Nonparametric roughness penalties for probability densities

BY I. J. GOOD

Virginia Polytechnic Institute

AND

R. A. GASKINS

Hampden-Sydney College

SUMMARY

Given a number of observations \( x_1, \ldots, x_N \), a nonparametric method is suggested for estimating the entire probability density curve. The method is to subtract a roughness penalty from the log likelihood, where the roughness penalty is a certain functional of the assumed density function \( f \). Those used are linear combinations of \( \int y'^2 \, dx \) and \( \int y''^2 \, dx \), where \( y = \int f \). The method appears to be consistent under wide conditions, although consistent methods can be rough. Numerical examples are given and show that for certain values of the coefficients in this linear expression the density function turns out to be very smooth even when \( N \) is small. Multivariate extensions are proposed, including one to distributions having some continuous and some discrete components, but numerical examples of these have not been tried. Some of the techniques are borrowed from quantum mechanics and tensor calculus.

1. INTRODUCTION

A fundamental problem in statistics is the estimation of a density function from a sample of observations. An example of a useful but not very familiar application would be for the estimation of probability densities of discriminant functions. This could be used, for example, for choosing between treatments when a discriminant function is available for success and failure in regard to each treatment. Another kind of application is for deciding whether a bump in the apparent density function is genuinely in the population, or more generally for deciding whether one density function is a better estimate than another one and by how much. According to Orear & Cassel (1971), ‘bump hunting’ is one of the major current activities of experimental physicists, and they requested aid from statisticians. When a bump is not due to random variation it is usually due to a new elementary particle or resonance.

The method of estimating a density function to be discussed in this paper arose from a proposal of Good (1971a, b) and is here modified, developed further, and exemplified numerically and graphically. A multivariate extension is proposed but no multivariate numerical calculations have yet been done.

2. THE BASIC APPROACH

For the estimation of a density function \( f \) of a random scalar or \( p \)-dimensional vector, \( x \), given \( N \) observations \( x_1, \ldots, x_N \), the likelihood or its logarithm

\[
L = \sum_i \log f(x_i)
\]

is relevant; but a naïve application of maximum likelihood methods would make the estimate the mean of a set of Dirac functions at the \( N \) observations. Nobody accepts this
solution because it is too rough. It must be better to subtract a penalty for roughness from
the log likelihood before maximizing. Accordingly our basic approach is to maximize
a score $\omega$, depending on $f$, $x_1, \ldots, x_N$, and defined by

$$\omega = \omega(f) = L - \Phi(f),$$

where the roughness penalty $\Phi$ is a functional $\Phi(f) = \Phi(f(\cdot))$ of $f$. Of course $L$ too is a functional of $f$ but it depends on the observations in addition. Note the mnemonics that $\omega$ is for
‘overall score’ and $\Phi$ is for ‘flamboyance functional’, flamboyant in the sense of having
a wavy edge.

Although this approach seems reasonable from a non-Bayesian point of view, it was
proposed for a Bayesian reason, and the Bayesian interpretation has the advantage that it
gives some measure for how much better one density function $f$ is than another one. It thus
contributes to bump theory.

A full-dress Bayesian approach would involve the maximization of the expected utility
of an assumed $f$. One would need, among other things, a formula for the utility of assuming
one density function when another one is true, as discussed by Good (1968b, 1969). Such an
approach would be very complicated so we use only a Bayesian approach in multi, in which
the aim is to maximize the final density of $f$ in function space. This is analogous to a familiar
method for estimating a scalar or vector, of which the ordinary method of maximum likeli-
hood is the special case where the prior is uniform. A uniform prior in function space is out
of the question, and would lead to the Dirac catastrophe mentioned before. The prior
density in function space implied by (2) is proportional to

$$e^{-\Phi}.$$  

When this Bayesian approach is used, with a specific functional, the final log odds in
favour of one density function $f_1$ as compared with another one $f_2$ is

$$\omega(f_1) - \omega(f_2) \quad \text{natural bans},$$

where a natural ban is the unit suggested in 1940 by A. Turing for log odds and ‘weights of
evidence’ when the base of the logarithms is $e$. When the base is 2, 10, $10^{91}$ and $10^{90}$, the
unit is a bit, a non, a deciban and a centiban, respectively.

When the aim is merely to obtain the ‘best’ smoothing, the Bayesian interpretation can
be ignored.

The choice of the penalty functional $\Phi$ will be considered in § 5, and, for the multivariate
case, in § 9. Section 10, on invariance, is relevant to both cases. We shall always assume that

$$\Phi(f) \geq 0,$$

whenever $\Phi(f)$ is defined, and that

$$\Phi(f_0) < \infty$$

for the true density $f_0$. We do not rule out infinite values for $\Phi(f)$, but such $f$’s will score
minus infinity so they will rule themselves out.

Density estimation is probability estimation for continuous distributions. For discrete
distributions an analogue of a roughness penalty was proposed by Good (1963, p. 931; 1965,
p. 76), namely the entropy, as a generalization of the principle of maximum entropy. The
entropy seems inappropriate for continuous distributions because they could have violent
small ripples with little effect on the entropy, and it is better that $\Phi$ should depend on the
derivatives of $f$. For multivariate distributions with some continuous and some discrete
components, see § 9.
3. Consistency

We shall give a ‘physicist’s proof’ of the following theorem. A rigorous proof might require a further constraint on the allowable density functions.

**PLAUSIBLE THEOREM.** Suppose that we consider only those density functions that are continuous, and satisfy the condition

\[
\int f(\log f)^2 \, dx < \infty. \quad (7)
\]

Then the estimate \( f \), obtained by maximizing \( w \), is consistent in the sense that in probability

\[
\int_b^a f(x) \, dx - \int_b^a f_0(x) \, dx \to 0 \quad \text{whenever } b < a, \text{ where } f_0 \text{ is the true density function.}
\]

Unspecified ranges of integration are taken to be \((-\infty, \infty)\). Consistency does not imply the absence of false bumps: they could become small and numerous as \( N \) increases.

**Proof.** To prove this, note that, by (5), in a sample of \( N \) observations,

\[
E[\omega(f_0) - \omega(f)] \geq N \int f_0(x) \log \frac{f_0(x)}{f(x)} \, dx - 2 \Phi(f_0).
\]

(9)

The integral is nonnegative and can vanish only if \( f = f_0 \). Moreover, if we assume

\[
\int_{-\infty}^{\infty} f_0(x) (\log f_0)^2 \, dx < \infty, \quad (7A)
\]

and we shall come back to this, the standard deviation of \( L \) is \( O(\sqrt{N}) \) so that, by Chebychev’s inequality, in probability \( \omega(f_0) - \omega(f) \) is ultimately positive. Therefore, by (6), in probability, any specified \( f \neq f_0 \) is ultimately beaten by \( f_0 \) when \( N \) is increased. In a large sample, in probability, those \( f \)’s that are not yet beaten will have the integral negligibly small so that the limit in (8) is true in probability.

Although (7A) does not follow from (7) it fails to follow only in virtue of the possibility that \( f \) is much smaller than \( f_0 \) at infinity. If \( f \) is replaced by another density function \( f^* \) for which this is not true, in which \( f \) is slightly increased at infinity and slightly decreased elsewhere, the expression \( \omega(f_0) - \omega(f^*) \) will be almost the same as \( \omega(f_0) - \omega(f) \) or will be smaller, whereas condition (7A) will be true with \( f \) replaced by \( f^* \). This justifies us in assuming (7A) in the course of the proof. Another point we have slurred over is that, under condition (7A),

\[
\left( \int f_0 \log f \, dx \right)^2 \leq \left( \int f_0 \sqrt{f_0 \log f} \, dx \right)^2 \leq \int f_0 \, dx \int f_0 (\log f)^2 \, dx < \infty.
\]

(10)

This was implicit when we claimed that the standard deviation of \( L \) was \( O(\sqrt{N}) \).

The condition (7) is weak and is satisfied, for example, if \( f \in L^{1, \infty}(\infty, \infty) \).

