Ann. Inst. Statist. Math.
36 (1984), Part B, 523-538

A BAYESIAN APPROACH TO THE PROBABILITY DENSITY ESTIMATION

MAKIO ISHIGURO AND YOSIYUKI SAKAMOTO

(Received Sept. 20, 1983; revised Aug. 2, 1984)

Summary

A Bayesian procedure for the probability density estimation is proposed. The procedure is based on the multinomial logit transformations of the parameters of a finely segmented histogram model. The smoothness of the estimated density is guaranteed by the introduction of a prior distribution of the parameters. The estimates of the parameters are defined as the mode of the posterior distribution. The prior distribution has several adjustable parameters (hyper-parameters), whose values are chosen so that ABIC (Akaike's Bayesian Information Criterion) is minimized.

The basic procedure is developed under the assumption that the density is defined on a bounded interval. The handling of the general case where the support of the density function is not necessarily bounded is also discussed. The practical usefulness of the procedure is demonstrated by numerical examples.

1. Introduction

The density estimation problem has been attracting the attention of many researchers at least for three decades. Rosenblatt [14], Whittle [15] and Parzen [13] proposed the use of 'window method' to obtain a smooth estimate. The main difficulty with this approach lies in the choice of the 'kernel', or the degree of smoothness, and any satisfactory criterion for the choice has never been proposed. Good and Gaskins [8], [9] developed the 'penalized likelihood' method for 'nonparametric' density estimation. It is misleading that they called their procedure 'nonparametric'. It really is a 'parametric' procedure. And, we think, therefore their argument on the choice of the parametrization is insufficient. Another point to be noted on their procedure is that it does not provide a clear-cut criterion for the choice of the penalty, or more precisely, the balancing of the penalty with the likelihood. Leonard [11]

carried through his nonparametric approach. However, his procedure is very much alike to that of Good and Gaskins, and suffers from the same difficulty of the penalty-likelihood balancing problem. Both of Good-Gaskins procedure and Leonard procedure are based on the Bayesian formulation of the preference for the smoother estimate. It seems that Bayesian approach is reasonable. However, when it is combined with nonparametric approach, various difficulties arise.

If we classify those methods into 'Bayesian-nonparametric' approach, the method proposed by Akaike [3] is to be classified into a 'non-Bayesian-parametric' approach. This procedure is implemented as the program GALTHY by Akaike and Arahata [4], and GALTHY is so far the only program which realizes a fully automated density estimation. A sensible part with their procedure is the adoption of the transformation $y=G(x)$ of the original data x, where G is a suitably chosen distribution function. By this device, they could convert any density estimation problem into a relatively easy problem of density estimation on a closed interval. Let $h^*(y)$ be the density function of the transformed data y on the finite closed interval $0 \leqq y \leqq 1$, then the density function of x is given by

$$f(x)=h^*(G(x))g(x)\ , \tag{1}$$

where $g(x)$ is the density function corresponding to the distribution function G. In GALTHY, the following parametrization for $h^*(y)$

$$\log h_K(y)=\sum_{k=0}^{K} a_k y^k\ . \tag{2}$$

is adopted. For each K, the order of the model, estimates of parameters are obtained by maximizing the likelihood. A weighted average of the maximum likelihood estimates of parametric models with a weight proportional to $\exp\{(-1/2)\,\mathrm{AIC}\}$ for each model is used as the final estimate of the density function.

GALTHY is a good example of the effectiveness of the entropy maximization principle which is proposed by Akaike [3] as an extension of the maximum likelihood principle. As a practical procedure, however, GALTHY sometimes produces estimates with siderobe-like spurious peaks. We consider that this is due to the parametrization adopted in GALTHY.

The purpose of the present paper is to propose a Bayesian approach that rectifies this limitation. Our procedure is realized as an extension of the Bayesian approach to the binary response curve estimation developed by the present authors [10] and works efficiently for any data set provided that the true density function is properly smooth. Our approach should be classified into a 'Bayesian-parametric' approach.

2. Bayesian model

2.1. *Histogram model*

The basic assumption here is that $h^*(y)$ is well approximated by a piecewise constant function, or a histogram, defined by

$$(3) \qquad h(y)=m \exp(h_j/m) \Big/ \sum_{k=1}^{m} \exp(h_k/m) \quad \text{for } a_{j-1} \leqq y < a_j \quad j=1, \cdots, m-1, \quad a_{j-1} \leqq y \leqq a_j \quad j=m$$

where $\{a_j\}$ are defined by

$$(4) \qquad a_j=j/m \qquad (j=0, 1, \cdots, m)$$

and h_j is a set of parameters which satisfy the relation;

$$(5) \qquad h_m=-\sum_{j=1}^{m-1} h_j \ .$$

Note that model (3) is flexible enough if it is possible to set m very large. See Appendix for the derivations of this model.

When a set of data $\{y_i;\ 0 \leqq y_i \leqq 1;\ i=1, \cdots, n\}$ is given, the likelihood of the model (3) as a function of $\boldsymbol{h}=(h_1, h_2, \cdots, h_{m-1})^t$ is given by

$$(6) \qquad L(\boldsymbol{h})=\prod_{i=1}^{n} h(y_i)=\prod_{j=1}^{m} \left\{ \frac{m \exp(h_j/m)}{\sum_{k=1}^{m} \exp(h_k/m)} \right\}^{n(j)}$$

where $n(j)$ is the number of data which satisfies $a_{j-1} \leqq y_i < a_j$ ($a_{m-1} \leqq y_i \leqq a_m$ for $j=m$). The maximum likelihood estimate $\boldsymbol{h}^+$ of $\boldsymbol{h}$ is obtained by maximizing (6) considering the restriction (5). The components of $\boldsymbol{h}^+$ is explicitly given by

$$(7) \qquad h_j^+=\log \left\{ \prod_{k=1}^{m} \frac{n(j)}{n(k)} \right\} \qquad (j=1, 2, \cdots, m) \ .$$

The maximum likelihood estimate $h^+(y)$ of $h^*(y)$ is obtained by replacing h_j's in (3) by corresponding h_j^+'s. However, it is well known that $h^+(y)$ is unstable or noisy when m is large compared to n. The introduction of a Bayesian procedure to solve this difficulty will be the subject of the following sections.

