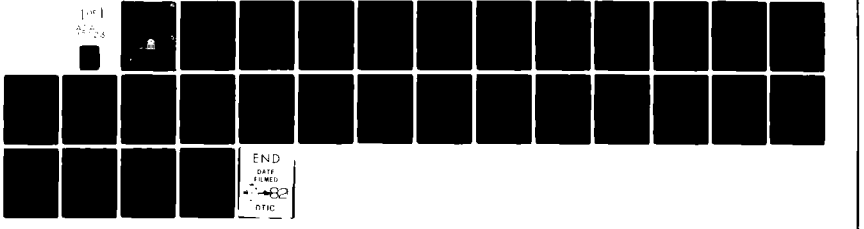
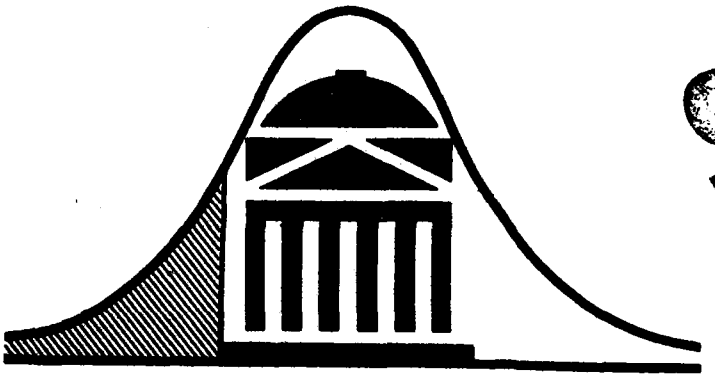
AD-A117 023 SOUTHERN METHODIST UNIV DALLAS TEX DEPT OF STATISTICS F/G 12/1
OPTIMAL GROUPINGS, SPACINGS, STRATIFICATION AND PIECEWISE CONSTAN--ETC(U)
JUN 82 R L EUBANK N00014-82-K-0207
UNCLASSIFIED TR-164 NL



SOUTHERN METHODIST UNIVERSITY



DEPARTMENT OF STATISTICS

DALLAS, TEXAS 75275

(12)

OPTIMAL GROUPING, SPACING, STRATIFICATION AND PIECEWISE CONSTANT APPROXIMATION

by

R. L. Eubank

Technical Report No. 164
Department of Statistics ONR Contract

June 1982

Research sponsored by the Office of Naval Research
Contract N00014-82-K-0207
Project NR 042-479

Department of Statistics
Southern Methodist University
Dallas, Texas 75275

OPTIMAL GROUPING, SPACING, STRATIFICATION AND PIECEWISE CONSTANT APPROXIMATION

by

R. L. Eubank[1]
Southern Methodist University

Abbreviated title: Grouping, Spacing and Stratification

Summary. A variety of statistical problems of optimal grouping, spacing and stratification are seen to be best $L_2[0,1]$ free knot piecewise constant approximation problems. This allows for the development of conditions that insure the existence and uniqueness of solutions, a computational algorithm and simple approximate solutions. In addition this approach is seen to provide insight into the relationships between the various problems considered.

[1]Research sponsored by Office of Naval Research contract N00014-82-K-0207.

AMS 1980 subject classification. Primary 62F10, 62F03, Secondary 65D07.

Key words and phrases. Grouping, spacing, splines, stratification.

Optimal Grouping, Spacing, Stratification and Piecewise Constant Approximation

by

R. L. Eubank
Southern Methodist University

1. Introduction and Summary. It has been recognized for some time that there is a structural similarity between certain problems of optimal grouping, spacing, and stratification. See, for example, Cox (1957),Kulldorff (1958a,b, 1961), Särndal (1961, 1962), Ekman (1969), Bofinger (1975), Bühler and Deutler (1975) and Adatia and Chan (1981). In this paper the underlying relationship between these and other problems is established. Specifically, it is shown that all these problems, when viewed in the quantile domain, become problems of optimal knot (breakpoint) selection for piecewise constant $L_2[0,1]$ approximation. This fact allows us to develop a unified approach to all such problems that includes i) conditions for existence and uniqueness of solutions ii) a computational algorithm and iii) simple approximate solutions. In addition, this approach provides insight into the geometry of and connection between these problem areas. Questions pertaining to the equivalence of certain problems, such as considered by Adatia and Chan (1981), become questions regarding the equivalence of certain function approximation problems.