4. Comparison with other methods

If the statistician makes the assumption that the true density function belongs to a family with a small finite number of parameters, then he can estimate these parameters, for example, by the method of moments, by maximum likelihood, or by a Bayesian method.
When the assumption can be safely made it is a pity to waste it. Our procedure, being non-parametric, does not depend on such an assumption. Other nonparametric methods are the 'window methods' (Rosenblatt, 1956; Parzen, 1962), a Bayesian method of Whittle (1958), the use of the Gram-Charlier or Edgeworth series, and a method of Boneva, Kendall & Stefanov (1971), which came to our attention when the present work was almost complete. Although the window methods are consistent they perhaps require large samples, whereas our method makes more use of all the observations at once. Its applicability to small samples might be a valuable feature. It has a Bayesian interpretation, which enables it to come to grips with bumps. It breaks away from the window methods in treating the density function nonlinearly; in this respect it has something in common with the method of Boneva, Kendall and Stefanov. That method also puts a premium on small values of a functional of the density, namely \( \int f'^2 \, dx \) for the one-dimensional case \( (p = 1) \). One important distinction is that their purpose is to smooth without much other change to the histogram, so they do not remove its bumps. Their results are locally smooth but oscillatory in the large. This is not a criticism of their method as an analyst's aid for the apprehension of the data more or less undoctored. Our method, on the other hand, aims to estimate the true density and to iron out nonsignificant bumps: compare our very smooth Fig. 5 with their bumpy Fig. 2, both of which use the same data. We obtain no negative estimates. Also our method seems to be much more widely applicable than the Gram-Charlier or Edgeworth series. See also the paragraph following equation (16). Whittle's method depends on the assumption of a prior distribution for the density at each point \( x \). This is a much less general-purpose assumption than ours, and requires more experience in a given problem area to justify it.

5. CHOICE OF THE ROUGHNESS PENALTY \( \Phi \)

One measure of the roughness or complexity of a density function \( f \) is the number of bumps in it, which can be defined as one half of the number of points of inflexion (Good, 1950, p. 86). Somewhat in the spirit of the discussion of complexity by Jeffreys (1961, Chapter 3), who regards \( 2^{-n-1} \) as the prior probability that a law has \( n \) adjustable parameters, the roughness penalty could be taken as proportional to the number of points of inflexion, this being equivalent to assuming that the initial probability of \( b \) bumps is of the form \( e^{-b} \). Compare also the discussion of complexity by Good (1968a). But more information about \( f \) is taken into account if the penalty functional of \( f \) is assumed to depend on all the values of \( f' \) and \( f'' \) or on the curvature of the graph of \( f \).

Good (1971a, b) suggested that the roughness might be measured by the ease of discriminating the density curve or hypersurface from itself when it is slid bodily by a slight displacement. Although the suggestion was along Bayesian lines it led, in the one-dimensional case, to the proposal that the roughness penalty might be taken proportional to Fisher's information concerning the displacement or location, regarded as a parameter,

\[
\int (f'^2/f) \, dx. \tag{11}
\]

On putting \( f = \gamma^2 \), this integral becomes

\[
4 \int \gamma'^2 \, dx. \tag{12}
\]
Working with $y$ instead of $f$ replaces the awkward constraint $f \geq 0$ by the convenient one that $y$ is real. Moreover, (12) is slightly simpler than (11). The constraint $\int f \, dx = 1$ becomes $\int \gamma^2 \, dx = 1$ so that $\gamma \in L^2(-\infty, \infty)$, that is, it is of integrable square. Therefore, by the Riesz–Fischer theorem, $\gamma$ is the limit in mean square of its expansion according to any normal orthogonal system of functions such as the Hermite functions (Riesz, 1907a, b; Fischer, 1907). The theory of the expansion of functions belonging to $L^2$ is simpler than that of functions belonging to $L^1$.

Apart from the convergence in mean square, questions of ordinary convergence are also of interest and some attention will be paid to them.

When we tried the penalty $4\alpha \int \gamma'^2 \, dx$, we found it sometimes led to density curves having portions that looked too straight. This was not very surprising because curvature depends also on second derivatives. Since similar methods of analysis were available for $\int \gamma''^2 \, dx$, we went over to the more general roughness penalty

$$\Phi(f) = 4\alpha \int \gamma'^2 \, dx + \beta \int \gamma''^2 \, dx \quad (\alpha \geq 0, \beta > 0, \alpha + \beta > 0).$$

(13)

A referee has commented that if $A_1, A_2$ is a dichotomy of $(-\infty, \infty)$,

$$p_i = \int_{A_i} f \, dx$$

and $f_i$ is the density conditional on $x \in A_i$, then

$$\Phi(f) = p_1 \Phi(f_1) + p_2 \Phi(f_2),$$

which is a neat feature.

The results of the numerical work are given in § 7, where it is found that smooth density functions were obtained with $\alpha = 0$ and various values of $\beta$, even for samples as small as $N = 1, 2, 3$. An intuitively reasonable constraint on $\alpha$ and $\beta$ is given in Appendix D.

Some lower bounds for the penalty, in terms of the moments of $f$, are given in Appendix E.

A smooth density function does not remain smooth for all transformations of the $x$ axis, and would be expected to be smooth only if $x$ is chosen as a physically natural variable. This point is discussed in § 10.

6. METHOD OF CALCULATION

We now consider how to maximize the score $\omega(f) = L - \Phi(f)$ when $\Phi$ is given by (13). We then write $\omega(f) = \omega(\alpha, \beta; f) = \omega$. The problem of maximizing $\omega$, when $\alpha$ and $\beta$ and of course $x_1, \ldots, x_N$ are given, is a problem in the calculus of variations for which Euler's equations turn out to be inapplicable. We can, however, use the Rayleigh–Ritz method (Courant & Hilbert, 1953, p. 175) with any normal orthogonal system. We select the Hermite system; that is, we use the formal expansion

$$\gamma(x) = \sum_{m=0}^{\infty} \gamma_m \phi_m(x),$$

(14)

where the $\gamma_m$ are real and

$$\phi_m(x) = e^{-i x^2} H_m(x) 2^{-\frac{m}{2}} \sqrt{n} n^{-\frac{1}{4}} (m!)^{-\frac{1}{4}},$$

(15)

$$H_m(x) = (-1)^m e^{x^2} \left( \frac{d}{dx} \right)^m e^{-x^2}.$$
Since $\gamma \in L^2(-\infty, \infty)$, (14) is at least true in mean square. It is also true in the sense of Abel summability; see Appendix F. Moreover, if $\gamma \in L(-\infty, \infty)$ and is of bounded variation in every finite interval, then the series converges everywhere to $\frac{1}{2}(\gamma(x+0) + \gamma(x-0))$ and therefore to $\gamma(x)$ at all points of continuity of $\gamma$. This follows from a theorem of Galbrun and Cramér; see, for example, Kendall & Stuart (1969, p. 161). Their notation for $H_m$ differs from ours. These are much less restrictive conditions on $\gamma$ than the condition for $f$ to have a Gram–Charlier Type A expansion, which is that $\gamma^2 e^{ix} \in L(-\infty, \infty)$.

We shall also want to differentiate (14) term by term. Owing to the form of $\Phi(f)$, one at least of $\gamma'$ and $\gamma''$ must be of integrable square: we shall assume both of them to be. Then at any rate $\gamma'$ and $\gamma''$ are limits in mean square of their Hermite expansions. Some sufficient conditions for convergence are given in Appendix F.

In numerical work we selected the origin of co-ordinates at the mean $\bar{x}$ of the sample, and scaled the $x$-axis to force the usual unbiased estimate of the variance to be equal to $\frac{1}{2}$. A nonlinear transformation with the same result could be used in accordance with §10. This is the most natural choice because when $f$ is a normal distribution of zero mean and variance $\frac{1}{2}$ we have $\gamma = \phi_0$.

By means of the formalism that underlies the equivalence of Schrödinger’s and Heisenberg’s formulations of the quantum mechanics of the harmonic oscillator (Appendix A) it can be neatly shown that

$$\int \gamma'^2 dx = \sum_{m=0}^{\infty} (m+\frac{1}{2}) \gamma_m^2 - \sum_{m=0}^{\infty} ((m+1)(m+2))^{\frac{1}{2}} \gamma_m \gamma_{m+2},$$

$$\int \gamma''^2 dx = \frac{1}{2} \sum (2m^2 + 2m + 1) \gamma_m^2 - \sum (2m+3)(m+1)(m+2)^2 \gamma_m \gamma_{m+2}$$

$$+ \frac{1}{2} \sum (2m+3)(m+3)(m+4)^2 \gamma_m \gamma_{m+4}.$$  \hspace{1cm} (17)

Further

$$L = \sum_{i=1}^{N} \log \left( \sum_{m=0}^{\infty} \gamma_m \phi_m(x_i) \right)^2.$$  \hspace{1cm} (18)

