, but in other cases one would be interested in the distribution of runs of a given length (see Good 1979).

In a sequence of J trials, the probability that there will be no run of length ν or more, of either pluses or minuses, if we do not circularize is (Good 1978c, based on Uspensky 1937)

$$\frac{2-\xi}{(\nu+1-\nu\xi)\xi^{J+1}} + \theta\nu 2^{-J+3} \quad (|\theta| < 1) \ ,$$

where

$$\xi = 1 + 2^{-\nu} + \sum_{\ell=2}^{\infty} \frac{(\ell\nu+2)(\ell\nu+3)\ldots(\ell\nu+\ell)}{\ell!} 2^{-\nu\ell} \ .$$

A more accurate result, allowing for the sign sequence's not being precisely Bernoulli, is given by Good (1978c).

The runs-of-signs test is different from the X^2 test because the latter pays no attention to the ordering of the bins while the former attends to little else. Given the null hypothesis that our density function is the true one, it is therefore intuitively obvious that the probability of no run of length ν or more is nearly independent of X^2 provided that the number of bins is large (say 30 or more) and provided that we are not in the extreme tail of the distribution. So one might suggest that we could combine the X^2 test and the runs-of-signs test by Fisher's method for independent tests (Fisher 1938, p. 104). But these two tests are greatly correlated given the nonnull hypothesis, so the product of the tail areas fails to be a sensible criterion although Fisher's technique correctly approximates its distribution when the null hypothesis is true. The rough Bayesian justification for the use of products of tail-area probabilities (Good 1958, p. 804) depends on their independence given both the null and the nonnull hypothesis.

For data in two dimensions or more, if there were enough bins we could make use of the array of signs (see Good 1957; Krishna Iyer 1950).

The remaining work on the LRL data is described in Section 3, the main aim being the location and evaluation of all odds-on bumps. The chondrite data are further discussed in Section 4. We hope the work in these two sections will be of interest to statisticians who know nothing about scattering and meteorites.

3. RESULTS FOR THE SCATTERING (LRL) DATA

The LRL data are shown in Table 1 and in Figure A. There are positive frequencies in 172 bins and a somewhat sharp cutoff at both ends. There is a physical reason why the bins below 280 MeV necessarily have zero values (because the mass of a pion is 140 MeV, so that a two-pion spectrum has a threshold energy of 280 MeV). The sharp cutoff at the upper end is apparently because results greater than 2,000 MeV were not regarded as worth reporting. For simplicity we regard all empty cells (to the left and right of the 172 bins with positive frequencies) as unobserved, and we assume further that the bins are all nearly enough of exactly equal width (10 MeV). Physical experiments provide information in addition to the energies, but we have ignored this additional information in our analysis.

Probability density estimates were made, β^* having values ranging from .05 to .80, as shown in Table 3. For each value of β^*, convergence was attained in various numbers of iterations by using the Fourier method, with $r = 271$ (see Appendix D), usually starting the iterative process with the standard normal density. The result for $\beta^* = .20$ was confirmed by the Hermite method. In this case, after convergence was attained, $\gamma(x)$ was of constant sign over the 172 bins and over 5 bins to the left and right of these, and the maximum value of $f(x)$ where $\gamma(x) < 0$ was less than one-twenty-thousandth of the value of $f(x)$ at its mode.

To confirm that $r^F = 271$ is large enough, we repeated the run by using $r^F = 601$ and with the same value of β. The output was almost exactly the same; in fact the largest difference in expected frequency in any bin was .5 (near the mode) and .1 (near the tail).

On the basis of Table 3, which depends on X^2 and D (the Kolmogorov-Smirnov statistic), or more precisely on their summary statistic called R_0, we selected the hyperparameter $\beta^* = .225$ as the best. The β^* could be decreased to .175 or increased to .30 with little loss. For $\beta^* = .225$, .300, and .400, we applied in addition a sign test that is implicit in Section 2.

The sequence of signs of observed minus expected for $\beta^* = .225$ can be inferred from Table 1 and was

```
----++----+ -++-----+-- +++-----+++
     ++-+-++--- +++-++----+ ---+-++-++
--++-+----+ +++-++----- -++++-+-+-
     +--+++--+- -++--+++-+ -+--+-+-+-
++--+-+++- -++-++++-+ --+-++-++-
     +--++---+- +--++-++-- -+
```

In this sign sequence (circularized or not) the numbers of true runs of lengths 1, 2, 3, ... are, respectively, 46, 28, 12, 6, and 2, while the expected numbers, in a circularized head-and-tails Bernoulli sequence, are 43, 21.5, 10.8, 5.4, and 2.7. Up to runs of length 5, this is a good fit, but the probability of no run of length 6 or more (without circularization) is .053 (see Section 2). There is thus some indication that we should smooth a little more, but if we take $\beta^* = .30$ the effect is to change only two of the 172 signs, and the numbers of runs become 43, 25, 12, 7, and 3, respectively. On taking $\beta^* = .40$ only one more sign changes and the frequency count of runs is the same as for $\beta^* = .30$. Because there are still no runs of length 6 or more and because R_0 increases rapidly when β^* increases beyond .3 or .4, we decided to accept the value $\beta^* = .225$.

With $\beta^* = .225$ there are 13 bumps in the putative density function within the observed range of bins. The bumps are bracketed in Table 1 and are labeled (i) to (xiii) and are also indicated in Figure A. Each bracket in Table 1 is located within 10 MeV of a point of inflection and corresponds to a change in sign in the sequence of second differences of the expected contents of the 172 bins. (The expectations were calculated much more accurately than the rounded values shown in Table 1 and were based on the putative smoothed density function.) Mac Gregor's bump (at 655 MeV) does not show up (and by 1974 he said that he no longer believed that it could be identified from our LRL data by statistical methods alone). (See also some discussion near the end of the present section.) Bump (iv) is the main peak of the density curve and bump (vii) is also clearly real.

When $\beta^* = .3$, the same 13 bumps appear, except that one point of inflection moves by only 10 MeV. For $\beta^* = .4$, the 3 small bumps (iii), (viii), and (xii) drop out (i.e., their probabilities drop from just above $\frac{1}{2}$ to just below), and 7 of the remaining 20 boundaries (inflections) move by 10 MeV.