2.2. *Bayesian model*

To obtain a "stable" estimate of $h^*(y)$, we introduce, following Akaike [5], a prior density of the parameter $\boldsymbol{h}$, which is defined by

$$(8)\qquad \pi(\boldsymbol{h}\,|\,v^2, h_{-1}, h_0)=\prod_{j=1}^{m}\frac{1}{\sqrt{2\pi}\,v}\exp\left\{-\frac{1}{2v^2}(h_j-2h_{j-1}+h_{j-2})^2\right\},$$

where v^2, h_0 and h_{-1} are adjustable parameters. Our estimate of $\boldsymbol{h}$ is defined as the mode of posterior density which is proportional to $L(\boldsymbol{h})\cdot\pi(\boldsymbol{h}\,|\,v^2, h_{-1}, h_0)$ and relatively easy to calculate. The values of v^2, h_0 and h_{-1} are chosen so that they maximize the likelihood of a Bayesian model defined by

$$(9)\qquad \int L(\boldsymbol{h})\pi(\boldsymbol{h}\,|\,v^2, h_{-1}, h_0)d\boldsymbol{h}\,,$$

or, equivalently, to minimize

$$(10)\qquad \mathrm{ABIC}=-2\log\int L(\boldsymbol{h})\pi(\boldsymbol{h}\,|\,v^2, h_{-1}, h_0)d\boldsymbol{h}\,.$$

3. Numerical consideration

The maximization and integration of $L(\boldsymbol{h})\pi(\boldsymbol{h}\,|\,v^2, h_{-1}, h_0)$ are the numerically most difficult part with our approach. These are almost the same problem that we had to solve in Ishiguro and Sakamoto [10], and are likewise solved as follows.

The maximization is carried out by Newton method, which is very efficient for the present situation because of the negative definitness of the Hessian of the log-likelihood (see Appendix). Since the exact evaluation of the integration in (10) is practically impossible, we approximate it by

$$(11)\qquad \int \exp\{T(\boldsymbol{h})\}d\boldsymbol{h}\,,$$

where $T(\boldsymbol{h})$ is the Taylor expansion of $\log\{L(\boldsymbol{h})\pi(\boldsymbol{h}\,|\,v^2, h_{-1}, h_0)\}$ up to the second order term around its maximizing point, say $\hat{\boldsymbol{h}}$. When v^2 is sufficiently small, i.e., $\pi(\boldsymbol{h}\,|\,v^2, h_{-1}, h_0)$ is sharp compared to $L(\boldsymbol{h})$, the approximation will be good enough to practical purposes.

Now we will derive an approximation of ABIC based on the expression (11). Since $T(\boldsymbol{h})$ attains its maximum at $\hat{\boldsymbol{h}}$, the first order term vanishes and $T(\boldsymbol{h})$ takes the form

$$(12)\qquad T(\boldsymbol{h})=T(\hat{\boldsymbol{h}})-\frac{1}{2}(\boldsymbol{h}-\hat{\boldsymbol{h}})'H(\hat{\boldsymbol{h}})(h-\hat{\boldsymbol{h}})\,,$$

where $T(\hat{\boldsymbol{h}})$ is defined by

$$(13)\qquad T(\hat{\boldsymbol{h}})=\log L(\hat{\boldsymbol{h}})\pi(\hat{\boldsymbol{h}}\,|\,v^2, h_{-1}, h_0)$$

and $H(\hat{\boldsymbol{h}})$ can be decomposed into the form

$$H(\hat{\boldsymbol{h}})=\hat{G}+\frac{1}{v^2}D^tD\ . \tag{14}$$

Here D is a matrix defined by

$$D=\begin{pmatrix} 1 & & & & & 0 & \cdots & 0 \\ -2 & 1 & & & & \ddots & & \vdots \\ 1 & -2 & 1 & & & & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & & & & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & & & \\ \vdots & & \ddots & \ddots & \ddots & \ddots & & \\ \vdots & & & \ddots & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & 1 & -2 & & 1 & \\ 1 & \cdots & \cdots & 1 & 0 & & 3 & \end{pmatrix} \tag{15}$$

and (j, k) element of the matrix $\hat{G}$ is given by

$$\hat{G}_{jk}=\begin{cases} \dfrac{n}{m^2}\{\hat{p}_m-(\hat{p}_j-\hat{p}_m)(\hat{p}_k-\hat{p}_m)\} & \text{if } j\neq k \\ \dfrac{n}{m^2}\{\hat{p}_j+\hat{p}_m-(\hat{p}_j-\hat{p}_m)^2\} & \text{if } j=k \end{cases} \tag{16}$$

where $\hat{p}_j$ is defined by

$$\hat{p}_j=\frac{\exp(\hat{h}_j/m)}{\sum_{k=1}^{m}\exp(\hat{h}_k/m)}\ . \tag{17}$$

The approximate value of ABIC derived from (10), (11) and (12) is

$$\mathrm{ABIC}=-2T(\hat{\boldsymbol{h}})+\log\{\det H(\hat{\boldsymbol{h}})\}\ . \tag{18}$$

We can choose h_{-1}, h_0 and v_2 by minimizing (18) instead of (10). The situation is further simplified by the fact that the effect of h_{-1} and h_0 on ABIC (18) is almost solely through the first term. It means that when the value of v^2 is fixed, it is harmless to choose h_{-1} and h_0 simultaneously with $\boldsymbol{h}$, by maximizing

$$\log\{L(\boldsymbol{h})\pi(\boldsymbol{h}\,|\,v^2, h_{-1}, h_0)\}\ . \tag{19}$$

The optimization with respect to v^2 remains non-linear but the density estimation is not so sensitive to the choice of v^2, so we can adopt the grid search technique here.

4. Estimation procedure for a general data set

Our procedure can be applied to any kind of data $x_1, \cdots, x_n$ if the mapping $y_i = G(x_i)$ into the closed interval [0, 1] is suitably chosen. Once an estimate of $h(y)$ is obtained, the estimate of $f(x)$ is readily calculated using the expression (1). This often produces, however, an unattractive result that the final estimate $f(x)$ is saw-toothed. A cure for this difficulty is the adoption of a higher order spline (second order, for example) function as a model of $h^*(y)$. Such a procedure, however, needs more complicated calculation techniques, so we adopt the following procedure here.