Hence $\omega(x, \beta; f)$ is maximized by solving the equations

$$\Sigma \gamma_m^2 = 1,$$  \hspace{1cm} (20)

$$2 \sum_{i=1}^{N} \phi_k(x_i) \left( \sum_{m=0}^{\infty} \gamma_m \phi_m(x_i) \right)^{-1} - 4\lambda [2k \gamma_k - \{(k+1)(k+2)\}^{\frac{1}{2}} \gamma_{k+2} - (k-1)k \gamma_{k-2}]$$

$$- \beta [3k^2 \gamma_k + 3k \gamma_k - (2k+3)(k+1)(k+2)]^{\frac{1}{2}} \gamma_{k+2}$$

$$- (2k-1)(k-1)k^{\frac{1}{2}} \gamma_{k-2} + \frac{1}{2}((k+1)(k+2)(k+3)(k+4))^{\frac{1}{2}} \gamma_{k+4}$$

$$+ \frac{1}{2}((k-3)(k-2)(k-1)k)^{\frac{1}{2}} \gamma_{k-4} - 2\lambda \gamma_k = 0 \quad (k = 0, 1, \ldots)$$  \hspace{1cm} (21)

for $\lambda, \gamma_0, \gamma_1, \ldots$. For example, when $N = 1, x_1 = 0, \alpha = \frac{1}{2}$ and $\beta = 0$, and if we disregard $\gamma_k$ for $k \geq 4$, so that the upper limit of summation becomes 3 instead of infinity, then we find that $\gamma_0 = 1$, so that the ‘best’ $f$ is normal to this degree of approximation; see Appendix B. This is one of the few cases that can be solved by hand. Its reasonableness encouraged us to write computer programs. All the programming was done by R. A. Gaskins.

We used the following method for the computer solution. Take an integer $r$ and assume that $\gamma_r = \gamma_{r+1} = \ldots = 0$. Then proceed iteratively, beginning with $\gamma_0 = 1$. At any stage of the iteration we have current values for $\gamma_0, \gamma_1, \ldots, \gamma_{r-1}$ and $\lambda$. We now obtain a new value for $\gamma_k$ ($0 < k < r$) by substituting the current values of the other unknowns everywhere into (21) except where $\gamma_k$ occurs linearly. The equation for $\gamma_0$ is (20).
An equation for $\lambda$ is obtainable by multiplying (21) by $\gamma_k$ and summing. This gives

$$\lambda = N - \Sigma (4\alpha + \frac{3}{2}\beta(k + 1)) k\gamma_k^2 + \Sigma (4\alpha + \beta(2k + 3)) [(k + 1)(k + 2)] \frac{1}{4} \gamma_{k+2}$$

$$- \frac{1}{4} \beta \Sigma ((k + 1)(k + 2)(k + 3)(k + 4)) \frac{1}{4} \gamma_{k+4}$$

$$= N + 2\alpha + \frac{3}{2}\beta - 4\alpha \int \gamma'^dx - \beta \int \gamma'^2 dx.$$  \hspace{1cm} (22)

The $r + 1$ unknowns are computed in turn, cyclically, and the run is stopped when the same value is obtained for $\lambda$ to five significant figures. We tried $r = 20, 50, 100$ and 200. In most cases there was no change in the $\gamma_k$, to three places of decimals, in going from $r = 20$ to $r = 50$. Three samples were tried with $r = 50$ and $r = 100$ and the results were identical, with $\gamma_k = 0.000$, for $i > 50$. We therefore regard the calculations as robust, and we used $r = 50$ in most of the runs. As an example of the values of the coefficients we take those for Fig. 4 with $N = 20$. The $\gamma_{2k-1}$ all vanish and the $\gamma_{2k}$ were $0.986, -0.058, -0.124, -0.080, -0.040, 0.022, -0.011, 0.005, -0.002$ and all others were less than 0.0005 in absolute value.

This iterative procedure converged in all the examples tried, with a variety of choices for $\alpha$ and $\beta$.

7. Numerical examples and graphs

A large number of graphs have been prepared, but to save space many fewer have been given than would really have been desirable. For the reasons given in Appendix D the pairs $(\alpha, \beta) = (\frac{1}{4}, 0), (\frac{1}{4}, \frac{1}{4})$ and $(0, \frac{1}{4})$ were emphasized in the experiments.

In all the graphs the observations have been subjected to a linear transformation to achieve zero mean and unbiased sample variance $\frac{1}{N}$. In Fig. 1, $N = 3$ and $\beta = 0$ with $\alpha = \frac{1}{4}, \frac{1}{4}, 1$ and 2. All four graphs are too bumpy. The bumps could perhaps be ironed out by increasing $\alpha$, but then the variance becomes too large in our opinion, although this objection is not yet decisive.

Figs. 2a, 3a and others not given here show again that $\alpha = \frac{1}{4}$ and $\beta = 0$ is unsatisfactory, although $\frac{1}{4}$ was the value of $\alpha$ preferred $a$ priori after standardization when $\beta = 0$. Apparently smoothness is not captured by the first derivative alone.

Figs. 2-5 show that $\alpha = 0$ and $\beta = \frac{1}{4}$ gives satisfactory smoothings in all examples tried. It is very much better than $(\frac{1}{4}, 0)$ and appears to be slightly better than $(\frac{1}{4}, \frac{1}{4})$. There is little to choose between the pairs $(0, \frac{1}{4}), (0, \frac{1}{4})$ and $(0, \frac{1}{4})$ and, on slight evidence, $(0, \frac{1}{4})$, except that the former was preferred $a$ priori. A scarcely perceptible bump in one of the graphs not given here, for $N = 6$, was the only graphical evidence against $(0, \frac{1}{4})$. If $(0, \frac{1}{4})$ is optimal it is to be expected that values of $(\alpha, \beta)$ that are close to it would be almost as good. The run for $\alpha = 0$, $\beta = \frac{1}{4}$ and $N = 3$ took 42 cycles of the iteration to converge whereas $\alpha = 0, \beta = \frac{1}{4}$ and $N = 3$ took only 11 cycles. This at least is a pragmatic argument against $\alpha = 0$ and $\beta = \frac{1}{4}$.

In case $\beta$ should be smaller than $\frac{1}{4}$, we tried $\alpha = 0$ and $\beta = \frac{1}{16}$, and we obtained a bimodal curve for $x_1 = -1$ and $x_2 = 1$; a nose cone for $x_1 = -1, x_2 = 0$ and $x_3 = 1$; and a distinct bump on the right for the six observations $(-\frac{3}{4}, -1, -\frac{3}{4}, -\frac{1}{4}, \frac{1}{4}, 1)$. We consider these fits counter-intuitive.

A few runs were done in error with standardization of the sample variance to 1. It is of interest to note that they gave the same results when correctly repeated, with the $\alpha$'s divided by 2 and the $\beta$'s by 4. This is another indication of robustness.

In the example of Fig. 3 with $N = 10$, the graphs with $(\alpha, \beta) = (0, \frac{1}{4}), (0, \frac{1}{4}), (0, \frac{1}{16})$ and...
The data used in Fig. 5 are the same as used by Boneva et al. (1971) in their Example 1. We have assumed that when the frequency in one interval was \( n \) the observations are at the centres of the \( n \) equal subintervals into which the interval could be divided. Bliss (1987,
p. 97) fitted a normal distribution to the histogram and obtained a $\chi^2$ value of 10-33 with 10 degrees of freedom. Our fit is almost identical with Bliss's but is very slightly skewed to the left.

8. COMPARISON OF FITS

It is useful and essential for bump theory to be able to compare different assumptions for $f$ more objectively than by visual inspection alone, when the observations $x_1, \ldots, x_N$ are given.

![Graphs showing comparison of fits](image)

Fig. 3. $N = 10$. Observations at centres of equal probability intervals for standard normal distribution. (a) $\alpha = \frac{1}{2}$, $\beta = 0$; (b) $\alpha = 0$, $\beta = \frac{1}{2}$; (c) $\alpha = 0$, $\beta = \frac{1}{3}$.

When two assumptions both use the same penalty functional the comparison can be made under the Bayesian interpretation. Of course, in this case one of the $f$'s will not be the optimal one. We give the simple theory that is required for the comparison.

Let $f_1 = \gamma_1^2$ and $f_2 = \gamma_2^2$ be regarded as defining two hypotheses $H_1$ and $H_2$, and let $E$ denote the observations. Let the values of $L$ under $H_1$ and $H_2$ be $L_1$ and $L_2$, respectively. Let $O(H_1|H_2|E, \alpha_1, \beta_1)$ denote the final odds of $H_1$ as compared with $H_2$ given $E$ and assuming $\alpha = \alpha_1$ and $\beta = \beta_1$. Then, by (4),

$$
\log O(H_1|H_2|E, \alpha_1, \beta_1) = \omega(\alpha_1, \beta_1; f_1) - \left( \log L_2 - 4\alpha_1 \int \gamma_2^2 dx - \beta_1 \int \gamma_2^6 dx \right). \quad (24)
$$
Of course, this is positive because $f_1$ maximizes $\omega(x_1, \beta_1; f)$. To exemplify this formula we were opportunistic and made use of some of the machine output (Table 1). For brevity we consider only the case $N = 6$, and the comparisons are shown in Table 2.