3.1 "Surgery"

We now discuss the (Bayesian) technique for evaluating the bumps, that is, for finding the odds that each bump would be present in a sample of infinite size. To do the evaluation one would like to compare a hypothesis $f_1(x)$ that shows a specific bump with a similar hypothesis $f_2(x)$ but with just that bump cut off. Before discussing how to cut off a bump we consider two different ways of comparing $f_1(x)$ with $f_2(x)$. If some physical theory has predicted a bump in a specific place, it would be reasonable to subtract the log-likelihood for $f_2(x)$ from that for $f_1(x)$ and to regard the result as the weight of evidence (C.S. Peirce's name for the logarithm of a Bayes factor on the odds) in favor of f_1 as against f_2. (This is not the whole story because it ignores the question of whether the theory predicted the size of the bump.) If there is no such theory, then it is more appropriate to form $\omega(f_1) - \omega(f_2)$, the difference in the overall scores for the two hypotheses. The result will then be the final (posterior) log-odds in favor of the reality of the bump because the scores incorporate the prior that we have selected by fixing the hyperparameter β^*. In Table 4 both methods are used for all 13 bumps. A physicist who has a theory that predicts a bump should combine the weight of evidence with his or her own initial log-odds to obtain his or her own final log-odds if he or she wants to be consistent. In Table 4 we have measured weights of evidence in decibans (db), a term introduced by A.M. Turing in about 1940 by analogy with the decibel (the term was used in G & G′). This unit and the centiban (cb) enable one to use base 10 while avoiding decimal points.

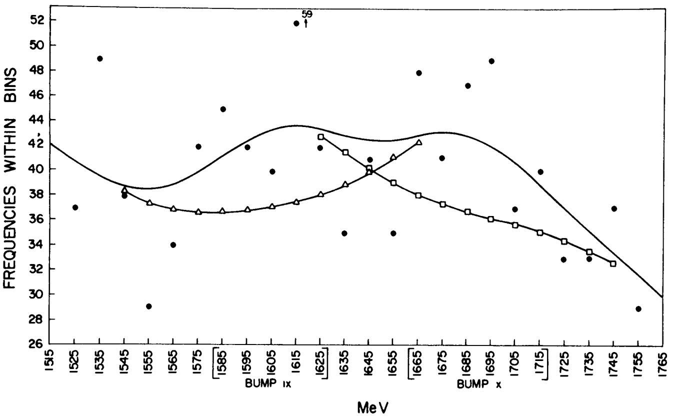
The surgery on each bump was performed by repeated smoothing, that is, by bringing the observations in the vicinity of the bump to the smoothed density, normalizing the total frequency to 25,752, fitting another smoothed density as if the observations were as so modified, and repeating this entire process up to 14 times or until convergence was achieved. (By the vicinity of the bump we mean the part of the curve lying between the two points of inflection that define the bump.) This method of ironing out bumps is intuitively appealing and more objective than making a separate judgment for each bump.

4. Results of Surgery for Evaluation of Bumps (LRL Data)

Bump	*Mid-bump MeV*	*4.343 Times Roughness*	*L Lost*	*RP*	*ω Lost*	*X^2*	*1,000D*
0	—	868	—	195	—	191	3.80
(i)	330	843	15	190	10	202	4.21
(ii)	435	863	9	194	8	196	4.45
(iii)	530	863	4	194	3	193	4.33
(iv)	770	754	58++	170	33++	217	10.86
(v)	940	820	27	184	16	204	4.05
(vi)	1,070	853	11	192	8	196	3.57
(vii)	1,280	840	47++	189	41++	213	6.86
(viii)	1,410	865	1	195	1	192	3.57
(ix)	1,605	865	13	195	13	198	3.45
(x)	1,670	869	13	195	13	197	3.46
(xi)	1,770	868	1	195	1	192	3.66
(xii)	1,845	868	.2	195	.3	191	3.74
(xiii)	1,940	813	89	183	77	278	3.59
Dip	1,005	760	123+	171	103+	243	8.17

NOTE: The first row, marked 0, refers to the smoothing adopted with $\beta^* = .225$. Rows (i) to (xiii) refer to the effects of performing surgery on the 13 bumps in turn, the surgery being very incomplete for bumps (iv) and (vii), which are overwhelmingly significant. The log-likelihoods (L), roughness penalties (RP), and overall scores (ω) are expressed in decibans. The roughness of the standardized density curves has been multiplied by 4.343 so that when further multiplied by β^* one obtains the roughness penalties in decibans. For the adopted smoothing we actually had L = −517,330 and ω = −517,525, and from these values we have subtracted the values of L and ω corresponding to bumps (i) to (xiii). The notations 58++ and 103+, for example, respectively mean that L lost far more than 58 db and that ω lost more than 103 db.

C. Surgery for Bumps (ix) and (x) of LRL *Data With $\beta^* = .225$. (The brackets are placed approximately at the relevant points of inflection. The results of the two surgeries are shown by the curves through the small triangles and squares.)*



The method led to the excision of all the bumps except numbers (iv) and (vii), for which the excision was far from complete after 14 operations. There is no objection to this incomplete surgery because both these bumps are obviously real, and the odds and Bayes factors already attained in Table 4 are in the thousands or tens of thousands. (If more ironing were done the Bayes factors for these two bumps would easily reach the trillions.) Bump (xiii), which is apparently highly significant statistically, may be an artifact caused by the lack of observations beyond the 172nd bin.

As an example, the results of the surgery on the adjacent bumps (ix) and (x), in the vicinities of 1,605 and 1,690 MeV, are shown in Figure C. The odds on the existence of bumps (i), (ii), (v), (vi), (ix), and (x) are substantial although none of these six bumps suggests itself when the raw histogram is examined by eye. We also evaluated the dip, with midpoint at 1,005 MeV, between bumps (v) and (vi), and found it scored well over 100 db, that is, odds of at least 10^{10} to 1 on. This must be the dip at about 980 MeV that is mentioned, with further supporting evidence, by Alston-Garnjost et al. (1971).

We did surgery on bumps (v), (ix), and (x) with $\beta^* = .3$ also. As compared with $\beta^* = .225$, the effect on the log-likelihoods was to increase them by 2.7 db, -1.0 db, and -2.6 db, and the increases in the overall scores were 1.8 db, $-.8$ db, and -1.5 db, respectively. Thus, this change in β^* does not much affect the evaluations of the bumps. For $\beta^* = .4$, we cut off just bump (v) to check whether the further shifts in L and ω were about the same as when β^* was changed from .225 to .3. They were. In fact, for $\beta^* = .4$, the log-likelihood for bump (v) is 4.6 db higher than for $\beta^* = .225$, and the overall score is 3.4 db up (a factor of about 2). Thus, the evaluation of the bumps is not much affected even when the prior density is roughly squared, which is a good indication of Bayesian robustness.