Step 1 To transform a given set of data $x_1, \cdots, x_n$ by $y = G(x)$

Step 2 To find the Bayesian estimate $\hat{h}(y)$ of y

Step 3 To obtain a second order spline estimate $\tilde{h}(y)$ substituting the basis of 2nd order B-spline for the corresponding 0th order B-spline basis. Here we keep the coefficients intact

Step 4 To estimate the probability density function of x by

$$f(x) = \tilde{h}(G(x))g(x) .$$

In Step 1 we have to choose $G(x)$. If the data contain some negative values, we adopt as $G(x)$ the normal distribution with the parameters estimated from the data. If all the data are positive, the normal distribution, the log-normal distribution or the exponential distribution are conceivable as $G(x)$. In such cases, the choice of $G(x)$ is realized by minimum AIC procedure by Akaike [2]. AIC's for those three models are respectively given by

$$\text{AIC}\,(1) = n \log 2\pi + n \log \hat{\sigma}^2 + n + 4$$

$$\text{AIC}\,(2) = n \log 2\pi + n \log \hat{\theta}_2^2 + 2n\hat{\theta}_1 + n + 4$$

$$\text{AIC}\,(3) = 2n \log \bar{x} + 2n + 2 ,$$

where the necessary maximum likelihood estimates are defined by

$$\bar{x} = \sum_{i=1}^{n} x_i / n$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n$$

$$\hat{\theta}_1 = \sum_{i=1}^{n} \log x_i / n$$

$$\hat{\theta}_2^2 = \sum_{i=1}^{n} (\log x_i - \hat{\theta}_1)^2 / n .$$

If it is known that the sample space is bounded, we adopt as $G(x)$ the uniform distribution defined in the corresponding sample space.

In Step 2 we have to choose the number of classes m. Although it is desirable to choose it as large as possible not to distort the information of the data by discretization, setting of $m=100$ would be adequate for most practical purposes.

5. Numerical examples

We first apply the present procedure to the data shown in Table 1 which are taken from Table 1 of the paper by Maguire, Pearson and

Table 1

378	203	286	113	871	228	66	37
36	176	114	32	48	271	291	19
15	55	108	23	123	208	4	156
31	93	188	151	457	517	369	47
215	59	233	361	498	1613	338	129
11	315	28	312	49	54	336	1630
137	59	22	354	131	326	19	29
4	61	61	58	182	1312	329	217
15	1	78	275	255	348	330	7
72	13	99	78	195	745	312	18
96	189	326	17	224	217	171	1357
124	345	275	1205	566	120	145	
50	20	54	644	390	275	75	
120	81	217	467	72	20	364	

From Table 1 of [12], pp. 169. (Read columnwise)

Wynn [12] on the time intervals between severe explosions in mines and were earlier analyzed by Boneva, Kendall and Stefanov [6]. By applying the rule in the preceding section to the above data we get AIC (1)=1563.54, AIC (2)=1418.69 and AIC (3)=1415.64. Following MAICE procedure we adopt the exponential distribution as $g(x)$ for this case. Under this choice, ABIC takes its minimum 1412.07 for $v^2=0.0078125^2$ and our procedure produces probability density shown in Figure 1 as the final estimate. This figure shows the values of the density function at 50 equally spaced points. This result is clearly better than the result by Boneva et al. which does not always take positive value.

We next reanalyze the data for the distribution of silica in 22 chondrite meteors reported by Ahrens [1] and Burch and Parsons [7]; 0.04, 0.15, 0.16, 0.18, 0.40, 0.44, 0.45, 0.46, 0.47, 0.49, 0.54, 0.59, 0.64, 0.75, 0.81, 0.83, 0.84, 0.84, 0.85, 0.87, 0.87, 0.92. This data set was

***** FINAL ESTIMATES *****

NO.	VALUES OF X	ESTIMATED DENSITIES
1	0.11096D+02	0.47545D-02
2	0.33288D+02	0.41873D-02
3	0.55480D+02	0.36991D-02
4	0.77671D+02	0.32770D-02
5	0.99863D+02	0.29105D-02
6	0.12206D+03	0.25911D-02
7	0.14425D+03	0.23116D-02
8	0.16644D+03	0.20662D-02
9	0.18863D+03	0.18502D-02
10	0.21082D+03	0.16594D-02
11	0.23301D+03	0.14905D-02
12	0.25521D+03	0.13406D-02
13	0.27740D+03	0.12072D-02
14	0.29959D+03	0.10883D-02
15	0.32178D+03	0.98214D-03
16	0.34397D+03	0.88715D-03
17	0.36617D+03	0.80202D-03
18	0.38836D+03	0.72562D-03
19	0.41055D+03	0.65696D-03
20	0.43274D+03	0.59518D-03
21	0.45493D+03	0.53953D-03
22	0.47712D+03	0.48934D-03
23	0.49932D+03	0.44403D-03
24	0.52151D+03	0.40310D-03
25	0.54370D+03	0.36609D-03
26	0.56589D+03	0.33260D-03
27	0.58808D+03	0.30228D-03
28	0.61028D+03	0.27481D-03
29	0.63247D+03	0.24990D-03
30	0.65466D+03	0.22731D-03
31	0.67685D+03	0.20681D-03
32	0.69904D+03	0.18819D-03
33	0.72124D+03	0.17129D-03
34	0.74343D+03	0.15593D-03
35	0.76562D+03	0.14198D-03
36	0.78781D+03	0.12929D-03
37	0.81000D+03	0.11775D-03
38	0.83219D+03	0.10725D-03
39	0.85439D+03	0.97700D-04
40	0.87658D+03	0.89009D-04
41	0.89877D+03	0.81099D-04
42	0.92096D+03	0.73899D-04
43	0.94315D+03	0.67342D-04
44	0.96535D+03	0.61372D-04
45	0.98754D+03	0.55935D-04
46	0.10097D+04	0.50982D-04
47	0.10319D+04	0.46471D-04
48	0.10541D+04	0.42360D-04
49	0.10763D+04	0.38615D-04
50	0.10985D+04	0.35203D-04