### Table 1. Some machine output

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>I</th>
<th>log $L$ in cb</th>
<th>Penalty</th>
<th>Score</th>
<th>$\int \gamma^2 , dx$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>18</td>
<td>-264</td>
<td>61</td>
<td>-325</td>
<td>0-362</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{1}{6}$</td>
<td>0</td>
<td>20</td>
<td>-241</td>
<td>46</td>
<td>-287</td>
<td>0-538</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{1}{6}$</td>
<td>0</td>
<td>12</td>
<td>-223</td>
<td>35</td>
<td>-259</td>
<td>0-822</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{1}{6}$</td>
<td>$\frac{1}{4}$</td>
<td>10</td>
<td>-240</td>
<td>23</td>
<td>-264</td>
<td>0-607</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>$\frac{1}{6}$</td>
<td>11</td>
<td>-236</td>
<td>13</td>
<td>-250</td>
<td>0-701</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
<td>14</td>
<td>-244</td>
<td>16</td>
<td>-260</td>
<td>0-698</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>$\frac{1}{6}$</td>
<td>21</td>
<td>-252</td>
<td>20</td>
<td>-272</td>
<td>0-605</td>
</tr>
</tbody>
</table>

The column I gives the number of cycles in the iteration before convergence. The log likelihood, penalty, and score, are in centibans (cb), so that the score is $43-43\omega(\alpha, \beta; f)$.

### Table 2. Final log-odds in centibans for discriminating between pairs of assumed density functions, $N = 6$

<table>
<thead>
<tr>
<th>Hypotheses</th>
<th>(1, 0)</th>
<th>($\frac{1}{6}$, 0)</th>
<th>($\frac{1}{6}$, $\frac{1}{4}$)</th>
<th>(0, $\frac{1}{4}$)</th>
<th>(0, $\frac{1}{4}$)</th>
<th>(0, $\frac{1}{6}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 0)</td>
<td>0</td>
<td>10</td>
<td>63</td>
<td>23</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>($\frac{1}{6}$, 0)</td>
<td>8</td>
<td>0</td>
<td>7</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>($\frac{1}{6}$, $\frac{1}{4}$)</td>
<td>20</td>
<td>0</td>
<td>7</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0, $\frac{1}{4}$)</td>
<td>17</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0, $\frac{1}{6}$)</td>
<td>23</td>
<td>7</td>
<td>9</td>
<td>17</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>(0, $\frac{1}{6}$)</td>
<td>22</td>
<td>7</td>
<td>114</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

The hypotheses are defined by the 'best' density functions corresponding to the values of $(\alpha, \beta)$ shown. The entry in a row labelled $(\alpha_1, \beta_1)$ and column labelled $(\alpha_2, \beta_2)$ is the log-odds in centibans in favour of the best density function corresponding to $(\alpha_1, \beta_1)$ as against the best density function corresponding to $(\alpha_2, \beta_2)$, when the coefficients in the penalty are assumed to be $(\alpha_2, \beta_2)$.

The above argument is not rigorous because the density functions $\exp(-\Phi)$ are improper, in the novel sense that their normalizing constants are infinite, not zero. Intuitively this is because almost all density functions are very rough and have

$$
\exp\left(-\int \gamma^2 \, dx\right) = \exp\left(-\int \gamma^2 \, dx\right) = 0.
$$

The integrals of $\exp(-\int \gamma^2 \, dx)$ over the spheres $\gamma_0^2 + \gamma_1^2 + \ldots + \gamma_{m-1}^2 = 1$, for $m = 4, 8, 12$ and 16, were worked out in an unpublished report by I. J. Good and were found to be $0.625$, $0.0150$, $6.64 \times 10^{-6}$ and $6.27 \times 10^{-8}$, respectively, fairly obviously tending to zero as $m \to \infty$.

The argument can be made more rigorous by imagining the Hermite expansions to be carried out to only a trillion terms which makes the normalizing constants finite. An estimation method with a trillion parameters is nonparametric! Another way to force propriety would be to divide the roughness penalties by $\int \gamma^2 \, dx + 0.01 \int \gamma^2 \, dx$, although this is ad hoc. The normalizing constants of the prior distribution could then be deduced in principle from the theory of the average of the exponential of a homogeneous function on a hypersphere, given in the same report. This is why we wrote $1$ as $\int \gamma^2 \, dx$. 


The need for natural finite normalizing constants is more serious when we wish to compare two different pairs \((\alpha_1, \beta_1)\) and \((\alpha_2, \beta_2)\). It would be convenient to have a ('Type III') initial distribution of \((\alpha, \beta)\), combined with propriety with the (Type II) priors in function space. Failing this, we can only fall back on a strong intuitive feeling that if \(f_1\) beats \(f_2\) easily, assuming \((\alpha_1, \beta_1)\), but \(f_2\) does not beat \(f_1\) easily, assuming \((\alpha_2, \beta_2)\), then we have obtained

Fig. 4. \(N = 20, 50\) and 100. Observations at the centres of equal probability intervals for the density corresponding to equal mixture of two normal distributions, means \(\pm 2/\sqrt{\pi}\) and variance \(\pi\). \(\alpha = 0, \beta = 1/4\). The least bimodal curve is \(N = 20\).

Fig. 5. \(N = 578, \alpha = 0, \beta = 1/4\). Lengths of ears of corn, derived from the histogram of Bliss (1967), who was quoting Emerson and East.

useful evidence in favour of \(f_1\). For example, in Table 2, \((0, \frac{1}{2})\) on its own ground beats \((\frac{1}{2}, 0)\) by 254 cb, a factor of over 300, but loses only 15 cb on its opponent's ground. The pairs \((\frac{1}{2}, \frac{1}{2}), (0, \frac{1}{2}), (0, \frac{1}{2})\) and \((0, \frac{1}{2})\) never had as much as 3 cb scored against them in calculations more extensive than those shown in Table 2, but \((\frac{1}{2}, 0)\) was over 8 bans down in its 'away encounter' with \((0, \frac{1}{2})\) for \(N = 3\), corresponding to log odds of 100,000,000 to 1 against. This method for comparing priors is controversial but we believe in it. Philosophically it is similar to the use of Type II maximum likelihood as described by Good (1967).
Let \( x \) be a random vector with components \((x_1, \ldots, x_p)\). This notation should not lead to confusion because vector observations would be denoted by \( x_1, \ldots, x_N \). The use of superscripts to denote components, as in some treatises on tensors, has other disadvantages, and the usual justification for it is untenable as pointed out by Schrödinger (1950, p. 7).

If \( f \) is a multivariate density of \( x \) we again write \( \gamma = \sqrt{f} \). We note for future reference that Fisher's information matrix, regarding the location of the density hypersurface, when it is slid bodily, can be written in the form, generalizing (12),

\[
F = \left\{ 4 \int \frac{\partial \gamma}{\partial x_i} \frac{\partial \gamma}{\partial x_j} \, dx \right\} \quad (i, j = 1, \ldots, p).
\]  

We should like to find appropriate and natural generalizations of the one-dimensional functionals \( \int \gamma^2 \, dx \) and \( \int \gamma'^2 \, dx \). A desideratum is 'additivity' in a sense now to be explained for the case \( p = 2 \).

Let the roughness penalty for a \( p \)-dimensional density \( f = \gamma^2 \) be denoted by \( \Phi_p(f) \) when \( p > 1 \). Assume that \( \Phi_2 \) is the integral over the plane \(-\infty < x < \infty, -\infty < y < \infty\), of some expression involving \( \gamma \) and its partial derivatives. Now when \( \gamma(x, y) \) is of the form \( \gamma_1(x) \gamma_2(y) \), that is, if \( x \) and \( y \) are assumed to be statistically independent, \( \Phi_2 \) should be the sum of the separate penalties for \( \gamma_1(x) \) and \( \gamma_2(y) \). Expressed in terms of Alonzo Church's \( \lambda \)-calculus, we should have

\[
\Phi_2((\lambda x, \lambda y) \gamma_1(x) \gamma_2(y)) = \Phi(\gamma_1) + \Phi(\gamma_2),
\]

a condition we describe as additivity. The \( \lambda \) in \( \S 6 \) is different! The generalization of this definition to \( p \) dimensions does not need to be spelt out. The condition of additivity is required for logical consistency; otherwise the 'best' density function, assuming independence, would not be the same as that obtained by treating \( x \) and \( y \) separately. The need for this desideratum depends of course on the additivity of the log likelihoods from which the penalties are to be subtracted.