We decided also that any further surgery would not give enough information to justify the computer expense of $30 per surgical run at the lowest priority charge in the VPI&SU computer center, and we consider that it is adequate to evaluate the bumps in accordance with Table 4.

The run with $r^F = 601$ provided another microscopic bump (worth less than 3 cb). This bump occurred at 1,135 MeV, where the putative density curve found with $r = 271$ had a second derivative that was smaller than anywhere else, and so could be regarded as providing a bump of almost zero score, but slightly negative instead of slightly positive. In other words, the effect of increasing r was to raise the probability of this bump from just below to just above .5, a negligible change.

When this work was completed we referred to the Meson Table in Particle Data Group (1976, pp. S26 and S27), which contains all substantial claims for meson resonances. The relevant entries in the table are those of mode (of decay) $\pi\pi$, and the corresponding masses in MeV are 773, 993, 1,270, and 1,600. These masses correspond to our bumps (iv), (v), (vii), and (ix). The meson of 1,600 MeV is not a well-established resonance. The reader should keep in mind that the results in this Meson Table are based on much more information than our Table 1, including information of a different kind, and that a small bump can be real (and of value) in the sense that it would be present in a sample of infinite size without necessarily corresponding to an elementary particle or resonance.

3.2 Comparison With Other Methods

A referee suggested the following quick and dirty methods for smoothing the LRL data, motivated by Tukey (1977, Ch. 7). He took medians of five consecutive bin frequencies, repeated this process on the resulting sequence, and then averaged adjacent values twice and four times. (Doing so reduces the length of the sequence from 172 to 162 and 160 in our interpretation.) When averaging twice this process located our 13 bumps, 5 being split into 2, plus 13 new bumps; while averaging four times gave 12 of our bumps, 1 being split, with only 6 additional bumps. When averaging twice we obtained $X^2 = 135.3$ (161 df) and $D = .002337$, with a left tail-area probability $P(D) = .00104$ (after normalizing over the 163 bins); while, when averaging four times, $X^2 = 140.8$ (159 df), $D = .002502$, $P(D) = .00293$. Thus, these two related methods of smoothing lead to results that are too good to be true. Of course, by decreasing β^* we also increase our number of bumps, the extreme case being to take $\beta^* = 0$, which means no smoothing.

We tried modifying these methods by doing more averaging of adjacent values, because averaging is a smoothing operation and is therefore analogous to increasing β^*. Moreover, criteria of goodness of fit can again be usefully applied. The respective values of X^2 for 6, 8, and 10 averagings were 147.6 (157 df), 154.7

(155 df), and 160.7 (153 df), while the values of D and their left tail-area probabilities were .003109, P = .035 and .003714, P = .128 and .28889, P = .289. In each case 11 of our bumps were located (but not evaluated), together respectively with 5, 4, and only 2 extra bumps. Thus, by repeated averaging, results were obtained that were neither too bad nor sanctimonious. An apparent bump at 665 MeV, near where Mac Gregor wanted one, was not ironed out until the ninth averaging.

A second referee drew a reasonable-looking freehand curve to fit the data from 505 to 875 MeV, and it appeared at first blush to show Mac Gregor's bump. But for these bins we find $X^2 = 11.3$ with at least 37 df so $P \approx 1/70{,}000$. Thus, reasonable-looking freehand fits can be far too good to be true.

3.3 Evaluation of Mac Gregor's Bump

A bump centered at about 665 MeV suggests itself to the eye when looking at the histogram, and, as we have just seen, it does not soon get ironed out when the averaging is carried out by the quick and dirty method. On the other hand, the bump did not appear in our best smoothing, which means that we would regard it as odds against based on our prior. But, as already indicated, because a bump at 655 MeV is suggested by Mac Gregor's theory, we would like to work out the weight of evidence W in favor of the theory, provided by the part of the evidence in the vicinity of the predicted bump. The weight of evidence is given by

$$W = \sum_{i=1}^{172} n_i \log \left\{ \int_{B_i} f_2(x)dx \Big/ \int_{B_i} f_1(x)dx \right\} ,$$

where $f_1(x)$ has no bump at 655 MeV and $f_2(x)$ has one there (or a dip). We take $f_1(x)$ as our best estimate of the probability density, and we convert it to $f_2(x)$ by implanting a bump or dip centered at 655 MeV. We must keep in mind that Mac Gregor's theory does not predict the size of the bump or dip.

For testing Mac Gregor's theory we decided that it was reasonable to implant bumps of the form (.15k, 1.65k, 7.5k, 12.4k, 7.5k, 1.65k, .15k) for various k (positive for bumps and negative for dips), where the seven components here correspond to 625 MeV, 635 MeV, ..., 685 MeV. Denote the seven components by c_i and define c_i as 0 for the other bins, so that i runs from 1 to 172. Also let n_i' denote our smoothing. Then

$$\int_{B_i} f_1(x)dx = n_i'/N ,$$

$$\int_{B_i} f_2(x)dx = (n_i' + c_i)/(N + \sum c_i) .$$

We find the following values for W in centibans:

k	−5	−4	−3	−2	−1	0	¼	½	¾	1	1½	2	3
W	−176	−115	−67	−33	−10	0	1	−1	−6	−13	−35	−68	−161

If we shift the bump 10 MeV to the right, the figures become

k	−½	−¼	0	¼	½	¾	1	1¼	1½	2	2½	3	3½	4
W	−25	−11	0	9	15	18	19	18	14	0	−24	−56	−115	−156

Even if Mac Gregor's theory had predicted a bump with $k = 1$ centered at 665 it would have gained only about 19 cb, corresponding to a Bayes factor of about 1.6. On balance the evidence for or against Mac Gregor's theory, from our analysis, is very slight.