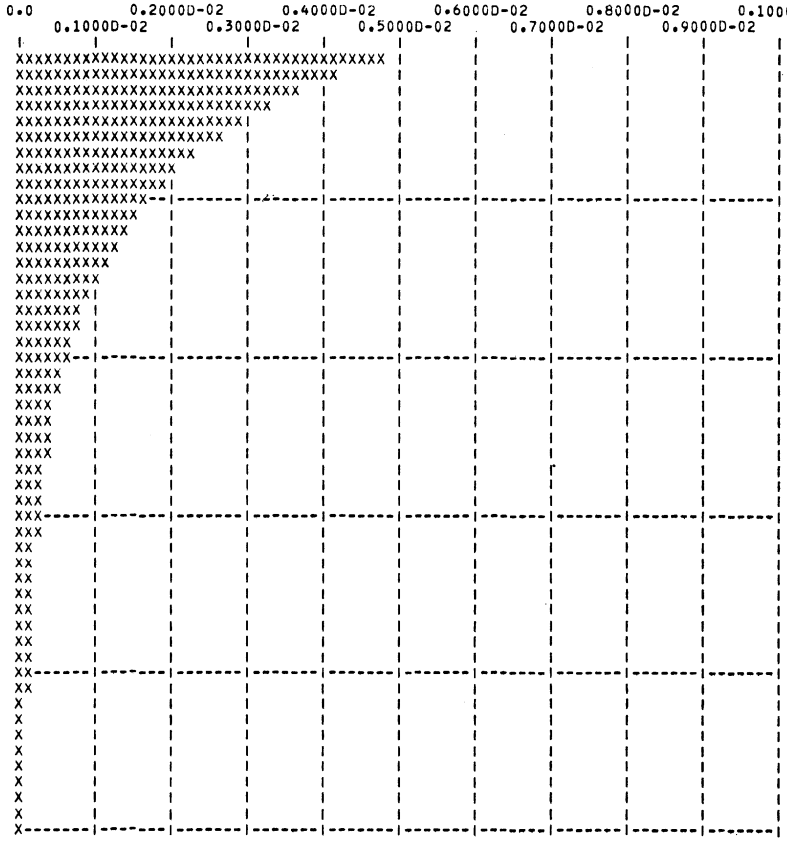

Figure 1

***** FINAL ESTIMATES *****

NO.	VALUES OF X	ESTIMATED DENSITIES
1	-0.20976D+00	0.32581D-01
2	-0.17784D+00	0.45995D-01
3	-0.14592D+00	0.63951D-01
4	-0.11400D+00	0.87506D-01
5	-0.82078D-01	0.11767D+00
6	-0.50158D-01	0.15512D+00
7	-0.18239D-01	0.20013D+00
8	0.13681D-01	0.25300D+00
9	0.45600D-01	0.31297D+00
10	0.77520D-01	0.37802D+00
11	0.10944D+00	0.44481D+00
12	0.14136D+00	0.50826D+00
13	0.17328D+00	0.56108D+00
14	0.20520D+00	0.59592D+00
15	0.23712D+00	0.61290D+00
16	0.26904D+00	0.62307D+00
17	0.30096D+00	0.64625D+00
18	0.33288D+00	0.70973D+00
19	0.36480D+00	0.84895D+00
20	0.39672D+00	0.10986D+01
21	0.42863D+00	0.14378D+01
22	0.46055D+00	0.17002D+01
23	0.49247D+00	0.16384D+01
24	0.52439D+00	0.13693D+01
25	0.55631D+00	0.11049D+01
26	0.58823D+00	0.90812D+00
27	0.62015D+00	0.78667D+00
28	0.65207D+00	0.74642D+00
29	0.68399D+00	0.79106D+00
30	0.71591D+00	0.95355D+00
31	0.74783D+00	0.12383D+01
32	0.77975D+00	0.15778D+01
33	0.81167D+00	0.18320D+01
34	0.84359D+00	0.17999D+01
35	0.87551D+00	0.14737D+01
36	0.90743D+00	0.10771D+01
37	0.93935D+00	0.75118D+00
38	0.97127D+00	0.51828D+00
39	0.10032D+01	0.36012D+00
40	0.10351D+01	0.25333D+00
41	0.10670D+01	0.18039D+00
42	0.10989D+01	0.12969D+00
43	0.11309D+01	0.93850D-01
44	0.11628D+01	0.68116D-01
45	0.11947D+01	0.49438D-01
46	0.12266D+01	0.35972D-01
47	0.12585D+01	0.26153D-01
48	0.12905D+01	0.18889D-01
49	0.13224D+01	0.13504D-01
50	0.13543D+01	0.95376D-02

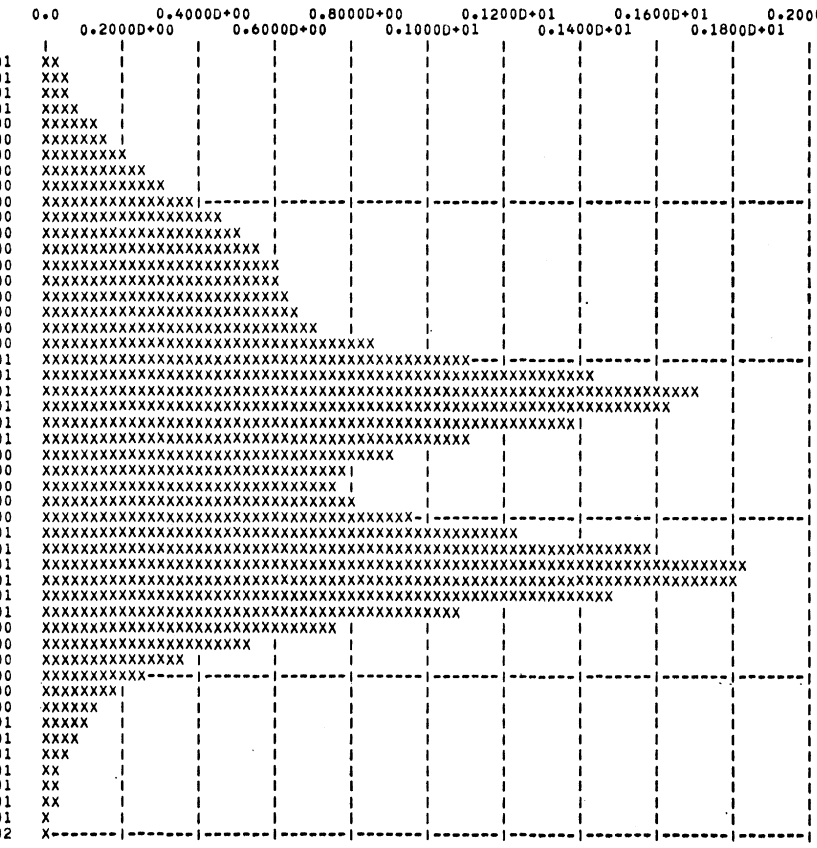

Figure 2

also analyzed by T. Leonard [11] and I. J. Good and R. A. Gaskins [9].

For this data set the normal distribution is adopted as $g(x)$ and the final estimate by our procedure is shown in Figure 2, which shows that our result is intermediate in shape between the tri-model results by Good et al. and the uni-model result by Leonard. Since we do not know the real density function the evaluation of these three results is difficult.