From now on we shall assume the existence and continuity of all derivatives that occur, and also that the densities and all their mentioned derivatives tend to zero at infinity.

For any positive integer \( r \), consider the following functional of \( \gamma(x, y) = \gamma_1(x) \gamma_2(y) \):

\[
\iint \left( \left( \frac{\partial \gamma}{\partial x} \right)^2 + \left( \frac{\partial \gamma}{\partial y} \right)^2 \right) \, dx \, dy = \iint \left[ (\gamma_1')^2 \gamma_2^2 + (\gamma_2')^2 \gamma_1^2 \right] \, dx \, dy
\]

\[= \int (\gamma^2)'^2 \, dx + \int (\gamma^2)'^2 \, dy.
\]

Thus the penalty (13) can be additively generalized in a straightforward way to

\[
\Phi_p(f) = 4a \int (\nabla \gamma)^2 \, dx + \beta \int \sum_{i=1}^p \left( \frac{\partial \gamma}{\partial x_i} \right)^2 \, dx,
\]

where \( dx \) means \( dx_1 \ldots dx_p \) as usual, and the bold integral sign denotes a multiple integral. The first term is \( a \text{tr} (\mathbf{F}) \). By means of some partial integrations we can write \( \Phi_p \) as

\[
\Phi_p(f) = -4a \int \gamma \nabla^2 \gamma \, dx + \beta \int \gamma \sum_{i=1}^p \left( \frac{\partial \gamma}{\partial x_i} \right)^2 \, dx
\]

\[= -4a E(\gamma^{-1} \nabla^2 \gamma) + \beta E \left( \gamma^{-1} \sum_{i} \left( \frac{\partial \gamma}{\partial x_i} \right)^2 \right).
\]
The first term can also be written as

\[ +4\alpha E(y^{-2}(\nabla y)^2). \] (30)

An example of a penalty that is not additive is \(E(\nabla^2 y)^2\).

When the above penalty is used, we would select the origin at the mean \(X\) of the observations and we might apply a linear transformation so as to force the sample covariance matrix to be \((1/d^2)I\). Then Hermite functions in several variables could be used but with \(H_m(x)\) equal simply to \(x_1^{n_1} \cdots x_p^{n_p}\). Alternatively, we could use the generalized Hermite functions defined in Appendix A, with the matrix \(C\) equal to the usual unbiased estimate of the covariance matrix: remember the notational inconsistency between the univariate and multivariate Hermite functions. The numerical devices of § 6 can all be extended to the multivariate problem; see especially Appendix A. The calculations are clearly simplest when \(C\) is a scalar multiple of the identity, and this case might often be adequate in view of the robustness noted at the end of § 7.

An objection that can be brought against the penalty (28), is that the various co-ordinates \(x_1, \ldots, x_p\) should not necessarily be given equal weight, whether they are left untransformed or are transformed as above. For example, if the scales of some of the co-ordinates are changed, the use of the penalty (28) would produce a different 'best' density function after transforming back to the original co-ordinates. Some condition of invariance is therefore desirable and is discussed in § 10.

If a density \(f(x)\) has some continuous components and some discrete ones \(i = (i_1, i_2, \ldots)\), such as will occur in medical diagnostic problems (see Dickey, 1968, who used a kind of window method), we could try a roughness penalty consisting of a linear combination of the entropy (see § 2) and of \(\Phi(x)\)'s corresponding to each value of \(i\). This raises computational problems not yet examined.

10. INVARIANCE

Even in one dimension \((p = 1)\), a transformation of co-ordinates raises problems. A density function cannot remain smooth under all transformations, even when they are nonsingular and differentiable as many times as one requires, a condition that we assume throughout. Belief in smoothness is rational only if the random variable \(x\) is reasonably natural from a physical or from some other point of view. The question arises whether a 'best' invariant density can be defined under all well-behaved transformations of \(x\) or \(x\).

Whittle (1958) pointed out that his method had this invariance property. An invariant method can be achieved in terms of the tensor calculus if a natural fundamental tensor can be introduced into \(x\) space. This can be done, except when ignorance is bliss, see below, in terms of the 'biutility' \(v(x, y)\), defined as the utility of assuming that a random variable has the value \(y\) when the true value is \(x\) (Good, 1968b, 1969, 1971a, b). It seems to be the most natural method that does not go outside the probability and utility structure of a given density-estimation problem.

As explained in Appendix C, when ignorance is not bliss, we can define a fundamental tensor and its matrix by

\[ g_{ij} = \left[ -\frac{\partial^2 v(x, y)}{\partial y_i \partial y_j} \right]_{x=x}, \quad G = \{g_{ij}\}. \] (31)

The matrix \(G\) is positive semidefinite if \(v(x, x) \geq v(x, y)\), and we shall assume it to be positive definite, a fairly small assumption. In particular \(v(x, x) - v(x, y)\) must not be only a fourth-
order infinitesimal when \( y \) is close to \( x \). If \( f \) is any probability density function, \( g_{ij} \) is a covariant tensor of the second order and \( f g^{-1} \) is invariant, where \( g = \det (G) \). In the standard terminology of the tensor calculus, \( f \) is a scalar density (for example, Eddington (1924, p. 111); or Weatherburn (1938, p. 33)). Hence \( \zeta \), defined as \( \gamma g^{-1} \), is a relative scalar of weight \( \frac{1}{2} \).

We can now redefine the penalty (28) in an invariant form. The logic is slightly easier in terms of expectations than in terms of integrals because the expected value of an invariant, unlike its integral, is again an invariant. In these expectations we first replace \( \gamma \) by \( \zeta \). Now the expressions \( \nabla \zeta \) and \( \nabla^2 \zeta \) have well-known tensorial expressions (for example, Weatherburn (1938, pp. 43, 67)), namely \( g^{ij} \zeta_{,i} \zeta_{,j} \) and \( g^{ij} \zeta_{,ij} \), where the commas imply covariant differentiation with respect to all the suffixes following them, and where of course we are using the summation convention. The first covariant derivative of an invariant is equal to its ordinary derivative, but the second one involves a Christoffel three-index symbol. Hence the generalization of the first term of \( \Phi_p(f) \) is \( 4\alpha \) times

\[
E(\zeta^2 g^{ij} \zeta_{,i} \zeta_{,j}) = \int g^{ij} \zeta^2 g^{ij} \zeta_{,i} \zeta_{,j} d\mathbf{x} = \int \zeta_{,i} \zeta_{,j} g^{ij} \sqrt{g} d\mathbf{x}
\]

(32)

\[
= -E(\zeta^{-1} g^{ij} \zeta_{,ij}) = -\int g^{ij} \zeta^{-1} g^{ij} \zeta_{,ij} d\mathbf{x} = - \int \zeta_{,ij} g^{ij} \sqrt{g} d\mathbf{x}.
\]

(33)

The equality of the expressions on the right of (32) and (33) can be verified by what may be called covariant partial integration, the justification of which is Ricci's lemma that \( g_{ij} \) and \( g \) behave like constants in covariant differentiation. For the same reason we have

\[
\zeta_{,i} = g^{-1} \gamma_{,i}, \quad \zeta_{,ij} = g^{-1} \gamma_{,ij},
\]

(34)

where the covariant differentiation of relative scalars and tensors is defined, for example, by Korn & Korn (1961, p. 490). A colon is used here to denote this form of covariant differentiation to distinguish it from the more familiar covariant differentiation of absolute scalars and tensors. The reader might prefer the formulæ expressed in terms of \( \zeta \) instead of \( \gamma \).

The invariant form of the first term of \( \Phi_p(f) \) can be written as

\[
4\alpha \int \gamma_{,i} \gamma_{,j} g^{ij} d\mathbf{x} = -4\alpha \int \gamma \gamma_{,ij} g^{ij} d\mathbf{x}.
\]

(35)

Similarly, the invariant generalization of Fisher's information matrix with respect to location of the slid density hypersurface is

\[
\text{F} = \left\{ 4 E(\zeta^2 \zeta_{,i} \zeta_{,j}) \right\} = \left\{ 4 \int \zeta_{,i} \zeta_{,j} g^{ij} d\mathbf{x} \right\} = \left\{ 4 \int \gamma_{,i} \gamma_{,j} d\mathbf{x} \right\}.
\]

(36)

The previous formula (25) is recovered when \( G \) is the identity matrix; in fact this is always the condition under which the tensorial expressions specialize to those in § 8. Compare the invariant form of entropy in Good (1968b, 1969). The first term of \( \Phi_p(f) \) can be expressed as \( \alpha \text{tr} (FG) \), when \( G \) is a constant matrix.