4. RESULTS FOR THE CHONDRITE DATA

For the chondrite data as given in row (ii) of Table 2, but standardized to have mean 0 and variance ½, we used,

5. Analysis of Chondrite Data Where N = 22

β^*	$X_{[6]}^2$	$P_{[6]}$	$X_{[7]}^2$	$P_{[7]}$	$X_{[8]}^2$	$P_{[8]}$	D	$P(D)$	$R(\ )$[a]	R_0
2.56(1)[b]	16.18	.987	8.55	.713	25.45	.999	.1919	.653	.996	.986
1.28(1)[b]	17.45	.992	11.45	.880	11.55	.828	.1634	.454	.979	.723
9.05	17.45	.992	15.82	.973	13.18	.894	.1481	.334	.983	.686
6.40[b]	13.00	.957	18.00	.988	16.45	.964	.1349	.230	.978	.567
4.52	13.00	.957	18.00	.988	16.45	.964	.1428	.291	.978	.618
3.20[b]	16.82	.990	14.36	.955	23.82	.998	.1491	.342	.994	.810
2.26	16.82	.990	14.36	.955	18.09	.979	.1543	.383	.982	.710
1.60[b]	13.64	.966	14.36	.955	15.64	.952	.1585	.416	.959	.605
8.00(−1)[b]	7.27	.703	14.36	.955	15.64	.952	.1637	.457	.935	.554
4.00(−1)[b]	7.27	.703	14.36	.955	16.45	.964	.1633	.454	.943	.576
2.00(−1)[b]	5.36	.501	14.36	.955	16.45	.964	.1582	.414	.942	.544
1.00(−1)[b]	5.36	.501	17.27	.984	16.45	.964	.1479	.332	.968	.589
5.00(−2)[b]	5.36	.501	17.27	.984	9.91	.729	.1328	.215	.957	.421
4.21(−2)	5.36	.501	13.64	.942	9.91	.729	.1286	.184	.869	.101
3.54(−2)	5.36	.501	13.64	.942	9.09	.665	.1242	.154	.865	.039
3.00(−2)	5.36	.501	13.64	.942	9.09	.665	.1199	.127	.865	.017
2.72(−2)	5.36	.501	10.00	.811	9.09	.665	.1173	.112	.709	.288
2.50(−2)[b]	5.36	.501	10.00	.811	9.09	.665	.1151	.099	.709	.318
1.77(−2)	3.45	.249	8.55	.713	11.55	.828	.1061	.057	.621	.522
1.25(−2)[b]	3.45	.249	8.55	.713	5.82	.333	.0974	.028	.423	.895
6.25(−3)[b]	0.91	.011	9.27	.766	5.82	.333	.0857	.008	.105	.972
3.12(−3)[b]	0.91	.011	7.09	.580	5.82	.333	.0754	.001	.081	.995

[a] This column gives the value of $R(X_{[6]}^2, X_{[7]}^2, X_{[8]}^2)$.

[b] Initial values arrived at by starting with $\beta^* = .4$ and doubling (or halving) to generate successively larger (or smaller) values in the table. All the other values of β^* were generated by taking the geometric mean of two other β^*'s.

NOTE: The best value of β^* is .030.

as we did in the final determination of β^* for the LRL data, the X^2 statistic together with the one-sample Kolmogorov-Smirnov statistic D. More precisely, for various values of β^*, we used the X^2 statistic with three different numbers of class intervals, namely, 7, 8, and 9, these being mutually prime in pairs (see Section 2). For each value of β^* we thus had the four statistics $X_{[6]}{}^2$, $X_{[7]}{}^2$, $X_{[8]}{}^2$, and D, the values of which are given in Table 5. These statistics were analyzed as described in Section 2, and the results are shown in Table 5.

The best value of β^* is about .030, and it corresponds to the trimodal curve shown in Figure B. Thus, there seem to be (at least) three different kinds of chondrites, which of course does not allow for evidence from other sources. Even when β^* is .8, for example, the curve still has three bumps but is only bimodal.

APPENDIX A. The Roughness Penalty

The historical reasons for selecting the specific roughness penalty of equation (2.1) are given in G & G′. It would be interesting to try replacing γ'' by f'', in which case we replace ω by ω_1. The f is not necessarily of integrable square, so that some of the theory would be less appealing. On the other hand, ω_1 can be seen to be a concave functional, so that it could have no merely local maximum (with r infinite, where r is defined just after equation (2.3)). The constraint $\int \gamma^2 dx = 1$ prevents ω from being exactly concave, although it may well be very nearly so, but the constraint $\int f dx = 1$, being linear, preserves concavity of ω_1. In practice, local maxima may well be due to taking r finite, whether ω or ω_1 is used, so the apparent advantage of ω_1 over ω may be practically illusory. Further experience on this point would be of interest.

APPENDIX B. The Fourier Normal Orthogonal System

Let $[x]$ denote the greatest integer not exceeding x.

In the Fourier system, where $\gamma(x)$ and $\phi_k(x)$ are defined by (2.3) and (2.4), it can readily be shown that

$$\int \{\gamma''(x)\}^2 dx = b^{-4} \sum_{n=1}^{\infty} \left[\frac{n+1}{2}\right]^4 \gamma_n{}^2 \qquad \text{(B.1)}$$

(compare (B.1) with (18) in G & G′) and that ω is maximized by solving (as in G & G′) the equations

$$\sum \gamma_m{}^2 = 1 \qquad \text{(B.2)}$$

and

$$\sum_{i=1}^{N} \phi_k(x_i) \{ \sum_{m=0}^{\infty} \gamma_m \phi_m(x_i) \}^{-1}$$
$$- \beta b^{-4} \left[\frac{k+1}{2}\right]^4 \gamma_k - \lambda \gamma_k = 0 \quad (k = 0, 1, \ldots) \ . \qquad \text{(B.3)}$$

(Compare (B.2) and (B.3) with (20) and (21) in G & G′.) Also,

$$\lambda = N - \beta b^{-4} \sum_{k=1}^{\infty} \left[\frac{k+1}{2}\right]^4 \gamma_k{}^2 \ . \qquad \text{(B.4)}$$

(Compare (B.4) with (22) and (23) of G & G′.) These comparisons show that the Fourier system is analytically simpler than the Hermite system.

APPENDIX C. Tail Trouble

In our experience $\gamma(x)$ does not change sign in realistic examples except in the extreme tails of the distribution and except for obviously incorrect local maxima in function space. As de Montricher, Tapia, and Thompson (1975) pointed out, when $\gamma(x)$ changes sign there is a discontinuity in the derivative of $f(x)$, which causes our maximization procedure to fall slightly short of maximizing $\omega(f)$. The example of de Montricher, Tapia, and Thompson concerns the case $N = 1$, which is sufficient to demonstrate their theoretically important point, but of course this is not in itself an example of practical concern. Thus, our method of calculation does not necessarily quite attain the MPL and has its own tail trouble. (In some methods of density estimation the estimated density becomes negative in the tails.) We must be careful not to take too seriously extrapolations into the extreme tails, as in any other conceivable nonparametric method.