For the comparison of these procedures, we suggest that they will analyze the following artificial data. It is a set of 200 measurements in Table 2 that were drawn from the distribution defined by $f^*(x)=0.425\phi(x-\mu_1)+(0.15/\sigma)\phi(x/\sigma)+0.425\phi(x-\mu_2)$ with $\mu_1=-1.4$, $\mu_2=1.7$ and $\sigma=0.5$, where $\phi(x)$ is the density function of the standard normal distribution. The frequencies under 50 classes are represented graphically in Figure 3. The final estimate for this set of data by our procedure is shown in Figure 4 and the graph of the real distribution is shown

Table 2

3.989	2.461	0.057	2.957	0.271
1.880	-1.949	0.370	-2.071	-2.076
-2.110	0.820	-1.847	-2.718	-0.604
1.713	-0.259	-1.857	-2.431	-0.039
1.006	-2.267	-4.300	3.334	-0.395
2.962	-2.339	0.104	0.548	1.862
-0.846	-1.455	1.322	1.040	-1.753
1.564	-1.406	0.072	2.886	-1.407
-2.985	3.398	-1.797	-0.293	-1.675
-2.733	1.963	-0.272	-0.965	0.101
3.294	-2.102	-1.717	-0.404	0.008
3.326	-1.674	-1.362	2.404	-1.879
1.336	-0.562	0.134	-3.003	-1.824
-0.344	1.565	-2.135	1.131	2.298
-1.819	1.912	-0.980	2.443	1.443
1.830	0.547	-0.999	-0.131	1.773
3.148	-0.560	2.377	1.556	-0.343
-0.460	0.437	-0.184	-0.508	2.840
4.229	1.963	-0.706	1.573	2.756
0.977	0.607	1.253	-2.478	0.026
-1.195	-2.588	-2.541	-0.203	-1.474
0.594	0.899	1.855	2.778	2.853
0.508	0.357	1.449	-2.939	0.690
-0.949	1.916	-1.765	-0.039	0.688
0.621	2.051	-0.248	-2.684	3.358
-1.959	0.820	0.565	-0.669	2.583
2.157	-0.139	2.528	1.220	0.001
0.452	2.440	-0.839	-1.757	1.338
-0.420	-1.854	1.890	-3.279	2.759
-1.362	-1.854	-0.283	1.707	-3.182
1.529	-0.114	0.182	-1.132	0.360
-1.166	-0.261	1.541	-0.303	-0.077
2.484	3.150	2.747	2.164	-0.158
0.378	-1.607	1.931	-0.284	2.256
0.871	2.944	3.291	2.084	-0.440
-1.755	0.538	0.286	-0.075	1.234
-1.559	-2.789	0.703	0.004	2.603
-1.001	-0.019	-2.279	0.214	2.078
0.302	-0.606	0.200	2.373	3.902
-2.083	-3.290	-1.524	-1.016	1.835