To generalize the second term of \( \Phi_p \), we note that the fourth covariant derivative \( \zeta_{,ijkl} \) is symmetric in its first two suffixes so that a natural invariant is

\[
\Theta^* = \frac{1}{2} \zeta_{,ijkl} g^{ik} g^{jl} + \frac{1}{2} \zeta_{,ijkl} g^{ij} g^{kl}.
\]

(37)
Hence an invariant form for the second term of $\Phi_p$ is $\beta$ times

$$E(\zeta^{-1}\theta^*) = \int \gamma d\zeta^{-1} \theta^* \, d\mathbf{x}$$

$$= \int \zeta \theta^* g^1 \, d\mathbf{x}$$

$$= \int \gamma \theta \, d\mathbf{x},$$

where

$$\theta = \frac{3}{2} \gamma_{ijkl} g^{ik} g^{jl} + \frac{1}{2} \gamma_{ijkl} g^{ij} g^{kl}. \tag{39}$$

Thus, an invariant form for $\Phi_p(f)$ is

$$\Phi_p(f) = 4\alpha \int \gamma_{ijkl} g^{ij} d\mathbf{x} + \beta \int \gamma \theta \, d\mathbf{x} \tag{40}$$

$$= -4\alpha \int \gamma_{ijkl} g^{ij} d\mathbf{x} + \beta \int \gamma \theta \, d\mathbf{x}. \tag{41}$$

Note that the penalty vanishes if $f$ is proportional to $g^1$, so this is the best density when there are no observations, although it might be an improper density. It is not a prior in the usual sense because the distribution of $x$ that we are estimating is assumed to have a physical rather than only a logical or subjective existence; contrast Jeffreys (1946).

**Special cases.** If $v(x, y)$ is of the form $u((x - y)^2)$ with $u'(0) = -\frac{1}{2}$, then $G$ is the identity and the tensorial expressions of § 9 reduce to those of § 8. Again, if $v(x, y)$ is any well-behaved function of a quadratic form in $x - y$, such as the often reasonable form

$$1 - \exp \{-(x - y)' A(x - y)\},$$

with $A$ positive definite, then the fundamental tensor $g_{ij}$ becomes constant. Since it is positive definite, a transformation of co-ordinates can be made so as to reduce the first fundamental form to $dx_1^2 + \ldots + dx_n^2$, that is, the $x$ space is Euclidean, $g = 1$ and $\zeta = \gamma$, and (42) reduces to (28). More generally, the $x$ space can be made Euclidean, or in tensor terminology it already is Euclidean, if the Riemann–Christoffel curvature tensor vanishes (Eddington, 1924, p. 80; Weatherburn, 1938, p. 117).

In one dimension ($p = 1$), the Riemann–Christoffel tensor always vanishes: a one-dimensional manifold has no intrinsic curvature. Note that this is not to say that the density function has no curvature when regarded as embedded in the two-dimensional space in which $x$ is only one of the co-ordinates. A transformation that converts the metric $ds^2 = g_{11} \, dx^2$ into $d\zeta^2$ is clearly

$$\zeta = \int_0^x \left[ \frac{g_{11}(x, y)}{\gamma} \right]^{\frac{1}{2}} \, dz. \tag{43}$$

**11. CONTRAST WITH QUANTUM MECHANICS**

Since we have used some mathematical techniques that occur in quantum mechanics, it might be thought that our methods could be applied to the smoothing of probability densities of measurements in quantum-mechanical experiments with unknown wave functions but whose states have been prepared in the sense defined, for example, by Ballentine (1970, p. 366). But these densities are usually flamboyant because they arise as
the squares of absolute values of complex wave functions, whereas our functions \( \gamma(x) \) are real. Our methods could, without modification, be applied to such problems but the samples would have to be large enough to pick up the oscillations. Alternatively \( \gamma \) could be replaced by the unknown complex wave function \( \psi \), where \( |\psi|^2 = f \). We could then use the expansions

\[
\psi(x) = \sum_i (a_i + ib_i) \phi_i(x), \quad |\psi(x)|^2 = \sum_i (a_i a_i + b_i b_i) \phi_i^* \phi_i,
\]

and extend the methods of §§ 6 and 9 in the obvious way for estimating the coefficients. Here the \( \phi_i \) could be either the Hermite functions or the eigenfunctions of the specific Schrödinger equation. It might be worth noting that \( \nabla^2 \phi_i \) can be rewritten by Schrödinger’s equation.

**Appendix A.** Proof of equations (17) and (18), and multivariate extensions

We have

\[
\int \gamma'^2 dx = \sum_{r,s} \gamma_r \gamma_s \int \phi_r' \phi_s' dx
\]

\[
= \sum_{r,s} \gamma_r \gamma_s \int \left( (2r_1^2 \phi_{r_1} - x \phi_{r_1}) (2s_1^2 \phi_{s_1} - x \phi_{s_1}) \right) dx
\]

and (17) quickly follows from the orthogonality of \{\phi_m\} combined with (65) of Appendix E, and (Margenau & Murphy, 1943, p. 361)

\[
\int x^2 \phi_m(x) \phi_n(x) dx = \left( \frac{1}{2} + m \right) \delta_m^n + \frac{1}{2} \left( m(m-1) \right)^{\frac{1}{2}} \delta_m^{n+2} + \frac{1}{2} \left( n(n-1) \right)^{\frac{1}{2}} \delta_{m+2}^n.
\]

A proof of equation (18) along these lines occupies several pages, but, as we noted, there is a short-cut that uses some of the formalism of quantum mechanics which we now describe. There is a similar short-cut for the above proof, but it does not save much work in this case and the above calculation will serve as a check.

The right side of (18) is a quadratic form in an infinite number of variables with a matrix \( Q \) whose elements are

\[
\int \phi_r^* \phi_s^* dx = \int \phi_r \left( i \frac{\partial}{\partial x} \right)^4 \phi_s dx.
\]

In accordance with the equivalence of Schrödinger’s and Heisenberg’s formulations of the quantum mechanics of the harmonic oscillator (Schrödinger, 1926), \( Q \) must be the fourth power of the Hermitian momentum matrix

\[
\begin{bmatrix}
0 & 1 & 0 & 0 & \ldots \\
-1 & 0 & \sqrt{2} & 0 & \ldots \\
0 & -\sqrt{2} & 0 & \sqrt{3} & \ldots \\
0 & 0 & -\sqrt{3} & 0 & \ldots \\
\end{bmatrix}
\]

and can be quickly calculated.

For completeness and for convenience in generalization, we here give a proof of the required identities. We wish to prove that, for each \( m \geq 0 \), the matrix

\[
\left\{ \int \phi_r \left( \frac{d}{dx} \right)^m \phi_s dx \right\} = \left( \int \phi_r \phi_s^* dx \right)^m,
\]

where \( r \) and \( s \) index the rows and columns of the matrices, i.e. the matrix on the left hand side is expressed as the \( m \)th power of a matrix. For \( m = 0 \) this merely expresses the normal
orthogonality of the Hermite functions. To prove (48) it is sufficient to prove that for each pair of positive integers \( m \) and \( n \) we have

\[
\int \phi_\epsilon \left( \frac{dx}{dx} \right)^{m+n} \phi_\epsilon \, dx = \sum_{t=0}^{\infty} \int \phi_\epsilon(x) \left( \frac{dx}{dx} \right)^{m} \phi_\epsilon(x) \, dx \int \phi_\epsilon(y) \left( \frac{dy}{dy} \right)^{n} \phi_\epsilon(y) \, dy.
\]

(49)

By means of a number of partial integrations, together with some inversions of the orders of integration and summation, we can write the required identity in the form

\[
\int d^m \phi_{\epsilon} \, d^n \phi_{\epsilon} \, dx = \int \left( \int d^m \phi_{\epsilon}(x) \, d^n \phi_{\epsilon}(y) \, \phi_{\epsilon}(x) \phi_{\epsilon}(y) \, dx \, dy, \right.
\]

(50)

and this is formally correct since physicists opine that

\[
\sum_{t} \phi_{\epsilon}(x) \phi_{\epsilon}(y) = \delta(x - y).
\]

(51)

Equation (51) is an example of the closure property of a complete orthonormal system; see, for example, Messiah (1961, pp. 187, 492).