When $\gamma(x)$ changes sign in the tail beyond the observations it becomes obvious by eye that we could make the tail smoother without changing the likelihood and the smoothing could be done by eye if one wished. The effect would be to obtain slightly more reasonable estimates in the extreme tail where, however, no nonparametric estimate could be reliable.

For these reasons, throughout this article we have referred to maximization of $\omega(f)$ when "approximate maximization" would often be slightly more accurate.

APPENDIX D. Determination of *r* (See Equation (2.3))

The estimate of a density curve cannot be worsened by using an unnecessarily large value of r, but it can be made costly because the running time is asymptotically proportional to r^2. The use of too few coefficients can of course lead to a bad estimate and reduces sensitivity to bumps.

When one is considering whether to increase r from one value to a larger value, one criterion is that the two values should lead to density curves that are distinguishable to the eye. It is better to set a threshold on the squared-distance measure $\int [\gamma_1(x) - \gamma_2(x)]^2 dx$, where $\gamma_1(x)$ and $\gamma_2(x)$ are the two estimates of $\gamma(x)$ (Appendix H). This integral is conveniently expressed as $\sum_m (\gamma_{1m} - \gamma_{2m})^2$ where γ_{1m} and γ_{2m} are the two estimates of γ_m (and are 0 when m exceeds the appropriate value of $r - 1$). If $\sum (\gamma_{1m} - \gamma_{2m})^2 < 10^{-2\nu}$, then of course $|\gamma_{1m} - \gamma_{2m}| < 10^{-\nu}$ for all m.

For the Fourier system we may determine $r (= r^F)$ by deciding the shortest wavelength that we wish to detect. Let h be the width of the thinnest bump we wish to detect (before x is transformed to z); then we shall need sine and cosine terms having period as small as $2h$, because half a sinusoidal wave can be thought of as a bump. But because the sine and cosine terms in (2.4) have period $2\pi b/n$, n must attain the value $[\pi b/h] + 1$. So it is sufficient to take $r = 2[\pi b/h] + 3$ as the number of terms in the Fourier expansion of $\gamma^*(z)$. (In effect this argument is an intuitive form of the Whittaker-Shannon sampling theorem of communication theory (Whittaker 1915; Shannon 1949) although, as in all practical applications of this theorem, we have only a finite sequence of observations and so can expect only a very good approximation.) For continuous data h is at least as large as the accuracy of measurement; for example, in the chondrite data this accuracy was .01. (For the LRL data we took h as a bin width and so took $r^F = 271$. Also see Section 3 where the value $r^F = 601$ is also discussed.)

There might be no simple way to determine r^H in advance, but, as in the Fourier method, a thin bump would correspond to a high harmonic and so would be artificially smoothed out or filtered out of existence if r^H were too small. To see how many terms should be taken to avoid this filtering-out effect, we added various pips (i.e., artificial bumps), of Gaussian shape, to the original LRL data, with the centers of the pips at 665 MeV, and with various widths, and then applied the Hermite MPL program with various values of r^H and $\beta^* = 1/24$, a value that we used for reasons now obsolete. We later increased β^*, but the runs with $\beta^* = 1/24$ remain of value because the increase in β^* makes the putative density curve smoother so that it requires no larger a value of r^H, for approximating the form it would have when $r^H = \infty$.

The runs also gave the attained percentages of heights and of areas of pips. These percentages are given in Table 6 and show that $r^H = 2{,}000$ or $3{,}000$ is adequate but that $r^H = 500$ is inadequate for the width of bump that Mac Gregor hoped to see (when $\beta^* = 1/24$).

Further confirmation that $r^H = 2{,}000$ or $3{,}000$ (Hermite terms) is adequate for the LRL data was obtained by applying the MPL method to the data with $\beta^* = 1/24$, and with various values of r^H, and then tabulating the scores obtained at convergence. The results are shown in Table 7. It is clear from the column of differences that the overall score is within about 1 db of convergence by the time $r^H = 3{,}000$. When we later increased β^* to .225, $r^H = 2{,}000$ or $3{,}000$ gave output that was very accurately consistent with that using the Fourier system. Reconstructing Tables 6 and 7 with the new value of β^* did not seem worth the expense.

6. Percentages of the True Height (and of Areas in Parentheses) of Pips of Normal Shape With Center at 665 MeV*, Height .06, and Various Standard Deviations σ (in* MeV*), and for Various Values of r^H*

	σ		
r^H	7	11	15
500	28.0 (67.8)	55.2 (88.1)	77.0 (96.7)
1,000	47.9 (84.2)	79.2 (97.4)	94.2 (99.8)
2,000	72.0 (95.6)	95.0 (99.8)	99.3 (100.0)
3,000	84.5 (98.7)	98.5 (100.0)	99.7 (100.0)
5,000	94.7 (99.9)	99.5 (100.0)	99.8 (100.0)

NOTE: Computed with the obsolete value $\beta^* = 1/24$.

7. Log-Likelihoods, Roughness Penalties Φ, and Overall Scores ω, After Applying the Penalized Likelihood Method to LRL *Data With $\beta^* = 1/24$ and With Various Values of r^H*

r^H	Log-Likelihood in db	Roughness Penalty	Overall Score ω	Differences
500	−517,281	61	−517,342	25
1,000	−517,240	78	−517,318	10
1,500	−517,223	85	−517,308	4
2,000	−517,215	89	−511,304	2
3,000	−517,211	91	−517,302	

NOTE: The results are here expressed in decibans; that is, the natural units have been multiplied by 10 $\log_{10}e$ = 4.3429. The column of differences shows that there is little to be gained by increasing r^H beyond 3,000. This is true a fortiori for larger values of β^*.

APPENDIX E. On Robustness for Convergence to a Global Maximum: Hermite vs. Fourier

In the absence of any prior information, we have, for standardized data, always regarded $(\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_{r-1}) = (1, 0, 0, \ldots, 0)$ as a reasonable way to initialize the iterative process (for a given value of β) in the Hermite system (see G & G′, p. 177). In the dozen or so different examples in which this procedure has been applied, we have never known it to cause the iterative process to converge to a merely local maximum (even for a trimodal distribution in which the three modes were well separated and of equal height). On the other hand, initialization of the iterative process in the Fourier system at $(1, 0, 0, \ldots, 0)$ is less satisfactory because it is equivalent to starting with a uniform distribution for $\gamma(x)$. For example, when this initialization was used with the LRL data, the Fourier system (with $r = 271$) converged to a merely local maximum.