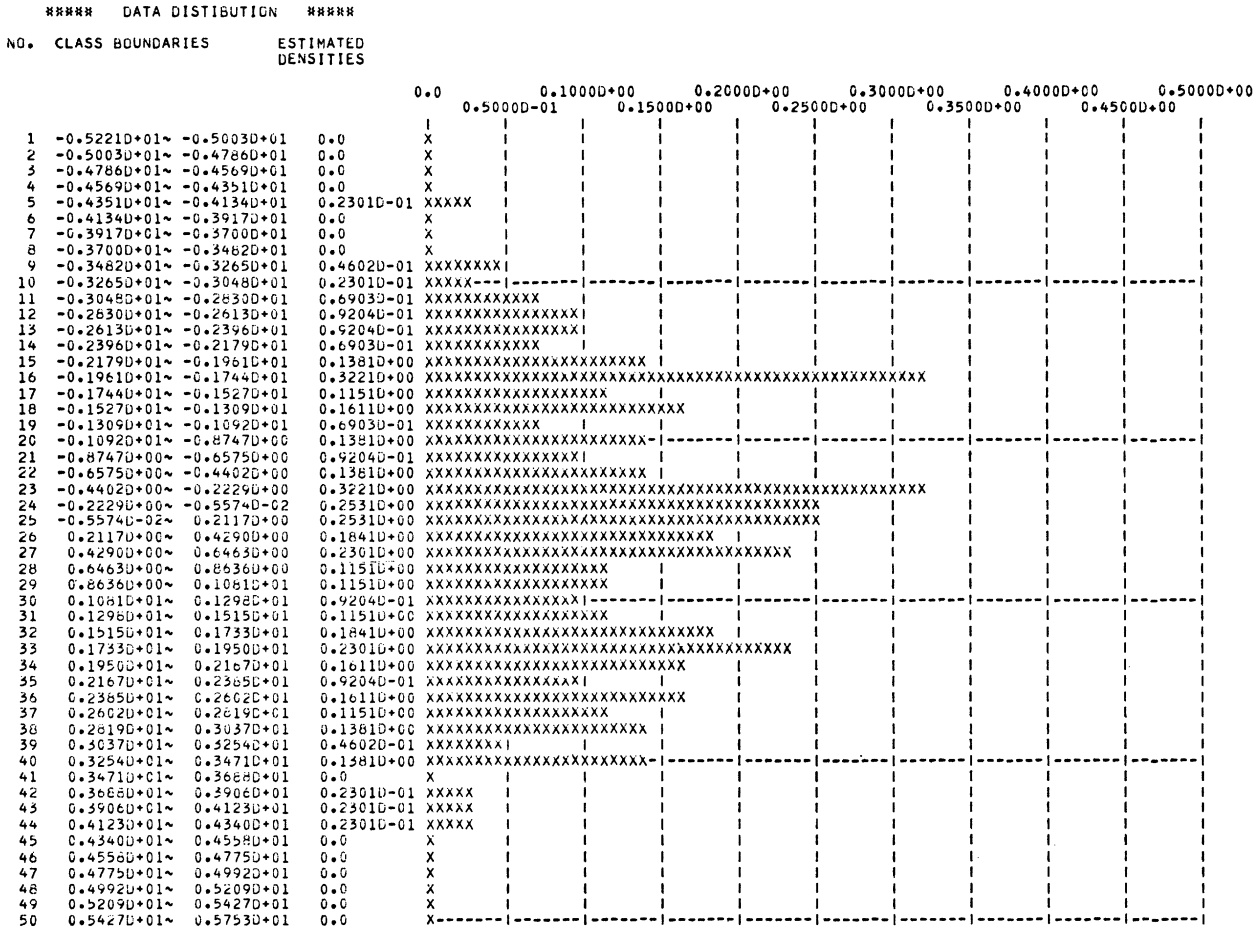
***** DATA DISTIBUTION *****
NO. CLASS BOUNDARIES ESTIMATED DENSITIES
0.0 0.1000D+00 0.2000D+00 0.3000D+00 0.4000D+00 0.5000D+00
0.5000D-01 0.1500D+00 0.2500D+00 0.3500D+00 0.4500D+00
1 -0.5221D+01~ -0.5003D+01 0.0
2 -0.5003D+01~ -0.4786D+01 0.0
3 -0.4786D+01~ -0.4569D+01 0.0
4 -0.4569D+01~ -0.4351D+01 0.0
5 -0.4351D+01~ -0.4134D+01 0.2301D-01 XXXXX
6 -0.4134D+01~ -0.3917D+01 0.0
7 -0.3917D+01~ -0.3700D+01 0.0
8 -0.3700D+01~ -0.3482D+01 0.0
9 -0.3482D+01~ -0.3265D+01 0.4602D-01 XXXXXXXX
10 -0.3265D+01~ -0.3048D+01 0.2301D-01 XXXXX
11 -0.3048D+01~ -0.2830D+01 0.6903D-01 XXXXXXXXXXX
12 -0.2830D+01~ -0.2613D+01 0.9204D-01 XXXXXXXXXXXXXXX
13 -0.2613D+01~ -0.2396D+01 0.9204D-01 XXXXXXXXXXXXXXX
14 -0.2396D+01~ -0.2179D+01 0.6903D-01 XXXXXXXXXXX
15 -0.2179D+01~ -0.1961D+01 0.1381D+00 XXXXXXXXXXXXXXXXXXXXXX
16 -0.1961D+01~ -0.1744D+01 0.3221D+00 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
17 -0.1744D+01~ -0.1527D+01 0.1151D+00 XXXXXXXXXXXXXXXXX
18 -0.1527D+01~ -0.1309D+01 0.1611D+00 XXXXXXXXXXXXXXXXXXXXXXX
19 -0.1309D+01~ -0.1092D+01 0.6903D-01 XXXXXXXXXXX
20 -0.1092D+01~ -0.8747D+00 0.1381D+00 XXXXXXXXXXXXXXXXXXXXX
21 -0.8747D+00~ -0.6575D+00 0.9204D-01 XXXXXXXXXXXXXX
22 -0.6575D+00~ -0.4402D+00 0.1381D+00 XXXXXXXXXXXXXXXXXXXXX
23 -0.4402D+00~ -0.2229D+00 0.3221D+00 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
24 -0.2229D+00~ -0.5574D-02 0.2531D+00 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
25 -0.5574D-02~ 0.2117D+00 0.2531D+00 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
26 0.2117D+00~ 0.4290D+00 0.1841D+00 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
27 0.4290D+00~ 0.6463D+00 0.2301D+00 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
28 0.6463D+00~ 0.8636D+00 0.1151D+00 XXXXXXXXXXXXXXXXXX
29 0.8636D+00~ 0.1081D+01 0.1151D+00 XXXXXXXXXXXXXXXXXX
30 0.1081D+01~ 0.1298D+01 0.9204D-01 XXXXXXXXXXXXXXX
31 0.1298D+01~ 0.1515D+01 0.1151D+00 XXXXXXXXXXXXXXXXXX
32 0.1515D+01~ 0.1733D+01 0.1841D+00 XXXXXXXXXXXXXXXXXXXXXXXXXXXX
33 0.1733D+01~ 0.1950D+01 0.2301D+00 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
34 0.1950D+01~ 0.2167D+01 0.1611D+00 XXXXXXXXXXXXXXXXXXXXXXXXXX
35 0.2167D+01~ 0.2385D+01 0.9204D-01 XXXXXXXXXXXXXXX
36 0.2385D+01~ 0.2602D+01 0.1611D+00 XXXXXXXXXXXXXXXXXXXXXXXXXX
37 0.2602D+01~ 0.2819D+01 0.1151D+00 XXXXXXXXXXXXXXXXXX
38 0.2819D+01~ 0.3037D+01 0.1381D+00 XXXXXXXXXXXXXXXXXXXXX
39 0.3037D+01~ 0.3254D+01 0.4602D-01 XXXXXXXX
40 0.3254D+01~ 0.3471D+01 0.1381D+00 XXXXXXXXXXXXXXXXXXXXXX
41 0.3471D+01~ 0.3688D+01 0.0
42 0.3688D+01~ 0.3906D+01 0.2301D-01 XXXXX
43 0.3906D+01~ 0.4123D+01 0.2301D-01 XXXXX
44 0.4123D+01~ 0.4340D+01 0.2301D-01 XXXXX
45 0.4340D+01~ 0.4558D+01 0.0
46 0.4558D+01~ 0.4775D+01 0.0
47 0.4775D+01~ 0.4992D+01 0.0
48 0.4992D+01~ 0.5209D+01 0.0
49 0.5209D+01~ 0.5427D+01 0.0
50 0.5427D+01~ 0.5753D+01 0.0