Formally the identity (51) holds for any complete orthonormal system, not merely for the Hermite functions. For the Hermite functions, the argument can be made rigorous by putting in a convergence factor \( z^t \) (\(|z| < 1\)), making use of Mehler's formula

\[
\sum_{t=0}^{\infty} z^t \phi_{\epsilon}(x) \phi_{\epsilon}(y) = \frac{1}{\pi i (1 - z^2)} \exp \left\{ \frac{4z^2 - (z^2 + y^2)(1 + z^2)}{2(1 - z^2)} \right\},
\]

(52)

see, for example, Erdélyi et al. (1953, p. 194); Wiener (1933, p. 62); Watson (1933), and finally letting \( z \to 1 - 0 \).

Rigour can also be attained by using the Christoffel–Darboux formula (see Appendix F), but the use of Mehler's formula has the advantage that it can be generalized to the multivariate case; see, for example, Erdélyi (1938), who gives further references to Koschmieder.

In this appendix we shall use the following multivariate notations:

\[
m = (m_1, m_2, ..., m_p)', \quad |m| = m_1 + ... + m_p, \quad m! = m_1! m_2! ... m_p!, \quad \delta_{mn} = \delta_{m_1 n_1} \cdots \delta_{m_p n_p}, \quad a^b = a_1^b_1 \cdots a_p^b_p,
\]

where \( a \) is a vector. The same convention applies if \( a \) is replaced by a vector matrix defined as \( A = (A_1, ..., A_p) \), where \( A_1, ..., A_p \) are commutative matrices. The matrix \( A \) can be regarded as a vector whose components are matrices and also as a matrix whose elements are vectors. The convention also applies to differentiation, that is,

\[
\nabla^m = \left( \frac{d}{dx} \right)^m = \left( \frac{\partial}{\partial x_1} \right)^{m_1} \cdots \left( \frac{\partial}{\partial x_p} \right)^{m_p}.
\]

The definitions of the multivariate Hermite polynomials are (see, for example, Erdélyi et al. (1953, p. 285))

\[
H_t(x) = H_t(x|C) = (-1)^t \exp \left( \frac{1}{2} x' C x \right) \left( \frac{d}{dx} \right)^{t} \exp \left( -\frac{1}{2} x' C x \right),
\]

(53)

\[
G_t(C^{-1}x) = G_t(C^{-1}x|C) = H_t(x|C^{-1}),
\]

(54)

where \( C = (c_{ij}) \) is a positive definite real \( p \times p \) matrix. The factor \( \frac{1}{2} \) in the index of the exponential is a convention differing from our convention for the case \( p = 1 \). Then the functions \( \phi_t(x) \) and \( x_t(x) \) defined by

\[
\frac{\phi_t(x)}{H_t(x)} = \frac{x_t(x)}{G_t(x)} = \frac{\exp \left( -\frac{1}{2} x' C x \right)}{(2\pi)^{p/2} ||C||^{1/2}(t!)}
\]

(55)
form a normal biorthogonal system, that is,
\[ \int \phi_i(x) \chi_j(x) \, dx = \delta_{ij}. \] (56)

The multivariate form of (48) is
\[ \left( \int \phi_i \nabla^m \chi_i \, dx \right) = \left( \int \phi_j \nabla \chi_j \, dx \right)^m, \] (57)
where the right side is the \( m \)th power of a vector matrix. It has a definite meaning because the relevant scalar matrices commute. This corresponds to the quantum-mechanical fact that components of momentum commute, and it also follows from the proof of (57) along the lines of that of the scalar case, rigour being provided by the Koschmieder–Erdélyi generalization of Mehler's formula.

The univariate argument of Margenau & Murphy (1943, pp. 120–121) can be conveniently generalized by using the multivariate notation. After some manipulation, we obtain
\[ \int x_i \phi_i(x) \chi_i(x) \, dx = \delta_{i}^{t+n} t_1 + \delta_{i+t+n} e_i \] (58)
where \( e_i \) is the vector whose components are all 0 except the \( i \)th which is 1. This generalizes (65) of Appendix E. But clearly
\[ \frac{\partial \phi_i}{\partial x_i} = \frac{i}{2} \phi_i \sum_j c_{ij} x_j - (t_1 + 1)i \phi_{i+t_1}, \] (59)
so the skew-symmetric matrix
\[ \left( \int \phi_i \frac{\partial}{\partial x_i} \chi_i \, dx \right) = - \left( \int \chi_i \frac{\partial}{\partial x_i} \phi_i \, dx \right) \] (60)
can be evaluated. Then (57) can be used to evaluate the penalty (28) by first writing
\[ \gamma = \sum_i \gamma_i \phi_i = \sum_i \delta_i \chi_i. \]

The calculations could be checked by noting that
\[ \left( \int x_0 \phi_0(x) \chi_0(x) \, dx \right) = \left( \int x \phi_0(x) \chi_0(x) \, dx \right)^m = M_m \] (61)
say, which generalizes Margenau & Murphy (1943, p. 361), combined with repeated use of (59). Here (61) is the position analogue of the momentum identity (57). The characteristic function of \( f \) is \( \gamma' e^{i \mathbf{a} \cdot \mathbf{s}} \), where \( \gamma \) and \( \mathbf{s} \) are the vectors with components \( y_t \) and \( \delta_t \).

**APPENDIX B. Example of a hand solution of (20) and (21)**

In equations (20) and (21) with \( \alpha = \frac{1}{4} \) and \( \beta = 0 \), take \( N = 1, x_1 = 0 \) and \( \gamma_k = 0 \) for \( k \geq 4 \). Then the equations become
\[ \frac{2 \phi_k(0)}{3} + \sum_{i=0}^{k} \gamma_{k+i} \psi_{k+i} = (2k + 2) \gamma_k \quad (k = 0, 1, 2, 3). \]

These equations are satisfied by taking \( \lambda = 1, \gamma_0 = 1 \) and \( \gamma_1 = \gamma_2 = \gamma_3 = 0 \) because \( \phi_0(0) = n^{-\frac{1}{4}}, \phi_x(0) = -n^{-\frac{1}{4}}, \) and \( \phi_2(0) = \phi_3(0) = 0. \)
APPENDIX C. Invariance of $f/\sqrt{g}$

The reader should look at §10 and compare it with Good (1968a, 1969).

If a nonsingular transformation $x = h(\xi)$ and $y = h(\eta)$ is made, where $h$ is some twice differentiable vector function, the same for both $x$ and $y$, then of course

$$v(x, y) = v(h(\xi), h(\eta)) = v^*(\xi, \eta),$$

where $v^*$ refers to the new co-ordinates. We have

$$\frac{\partial v^*(\xi, \eta)}{\partial \eta_i} = \sum_\mu \left( \frac{\partial v(x, y)}{\partial x_\mu} \frac{\partial x_\mu}{\partial \eta_i} + \frac{\partial v(x, y)}{\partial y_\mu} \frac{\partial y_\mu}{\partial \eta_i} \right) = \sum_\mu \frac{\partial v(x, y)}{\partial y_\mu} \frac{\partial y_\mu}{\partial \eta_i}.$$

At $y = x$, the second summation vanishes because $v(x, y)$, as a function of $y$, has a maximum at $y = x$ when truth is the best policy. Hence

$$\left[ -\frac{\partial^2 v(x, y)}{\partial y_i \partial y_j} \right]_{y=x}$$

is a covariant tensor of rank 2. The sign in Good (1968a) was misprinted. It also follows that

$$g(x) \left( \frac{\partial x}{\partial \xi} \right)^2 = g^*(\xi),$$

where $\partial x/\partial \xi$ denotes a Jacobian. If $g(x)$ never vanishes, then $\{g_{ij}\}$ is positive definite and $f(x)/[g(x)]^{1/2}$ is invariant: recall that $f$ is a density function so that $f(x) \partial x/\partial \xi = f^*(\xi)$.

APPENDIX D. An intuitive constraint on $\alpha$ and $\beta$

An intuitively reasonable constraint on $\alpha$ and $\beta$ will be derived here, namely

$$2\alpha \sigma^2 + \frac{1}{2} \beta = \sigma^4, \quad (62)$$

where $\sigma^2$ is either an initially guessed value of the variance or an estimate from the sample $x_1, \ldots, x_N$. In our numerical work we standardized the sample variance to $\frac{1}{2}$ so that (62) could be written

$$4\alpha + 3\beta = 1. \quad (63)$$

Some pairs of values that satisfy (63) are (\frac{1}{4}, 0), (0, \frac{1}{6}) and (\frac{1}{2}, \frac{1}{6}), and these were emphasized in our numerical work. Of all pairs of values for $\alpha$ and $\beta$ that satisfy the constraint, the simplest are the first two of these pairs, and (0, $\frac{1}{6}$) gave good results in all our computer runs.