By converting $(1, 0, 0, \ldots, 0)$ in the Hermite system to its equivalent in the Fourier system (see Appendix I) we are now able to use a normal distribution as an initial estimate of $\gamma^2(x)$ in both systems. (This method failed for the Fourier system for 6 values of β out of 32 values, but the sample of experiments was by no means random.) The actual normal distribution used has mean 0 and variance $\frac{1}{2}$. So, for consistency, one should standardize the sample data so that the mean and variance are 0 and $\frac{1}{2}$, respectively.

Another device incorporated into our program is to start with $r^H = 2$ or $r^F = 11$, to use the resulting coefficients $\gamma_0, \gamma_1, \ldots, \gamma_{r-1}$ as the initialization of a run with r doubled, and so on with repeated doubling. Doing so seemed to encourage convergence to a global maximum for the Fourier method and saved running time for both methods.

If trouble with convergence is encountered while the Fourier system is being used, we recommend restarting the iterative process (with small r) in the Hermite system and finishing it in the Fourier system (which is faster) or, by switching back and forth between the two systems, using the intermediate $\gamma(x)$ of one system to initialize the other system.

On eight separate occasions we tried initializing the Hermite system at the local maximum previously arrived at in the Fourier system and then had convergence to what we believe to be the global maximum.

If a global maximum appears to have been achieved for a given value of β, then the γ coefficients thus computed can be used to initialize runs for other values of β. This method has apparently always succeeded for both the Fourier and Hermite systems.

If a distribution appears, from a sample, to have very thick tails, it may be more appropriate to take the infinite range seriously and to use the Hermite system, but this possibility is conjectural and we have not verified it. Tarter and Kronmal (1976), who used a Fourier method for fitting a Gaussian density, quoted Tukey as saying that Fourier expansion "has its peculiarities."

APPENDIX F. Effect on β of Scaling

In estimating a probability density function

$$f(x) = \gamma^2(x) = \{\sum_{m=0}^{\infty} \gamma_m \phi(x)\}^2$$

from sample data by the method of MPL, we have sometimes found it necessary, especially in the Hermite system, to begin by applying the scale transformation $z = x/c$ to the observations. We then define f^* and γ^* by $f(x)dx = f^*(z)dz$ and $\gamma^*(z) = (f^*(z))^{\frac{1}{2}}$, and we estimate the coefficients $\gamma_m{}^*$ in

$$\gamma^*(z) = \sum_{m=0}^{\infty} \gamma_m{}^* \phi_m(z) \ .$$

Because $\gamma(x) = c^{-\frac{1}{2}}\gamma^*(z)$, we have $\gamma''(x) = c^{-\frac{5}{2}}\gamma^{*\prime\prime}(z)$ and

$$\beta_x \int \gamma''^2(x)dx = c^{-4}\beta_x \int (\gamma^{*\prime\prime}(z))^2 dz \ .$$

Thus, $c^4\beta_z = \beta_x$, where β_x and β_z are what we usually call β and β^*. For the chondrite data $\beta/\beta^* = .02198$, and for the LRL data $\beta/\beta^* = .03556$.

APPENDIX G. Explanation of Sign Correlations Among the Hermite Coefficients

Because the Hermite functions satisfy the recurrence relation

$$\phi_m(x) = (2/m)^{\frac{1}{2}} x \phi_{m-1}(x) - (1 - m^{-1})^{\frac{1}{2}} \phi_{m-2}(x)$$

and because $|\phi_n(x)| < .81605$ (Sansone 1959, p. 324, attributes this inequality to Charlier 1930, p. 52), it follows that, if $|x|m^{-\frac{1}{2}}$ is small,

$$\phi_m(x) \simeq -(1 - m^{-1})^{\frac{1}{2}} \phi_{m-2}(x) \ .$$

This provides a partial explanation of why the four sequences $(\gamma_0, \gamma_4, \gamma_8, \ldots)$, $(\gamma_1, \gamma_5, \gamma_9, \ldots)$, $(\gamma_2, \gamma_6, \gamma_{10}, \ldots)$, and $(\gamma_3, \gamma_7, \gamma_{11}, \ldots)$ are smooth and why long runs of $+ + - - + + - - + + - - \ldots$ occur in the signs of the sequence $(\gamma_0, \gamma_1, \gamma_2, \ldots)$ in the Hermite system. This argument also suggests that the information from γ_{m+4} largely duplicates that of γ_m if $\gamma(x)$ is small whenever $|x|m^{-\frac{1}{2}}$ is not small.

APPENDIX H. A Convergence Criterion for Estimating the Density Curve

For a fixed value of r, the iterative process for estimating the density curve $\gamma^2(x)$ can be considered to have converged if two successive estimates (denoted by $[\gamma^{(i)}(x)]^2$ and $[\gamma^{(i+1)}(x)]^2$) are indistinguishable to the eye. This will be the case if

$$|[\gamma^{(i)}(x)]^2 - [\gamma^{(i+1)}(x)]^2| < 10^{-\nu} \max_x \gamma^2(x) \quad (a_1 < x < a_2) \ , \qquad (\text{H.1})$$

where $\nu \geq 1$ is chosen as large as necessary ($\nu = 3$ should suit most purposes) and (a_1, a_2) represents any finite interval outside of which $\gamma^2(x)$ is assumed to vanish or to be negligible. Factorizing the left side of (H.1), we have

$$|\gamma^{(i)}(x) + \gamma^{(i+1)}(x)| \, |\gamma^{(i)}(x) - \gamma^{(i+1)}(x)| < 10^{-\nu} \max_x \gamma^2(x) \ ,$$

which, because $\gamma^{(i)}(x) \simeq \gamma(x)$, we can rewrite as

$$\gamma(x) |\gamma^{(i)}(x) - \gamma^{(i+1)}(x)| < 10^{-\nu} \max_x \gamma^2(x) \ , \qquad (\text{H.2})$$

(by throwing away a factor of 2 as insurance). Squaring both sides, replacing $\gamma^2(x)$ by its maximum, and integrating gives

$$\int_{a_1}^{a_2} [\gamma^{(i)}(x) - \gamma^{(i+1)}(x)]^2 dx < 10^{-2\nu}(a_2 - a_1) \max_x \gamma^2(x) \ . \qquad (\text{H.3})$$

But $(a_2 - a_1) \max_x \gamma^2(x) > 1$; hence, (H.3) is satisfied if

$$\int_{a_1}^{a_2} [\gamma^{(i)}(x) - \gamma^{(i+1)}(x)]^2 dx < 10^{-2\nu} \ . \qquad (\text{H.4})$$

Thus, if (for some ν) we are satisfied with the convergence criterion stated in (H.1), the iterative process can be halted whenever the sum of the squared differences between corresponding coefficients for two successive estimates of $\gamma^2(x)$ is less than $10^{-2\nu}$.