Figure 3

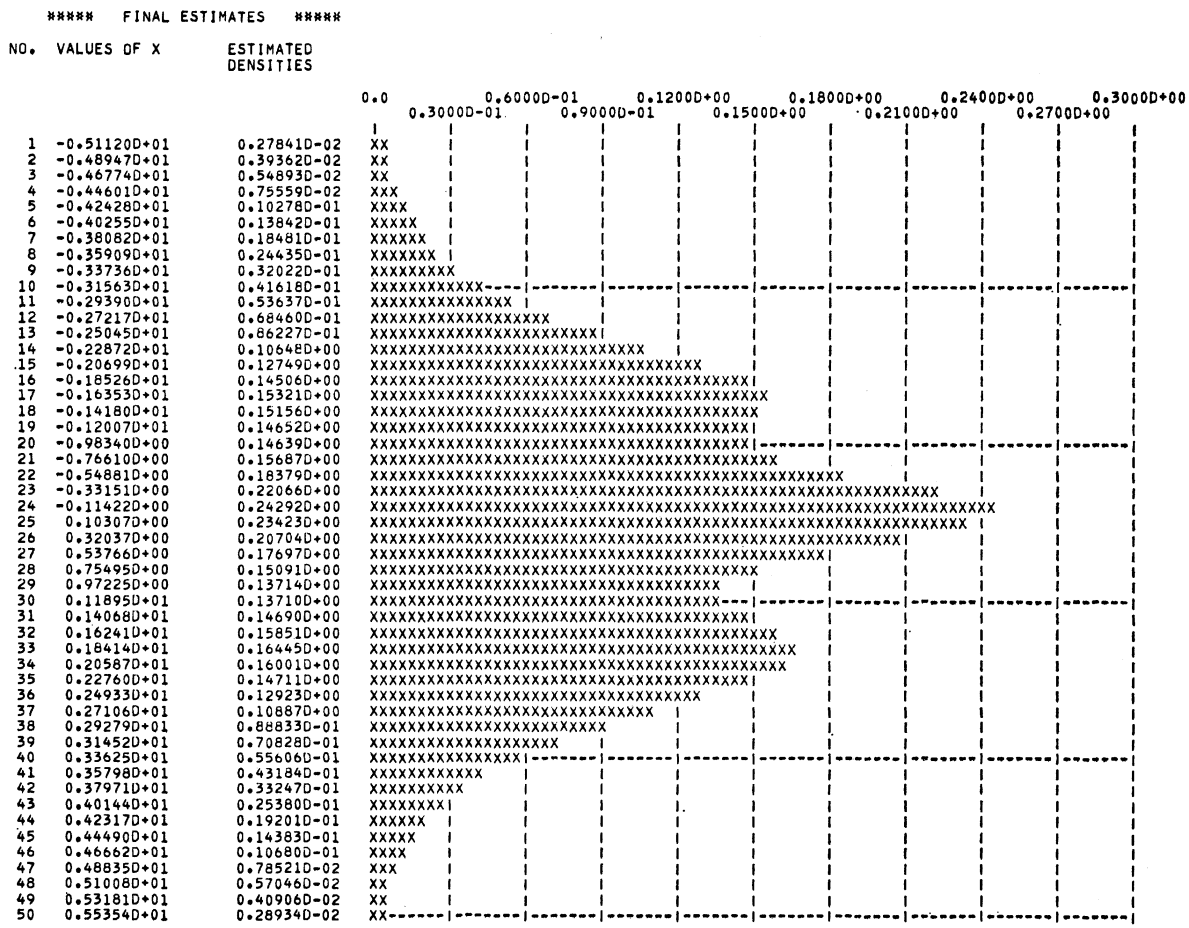
***** FINAL ESTIMATES *****
NO. VALUES OF X ESTIMATED DENSITIES
0.0 0.6000D-01 0.1200D+00 0.1800D+00 0.2400D+00 0.3000D+00
0.3000D-01 0.9000D-01 0.1500D+00 0.2100D+00 0.2700D+00
1 -0.51120D+01 0.27841D-02
2 -0.48947D+01 0.39362D-02
3 -0.46774D+01 0.54893D-02
4 -0.44601D+01 0.75559D-02
5 -0.42428D+01 0.10278D-01
6 -0.40255D+01 0.13842D-01
7 -0.38082D+01 0.18481D-01
8 -0.35909D+01 0.24435D-01
9 -0.33736D+01 0.32022D-01
10 -0.31563D+01 0.41618D-01
11 -0.29390D+01 0.53637D-01
12 -0.27217D+01 0.68460D-01
13 -0.25045D+01 0.86227D-01
14 -0.22872D+01 0.10648D+00
15 -0.20699D+01 0.12749D+00
16 -0.18526D+01 0.14506D+00
17 -0.16353D+01 0.15321D+00
18 -0.14180D+01 0.15156D+00
19 -0.12007D+01 0.14652D+00
20 -0.98340D+00 0.14639D+00
21 -0.76610D+00 0.15687D+00
22 -0.54881D+00 0.18379D+00
23 -0.33151D+00 0.22066D+00
24 -0.11422D+00 0.24292D+00
25 0.10307D+00 0.23423D+00
26 0.32037D+00 0.20704D+00
27 0.53766D+00 0.17697D+00
28 0.75495D+00 0.15091D+00
29 0.97225D+00 0.13714D+00
30 0.11895D+01 0.13710D+00
31 0.14068D+01 0.14690D+00
32 0.16241D+01 0.15851D+00
33 0.18414D+01 0.16445D+00
34 0.20587D+01 0.16001D+00
35 0.22760D+01 0.14711D+00
36 0.24933D+01 0.12923D+00
37 0.27106D+01 0.10887D+00
38 0.29279D+01 0.88833D-01
39 0.31452D+01 0.70828D-01
40 0.33625D+01 0.55606D-01
41 0.35798D+01 0.43184D-01
42 0.37971D+01 0.33247D-01
43 0.40144D+01 0.25380D-01
44 0.42317D+01 0.19201D-01
45 0.44490D+01 0.14383D-01
46 0.46662D+01 0.10680D-01
47 0.48835D+01 0.78521D-02
48 0.51008D+01 0.57046D-02
49 0.53181D+01 0.40906D-02
50 0.55354D+01 0.28934D-02

Figure 4

in Figure 5. Figure 4 illustrates that our procedure satisfactorily reproduces the real density function which could not be guessed from Figure 3 by eye.

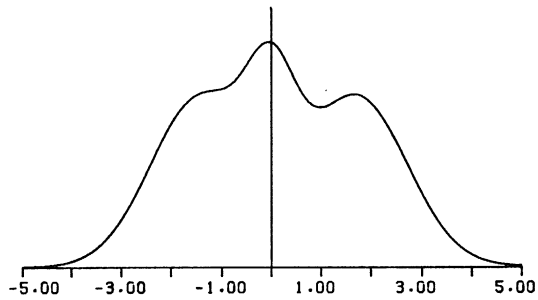

Figure 5

Of course, the performance of the present procedure depends on the sample size and the true structure. For instance, in case of the last example, we need much more sample size, say $n=1000$, to estimate the precise detailes of the true distribution. If the true distribution had simpler structure, the present procedure would work for data of size less than 200. To test further the performance of the procedure, we generated 100 sets of data each composed of random sample of size 100 from the density function defined by $f^*(x)=0.5\phi(x)+0.5\phi(x-4.0)$. Our procedure has detected 98 times out of 100 that the true distribution has two bumps. These observations imply that our procedure is good enough for most practical use provided that the density varies smoothly.

6. Concluding remarks

As is seen from the arguments in the previous sections, the present procedure will be practical and useful since it automatically gives stable estimate of density. In our procedure, however, we approximated the function $h^*(y)$ in (1) by a piece-wise constant function. As is noted in Section 4, we could have approximated $h^*(y)$ by a higher order spline function and develop a procedure to estimate its parameters. We think, however, that what we gain by such an approach will not be worth, at least for the estimation of unidimensional density, the amount of computational task required. We did not discuss the estimation of two and higher dimensional density in the present paper. Those aspects will be the subject of future study.