The basis for the constraint (62) is the feeling that the normal distributions form the smoothest class of distributions, the improper uniform distribution being a limiting form. Perhaps some justification for this feeling is that a normal distribution is the distribution of maximum entropy for a given mean and variance (Shannon, 1948, p. 629). This is not convincing because the result is different when more moments are given (Good, 1963, p. 915). The integral $\int \gamma^n dx$ is also minimized for a given variance when $f$ is normal; see Appendix E. In any case it seems reasonable to give the normal distribution special consideration. We therefore prefer $\alpha$ and $\beta$ to be such that, when $f$ is normal, the score $\omega(\alpha, \beta; f)$ is maximized by taking the mean and variance of $f$ equal to $\overline{x}$ and $\Sigma(x_i - \overline{x})^2/(N - 1)$.
If \( f(x) \sim \mathcal{N}(\mu, \sigma^2) \), we have
\[
\int \gamma'^2 \, dx = 1/(4\sigma^2), \quad \int \gamma'^2 \, dx = 3/(16\sigma^4),
\]
and therefore
\[
\omega(\alpha, \beta; f) = -N \log(\sigma^2 + \beta^2) - \sum_{i=1}^{N} (x_i^2 - \mu^2)/\sigma^2 - \alpha/\sigma^2 - 3\beta/(16\sigma^4).
\]
This is maximized by taking \( \mu = \bar{x} = \Sigma x_i/N \) and \( \sigma \) such that
\[
-N/\sigma + \Sigma (x_i - \bar{x})^2/\sigma^3 + 2\alpha/\sigma^3 + 3\beta/(4\sigma^5) = 0.
\]
This implies the constraint (62) if we put \( \sigma^2 = \Sigma (x_i - \bar{x})^2/(N - 1). \)

**Appendix E. A penalty for large variance and some lower bounds for the roughness penalty**

If we have some initial impression of how large the variance of our distribution should be and also if we estimate \( \sigma^2 \) from the sample, we might wish to take a penalty for any hypothesis with too large or too small a variance. Note that the variance can be expressed in terms of the \( \gamma \)'s, by using equation (45) together with the simpler formula
\[
E(x) = \frac{1}{2} \sum_{m=1}^{n} \gamma_m^2.
\]
We obtain
\[
\text{var}(x) = \frac{1}{2} \left[ \sum_{m=1}^{n} \gamma_m^2 + \sum_{m=1}^{n} \gamma_{m+1}^2 \right] - 2 \left\{ \gamma_m \gamma_{m+1}(m + 1) \right\}^2.
\]
These formulae are of interest apart from the possible use of (67) for a new roughness penalty. We should have printed these values in our computer runs. They can be generalized to \( p \) dimensions by using (61).

If the roughness penalty were taken proportional to \( \text{var}(x) \) the effect would be that \( f \) would vanish outside the range of the observations and would therefore be discontinuous. This would be unreasonable, but it might be worth while to consider using this penalty, or one proportional to the generalized variance in \( p \) dimensions, as an addition to the penalties considered in the main text. This would enable our dislike of large variance, which we used in §7, to be taken into account quantitatively. An invariant form of this penalty could perhaps be devised by regarding Riemannian co-ordinates (Weatherburn, 1938, p. 79) with the pole at \( \bar{x} \) as the most natural; but \( \bar{x} \) depends on the co-ordinates used so it would have to be found iteratively and might not be unique.

For \( p = 1 \), a penalty for small variances, and other moments, is already implicit in (13). For if we assume \( \gamma, \gamma', \gamma'', \ldots \) tend to 0 at \( \infty \), and write
\[
\mu_r' = \int x^r f(x) \, dx \quad (r = 0, 1, 2, \ldots),
\]
we can prove that
\[
\int \gamma'^2 \, dx \geq \frac{r^2 \mu_r^2}{4 \mu'_r}, \quad \int \gamma'^2 \, dx \geq \frac{r^2 \mu_r^2}{16 \mu'_r^2},
\]
and in particular
\[
\int \gamma'^2 \, dx \geq \frac{1}{4 \mu'_3}, \quad \int \gamma'^2 \, dx \geq \frac{1}{16 \mu'_4}.
\]
The first of the inequalities (69) is essentially equivalent to a form of the uncertainty principle; see for example, Weyl (1931, p. 393). It becomes an equality when \( f \) is normal. All these inequalities can be proved by using partial integrations and Schwartz's inequality.

The corresponding multivariate inequalities are

\[
4 \int (\nabla \gamma)^2 \, dx \geq (\mu'_{1200\ldots})^{-2} + (\mu'_{0120\ldots})^{-2} + \ldots,
\]

\[
16 \sum \left[ \frac{\partial^2 \gamma}{\partial x^2} \right] \, dx \geq (\mu'_{1200\ldots})^{-4} + (\mu'_{0120\ldots})^{-4} + \ldots.
\]

**APPENDIX F. Sufficient conditions for \( \gamma(x) = \Sigma \gamma_m \phi_m(x) \) and for its differentiation**

We have (for example, Erdélyi et al. (1953, p. 193) and Jackson (1941, p. 158)),

\[
\sum_{r=0}^{n} \phi_r(x) \psi_r(y) = \psi_n(x, y),
\]

where

\[
\psi_n(x, y) = \frac{\phi_{n+1}(x) \phi_n(y) - \phi_{n+1}(y) \phi_n(x)}{x - y}
\]

This is a special case of the Christoffel–Darboux identity (Szegö, 1959, p. 43). Therefore

\[
\sum_{r=0}^{n} \gamma_r \phi_r(x) = \int_{-\infty}^{\infty} \gamma(y) \sum_{r=0}^{n} \phi_r(x) \phi_r(y) \, dy = \int_{-\infty}^{\infty} \gamma(y) \psi_n(x, y) \, dy.
\]

Let us assume that the derivatives \( \gamma'(x), \gamma''(x), \ldots \), as far as the \( s \)th derivative, tend to 0 when \( x \to \pm \infty \), and that \( \gamma \) and each relevant derivative is of bounded variation in each finite interval and is absolutely integrable in \( (-\infty, \infty) \). We can then differentiate (74) \( s \) times, and after \( s \) partial integrations, we obtain

\[
\sum_{r=0}^{n} \gamma_r \phi_r^{(s)}(x) = (-1)^s \int_{-\infty}^{\infty} \gamma^{(s)}(y) \psi_n(x, y) \, dy \quad (s = 0, 1, 2, \ldots).
\]

Now (Plancherel & Rotach, 1929) if \( x \neq 0 \) and \( n \) is large,

\[
\phi_n(x) = \frac{2i}{n^{1/2}} \left[ \cos \left\{ \frac{1}{2} (n+1) \pi - x(2n)^{1/2} \right\} + O(n^{-1}) \right].
\]

Therefore

\[
\sum_{r=0}^{n} \gamma_r \phi_r^{(s)}(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\gamma^{(s)}(y) \sin \left\{ (x-y) (2n)^{1/2} \right\}}{x-y} \, dy + O \left( \frac{1}{\sqrt{n}} \right);
\]

see, for \( s = 0 \), Szegö (1959, p. 245). By the Riemann–Lebesgue theorem, when \( n \to \infty \) the integral depends only on values of \( \gamma^{(s)}(y) \), where \( y \) is in the neighbourhood of \( x \). It follows that

\[
\sum_{r=0}^{n} \gamma_r \phi_r^{(s)}(x) = \gamma^{(s)}(x) \quad (s = 0, 1, \ldots).
\]

When \( x = 0 \), one could base the proof on the asymptotic formula (2) of Erdélyi et al. (1953, p. 199).

For the multivariate case this argument does not generalize, but the series (78) can be
shown to be summable in Abel's sense (Hardy, 1949) by using Mehler's identity (52) and this result can be generalized by using the generalization of Mehler's identity mentioned in Appendix A. The multivariate notation $z^m$ gives the appropriate generalization of Abel summability.

References


Nonparametric roughness penalties for probability densities


[Received November 1970. Revised February 1971]

Some key words: Density estimation; Hermite functions; Multivariate density estimation; Nonparametric density estimation; Quantum-mechanical techniques; Roughness penalties; Smoothness; Tensors; Vector matrices.
学霸图书馆
www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，
提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。

图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：
图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具