APPENDIX I. Converting From One Normal Orthogonal System to Another

Given the coefficients $\gamma_0{}^F$, $\gamma_1{}^F$, $\ldots$ in the Fourier expansion of the square root of a density function

$$\gamma(x) = \sum_{m=0}^{\infty} \gamma_m{}^F \phi_m{}^F(x) \quad (-\pi b < x < \pi b) \ , \qquad (\text{I.1})$$

we wish to be able to compute the coefficients $\gamma_0{}^H$, $\gamma_1{}^H$, ... in the Hermite expansion

$$\gamma(x) = \sum_{m=0}^{\infty} \gamma_m{}^H \phi_m{}^H(x) \quad (-\infty < x < \infty) \qquad (\mathrm{I}.2)$$

and vice versa. (See (2.4) for the definition of $\phi_m{}^F(x)$ and (23.1) of G & G′ for $\phi_m{}^H(x)$.)

Defining $\gamma(x) = 0$ outside $[-\pi b, \pi b]$, we have

$$\gamma_n{}^H = \int_{-\infty}^{\infty} \gamma(x)\phi_n{}^H(x)dx$$

$$= \int_{-\pi b}^{\pi b} \gamma(x)\phi_n{}^H(x)dx \quad (n = 0, 1, \ldots) \ . \qquad (\mathrm{I}.3)$$

Now replace $\gamma(x)$ in (I.3) by its Fourier expansion in (I.1) and obtain

$$\gamma_n{}^H = \sum_{m=0}^{\infty} \gamma_m{}^F \int_{-\pi b}^{\pi b} \phi_m{}^F(x)\phi_n{}^H(x)dx \qquad (n = 0, 1, \ldots) \qquad (\mathrm{I}.4)$$

where we have assumed the same scaling for the two systems.

Unfortunately, there are no known explicit formulas for evaluating (I.4) (although there is such a formula for the interval $(-\infty, \infty)$), because the integrand, unlike $\gamma(x)$, is not negligible outside $(-\pi b, \pi b)$. Therefore, we must resort to numerical integration. First we can quarter the amount of work by noting that, because $\phi_k(x)$ is an even function of x when k is even and odd when k is odd, for both the Fourier and Hermite systems, we have

$$\int_{-\pi b}^{\pi b} \phi_m{}^F(x)\phi_n{}^H(x)dx = 2\int_{0}^{\pi b} \phi_m{}^F(x)\phi_n{}^H(x)dx \qquad (\mathrm{I}.5)$$

if m and n are both even or both odd and 0 otherwise. It follows from (I.4) and (I.5) that

$$\gamma_{2n'}{}^H = 2\sum_{m=0}^{\infty} \gamma_{2m}{}^F \int_{0}^{\pi b} \phi_{2m}{}^F(x)\phi_{2n'}{}^H(x)dx \qquad (n' = 0, 1, \ldots) \qquad (\mathrm{I}.6)$$

and

$$\gamma_{2n'+1}{}^H = 2\sum_{m=1}^{\infty} \gamma_{2m-1}{}^F \int_{0}^{\pi b} \phi_{2m-1}{}^F(x)\phi_{2n'+1}{}^H(x)dx \ .$$

Conversely, starting with $\gamma(x)$ expressed in the Fourier system as in (I.1) multiplying through by $\phi_n{}^F(x)$ and integrating over $[-\pi b, \pi b]$ gives us

$$\gamma_n{}^F = \int_{-\pi b}^{\pi b} \gamma(x)\phi_n{}^F(x)dx \quad (n = 0, 1, 2, \ldots) \ . \qquad (\mathrm{I}.7)$$

We now have only to reverse the roles of the Hermite and Fourier systems in the previous derivation in order to arrive at results identical with (I.4), (I.5), and (I.6), except that the superscripts H and F are interchanged.

Thus, given the Fourier (or Hermite) coefficients, we can efficiently compute the Hermite (or Fourier) coefficients in pairs by applying the midpoint (or other) method of numerical integration to (I.6) (or to (I.6) with the superscripts H and F interchanged).

APPENDIX J. Computer Programs for MPL

The numerical results quoted in this article are based on a set of 12 MPL computer routines together with various driver programs. All routines are written in FORTRAN IV and were run in double precision on the twin IBM 370/158 computers at the VPI&SU computer center. We hope to publish these routines elsewhere. They can be supplied at cost together with a driver program.

Double-precision core requirements are as follows. The Fourier system consisting of four Fourier, two Hermite, and six mixed routines requires $26{,}000 + 8(4.5r^F + 8J + 6N^I) + \mathrm{DP}$ bytes of computer storage, where N^I is the number of intervals used in computing X^2 and DP is the amount of storage required by the driver program (from 50,000 to 100,000 bytes in our work). The full Fourier-Hermite system consisting of four Fourier, four Hermite, and seven mixed routines requires $35{,}000 + 8(4.5r^F + 8r^H + 16J + 6N^I) + \mathrm{DP}$ bytes. Total storage requirements for our work with the chondrite and LRL data was between 128,000 and 256,000 bytes.

[*Received April 1978. Revised October 1978.*]

REFERENCES

Ahrens, L.H. (1965), "Observations on the Fe-Si-Mg Relationship in Chondrites," *Geochimica et Cosmochimica Acta*, 29, 801–806.

Alston-Garnjost, M., Barbaro-Galtieri, A., Flatté, S.M., Friedman, J.H., Lynch, G.R., Protopopescu, S.D., Rabin, M.S., and Solmitz, F.T. (1971), "Observation of an Anomally [sic] in the $\pi^+\pi^-$ System at 980 MeV," *Physics Letters*, 36B, 2, 152–156.

Boneva, L.I., Kendall, D.G., and Stefanov, I. (1971), "Spline Transformations: Three New Diagnostic Aids for the Statistical Data-Analyst" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 33, 1–71.