Appendix

Multinomial logit transformations of the parameters of the histogram model and its nature

Let $\boldsymbol{p}=(p_1, p_2, \cdots, p_m)^t$ denote a vector whose components are positive and add up to 1. Then the loglikelihood of the histogram on the interval [0, 1] is defined by

$$l(\boldsymbol{p})=\sum_{j=1}^{m} n(j) \log (mp_j) .$$

Define the multinomial logit transformation of $\boldsymbol{p}$ by $\boldsymbol{h}=(h_1, h_2, \cdots, h_{m-1})^t$ whose elements are given by

$$h_j=\log \left\{\prod_{k=1}^{m} \frac{p_j}{p_k}\right\}=m \log p_j-\sum_{k=1}^{m} \log p_k \qquad (j=1, 2, \cdots, m-1)$$

then the inverse transformation is given by

$$p_j=\frac{\exp (h_j/m)}{\sum_{k=1}^{m} \exp (h_j/m)} \qquad (j=1, 2, \cdots, m)$$

where h_m is defined by

$$h_m=-\sum_{j=1}^{m-1} h_j .$$

Then the log likelihood of the model as a function of $\boldsymbol{h}$ is given by

$$l(\boldsymbol{h})=\sum_{j=1}^{m} n(j) \log (mp_j)=\sum_{j=1}^{m} n(j) \log \left\{\frac{m \exp (h_j/m)}{\sum_{k=1}^{m} \exp (h_k/m)}\right\} .$$

The kth element of the gradient of $l(\boldsymbol{h})$ is given by

$$\frac{\partial l}{\partial h_k}=\frac{n(k)-n(m)}{m}-\frac{n}{m}(p_k-p_m) .$$

$\partial p_k/\partial h_j$ is given by

$$\frac{\partial p_k}{\partial h_j}=\begin{cases} \dfrac{1}{m}\{\delta_{kj}p_k-p_k(p_j-p_m)\} & \text{if } k\neq m \\ \dfrac{1}{m}\{-p_m-p_m(p_j-p_m)\} & \text{if } k=m \end{cases}$$

where δ_{kj} is Kronecker's delta. Therefore, the (k, j)th element of the Hessian of $l(\boldsymbol{h})$ is given by

$$\frac{\partial^2 l}{\partial h_k \partial h_j} = -\frac{n}{m^2}\{\delta_{kj}p_k + p_m - (p_k - p_m)(p_j - p_m)\} = -\frac{n}{m^2}G_{kj} .$$

Let G be a matrix whose (k, j)th element is G_{kj}, then it is explicitly given by

$$G = \begin{pmatrix} \sqrt{p_1} & & & & \sqrt{p_m} \\ & \sqrt{p_2} & & & \sqrt{p_m} \\ & & \ddots & & \vdots \\ & & & \sqrt{p_{m-1}} & \sqrt{p_m} \end{pmatrix} \begin{pmatrix} \sqrt{p_1} & & & \\ & \sqrt{p_2} & & \\ & & \ddots & \\ & & & \sqrt{p_{m-1}} \\ \sqrt{p_m} & \sqrt{p_m} & \cdots & \sqrt{p_m} \end{pmatrix}$$

$$- \begin{pmatrix} p_1 - p_m \\ p_2 - p_m \\ \vdots \\ p_{m-1} - p_m \end{pmatrix} (p_1 - p_m, p_2 - p_m, \cdots, p_{m-1} - p_m) .$$

Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_{m-1})^t$ be any vector, and defining x_m by

$$x_m = -\sum_{j=1}^{m-1} x_j$$

it follows:

$$\boldsymbol{x}^t G \boldsymbol{x} = \sum_{j=1}^{m} p_j x_j^2 - \left(\sum_{j=1}^{m} p_j x_j\right)^2 \geqq 0 .$$

Note that the equality implies, if $p_j > 0$ $j = 1, 2, \cdots, m$, that

$$x_j = 0 \qquad (j = 1, \cdots, m-1) .$$

This proves the positive definitness of G.

Acknowledgement

The authors are grateful to Dr. Akaike for his helpful comments and to Mr. K. Katsura of the Institute of Statistical Mathematics for computing assistance.

Their thanks are also to the referees for their valuable suggestions.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Ahrens, L. H. (1965). Observations on the Fe-Si-Mg relationship in chondrites, *Geochimica et Cosmochimica Acta*, **29**, 801-806.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, (eds. B. N. Petrov and F. Csaki), Akademiai Kiado, Budapest, 267-281.

[3] Akaike, H. (1977). On entropy maximization principle, *Applications of Statistics*, (ed. Krishnaiah, P. R.), North-Holland, Amsterdam, 27-41.

[4] Akaike, H. and Arahata, E. (1978). GALTHY, A probability density estimation, *Computer Science Monographs*, No. **9**, The Institute of Statistical Mathematics, Tokyo.

[5] Akaike, H. (1980). Likelihood and Bayes procedure, *Bayesian Statistics*, (eds. J. M. Bernardo, M. H. De Groot, D. U. Lindley and A. F. M. Smith), University Press, Valencia, Spain.

[6] Boneva, L. I., Kendall, D. G. and Stefanov, I. (1971). Spline transformations: three new diagnostic aids for the statistical data-analyst, *J. R. Statist. Soc.*, B, **33**, 1-71.

[7] Burch, C. R. and Parsons, I. T. (1976). Squeeze significance tests, *Applied Statistics*, **25**, 287-291.

[8] Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities, *Biometrika*, **58**, 255-277.

[9] Good, I. J. and Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data, *J. Amer. Statist. Ass.*, **75**, 42-56.

[10] Ishiguro, M. and Sakamoto, Y. (1983). A Bayesian approach to binary response curve estimation, *Ann. Inst. Statist. Math.*, B, **35**, 115-137.

[11] Leonard, Tom (1978). Density estimation, stochastic processes, and prior information, *J. R. Statist. Soc.*, B, **40**, 113-146.

[12] Maguire, B. A., Pearson, E. S. and Wynn, A. H. A. (1952). The time intervals between industrials accidents, *Biometrika*, **39**, 168-180.

[13] Parzen, E. (1962). On estimation of a probability density and mode, *Ann. Math. Statist.*, **33**, 1065-1076.

[14] Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function, *Ann. Math, Statist.*, **27**, 832-837.

[15] Whittle, P. (1958). On the smoothing of probability density functions, *J. R. Statist. Soc.*, B, **20**, 334-343.