Burch, C.R., and Parsons, I.T. (1976), "Squeeze" Significance Tests." *Applied Statistics*, 25, 287–291.

Charlier, C.V.L. (1930), *Les applications de la théorie des probabilités aux sciences mathématiques et aux sciences physiques; Application de la théorie des probabilités à l'astronomie*, Paris: Gauthier-Villars.

Fisher, R.A. (1938), *Statistical Methods for Research Workers* (7th ed.), Edinburgh and London: Oliver and Boyd.

Gončarov, V. (1943), "Sur la succession des événements dans une série d'épreuves indépendantes répondant au schème de Bernoulli," *Comptes Rendus (Doklady) de l'Académie des Sciences de l'URSS*, 38, 283–285.

——— (1962), "On the Field of Combinatory Analysis," *American Mathematical Society Translations*, Ser. 2, 19, 1–46 (trans. from the Russian version published in 1944).

Good, I.J. (1952), "Rational Decisions," *Journal of the Royal Statistical Society*, Ser. B, 14, 107–114.

——— (1953), "The Serial Test for Sampling Numbers and Other Tests for Randomness," *Proceedings of the Cambridge Philosophical Society*, 49, 276–284.

——— (1957), "On the Serial Test for Random Sequences," *Annals of Mathematical Statistics*, 23, 262–264.

——— (1958), "Significance Tests in Parallel and in Series," *Journal of the American Statistical Association*, 53, 799–813.

——— (1963), "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," *Annals of Mathematical Statistics*, 34, 911–934.

——— (1965), *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, Cambridge, Mass.: MIT Press.

——— (1967), "A Bayesian Significance Test for Multinomial Distributions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 29, 399–431.

——— (1971a), "The Probabilistic Explication of Information, Evidence, Surprise, Causality, Explanation, and Utility" (with appendix, discussion, and replies), in *Foundations of Statistical Inference*, eds. V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston of Canada, 108–141.

——— (1971b), "Non-parametric Roughness Penalty for Probability Densities," *Nature Physical Science*, 229, 29–30 (contains 21 misprints).

——— (1978a), "An Improvement to the Chi-squared Goodness-of-Fit Test," in "Comments, Conjectures and Conclusions," *Journal of Statistical Computation and Simulation*, 7, 79–80.

——— (1978b), "Moments of the Kolmogorov-Smirnov One-Sample Statistic," in "Comments, Conjectures and Conclusions," *Journal of Statistical Computation and Simulation*, 7, 289–290.

——— (1978c), "A Signs Test When Estimating Probability Densities," in "Comments, Conjectures and Conclusions," *Journal of Statistical Computation and Simulation*, 7, 290–292.

——— (1979), "A Comment on Runs of Signs," in "Comments, Conjectures and Conclusions," *Journal of Statistical Computation and Simulation*, 8, 311–312.

Good, I.J., and Crook, J.F. (1974), "The Bayes/Non-Bayes Compromise and the Multinomial Distribution," *Journal of the American Statistical Association*, 69, 711–720.

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

——— (1972), "Global Nonparametric Estimation of Probability Densities," *Virginia Journal of Science*, 23, 4, 171–193.

Krishna Iyer, P.V. (1950), "The Theory of Probability Distributions of Points on a Lattice," *Annals of Mathematical Statistics*, 21, 198–217.

Leonard, Tom (1978), "Density Estimation, Stochastic Processes, and Prior Information" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 40, 113–146.

Mardia, K.V. (1972), *Statistics of Directional Data*, London: Academic, Press.

de Montricher, G.F., Tapia, R.A., and Thompson, J.R. (1975), "Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods," *Annals of Statistics*, 3, 1329–1348.

Omnés, R., and Froissart, M. (1963), *Mandelstam Theory and Regge Poles*, New York and Amsterdam: W.A. Benjamin.

Orear, J., and Cassel, D. (1971), "Applications of Statistical Inference to Physics," in *Foundations of Statistical Inference*, eds. V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston of Canada, 280–288.

Particle Data Group (1976), "Review of Particle Properties," *Reviews of Modern Physics*, 48, 2, Part II, S1–S246.

Pelz, W. (1977), "Maximum Likelihood/Entropy Estimation," part of a doctoral dissertation, Virginia Polytechnic Institute and State University.

Sansone, G. (1959), *Orthogonal Functions* (trans. from Italian by A.H.D. Diamond), New York: Wiley Interscience.

Shannon, E.C. (1949), "Communication in the Presence of Noise," *Proceedings of the Institution of Radio Engineers*, 37, 10–21.

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

Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley.

Uspensky, J.V. (1937), *Introduction to Mathematical Probability*, New York: McGraw-Hill Book Co.

Wahba, Grace (1976), "Optimal Smoothing of Density Estimates," Technical Report No. 469, University of Wisconsin, Madison, Dept. of Statistics.

Whittaker, E.T. (1915), "On the Functions Which Are Represented by the Expansions of the Interpolatory Theory," *Proceedings of the Royal Society of Edinburgh*, 35, 181–194.

Comment

EMANUEL PARZEN*

It gives me great pleasure to discuss a paper on the estimation of probability density functions and the location of bumps. There is an extensive literature on density estimation, but many statisticians seem doubtful about the usefulness of these techniques because their application seems subjective and complicated. A major criticism I would make of this paper is that it does not help to dispel this negative attitude of statisticians toward density estimation. One cannot help but be impressed by the ingenuity of Good and Gaskins and even to believe that they may be able to successfully fit probability densities to data. But I have doubts if other statisticians would find their approach a practical method for daily use in statistical data analysis.

I find it strange that the authors would dismiss the cross-validation method of Grace Wahba on the ground that it requires statisticians "to understand much more advanced mathematical concepts than we need in the present work." I believe that statistical computing has changed the statisticians' criterion for what makes a statistical technique effective; it is not how easy it is to compute by hand, or how easy it is to understand in a rigorous sense the theory of the technique, but how easy it is for a statistician to communicate to a client how to interpret the computer output (one hopes graphical) of a program implementing the techniques. I conjecture that Good and Gaskins may have more work to do to convince statisticians of the practical effectiveness of their techniques of density estimation. The unique aspect

* Emanuel Parzen is Distinguished Professor at the Institute of Statistics, Texas A&M University, College Station, TX 77843. This research was supported in part by the Army Research Office, Grant DAAG-29-78-G-0180.

© Journal of the American Statistical Association
March 1980, Volume 75, Number 369
Applications Section