In the next section we examine a canonical form for the problems to be considered and establish our principal results concerning its solution. In subsequent sections these results are applied to various problems of optimal stratification and grouping, optimal spacing and grouping and some bivariate stratification and grouping problems

that have appeared in the literature.

2. An optimal grouping problem. Let X be a random variable with strictly increasing distribution function (d.f.) F and associated continuous probability density function (p.d.f.) $f = F'$. Define the quantile function (q.f.) for F as $Q(u) = F^{-1}(u)$, $0 < u < 1$, and density-quantile function by $fQ(u) = f(Q(u))$, $0 \leq u \leq 1$. Also let $a = x_0 < x_1 < \ldots < x_{k+1} = b$ (where we allow for either or both of $a = -\infty$, $b = \infty$) represent a partition of the range of X and note that the set of percentile points associated with the x_i's, $U = \{u_0,\ldots,u_{k+1}\}$, is uniquely defined by

$$\begin{aligned} u_0 &= 0 \\ Q(u_i) &= x_i, \quad i = 1,\ldots,k, \\ u_{k+1} &= 1. \end{aligned} \tag{2.1}$$

The probability mass corresponding to the ith interval can then be written as

$$F(x_i) - F(x_{i-1}) = u_i - u_{i-1} . \tag{2.2}$$

Suppose that instead of X the object of interest is a monotone increasing transformation T(X) which, for convenience of presentation, is discretized to obtain a new variable

$$T_U(X) = m_i , \quad x_{i-1} < x \leq x_i , \tag{2.3}$$

where m_i is the conditional mean of T(X) on the ith interval, i.e.,

$$m_i = (u_i-u_{i-1})^{-1}\int_{x_{i-1}}^{x_i} T(x)f(x)dx = (u_i-u_{i-1})^{-1}\int_{u_{i-1}}^{u_i} TQ(u)du, \tag{2.4}$$

and $TQ(u) = T(Q(u))$. The characteristics of T(X) are then summarized by (x_i,m_i), $i = 1,\ldots,k+1$. Observing that T(X) and $T_U(X)$ will have

identical expectations, whose common value may be taken without loss of generality as zero, the "within group variance" of this summarization scheme can be written as

$$V(T(X) - T_U(X)) = \int_0^1 TQ(u)^2 du - \Sigma_{i=1}^{k+1}(u_i - u_{i-1})m_i^2 \quad . \tag{2.5}$$

Since this variance is a function of the partition or grouping, the x_i's, or equivalently U, should be chosen to minimize (2.5). Therefore, let us define the set of all "k-point spacings" by

$$D_k = \{(u_0, u_1, \ldots, u_{k+1}): 0 = u_0 < u_1 < \ldots < u_{k+1} = 1\} \tag{2.6}$$

and consider the problem of selecting a $U^* \varepsilon D_k$ that satisfies

$$V(T(X) - T_{U^*}(X)) = \inf_{U \varepsilon D_k} V(T(X) - T_U(X)). \tag{2.7}$$

A spacing U* satisfying (2.7) will be termed an optimal spacing. It should be emphasized that choosing an optimal spacing is equivalent to choosing an optimal partition. In subsequent work we will, therefore, often indicate only how to obtain an optimal U with the use of (2.1) to obtain the corresponding grouping an implied next step.

Let $\langle \cdot, \cdot \rangle$ and $||\cdot||$ denote the usual $L_2[0,1]$ inner product and norm and note that $(u_i - u_{i-1})m_i^2 = \langle TQ, B_i \rangle^2$ where B_i is the ith normalized B-spline for the knot sequence u_i, $i = 1, \ldots, k+1$, with

$$B_i(u) = \begin{cases} (u_i - u_{i-1})^{-1/2} & , \quad u_{i-1} < u \leq u_i \\ 0 & , \quad \text{otherwise.} \end{cases}$$

Then, as $\langle B_i, B_j \rangle = \delta_{ij}$, we have

$$\begin{aligned} V(T(X) - T_U(X)) &= \int_0^1 TQ(u)^2 du - \Sigma_{i=1}^{k+1} \langle TQ, B_i \rangle^2 \\ &= ||TQ - P_U(TQ)||^2 \end{aligned} \tag{2.8}$$

where P_U is the projection operator for the linear span of $\{B_i: i = 1, \ldots, k+1\}$. Thus minimizing (2.5) with respect to U is

equivalent to finding the best set of knots (breakpoints) for the approximation of TQ by splines of order one (piecewise constants). Regarding this latter problem several results can be deduced from the approximation theory literature that, for the problem at hand, may be stated as follows.

Proposition (Barrar and Loeb (1970)). If TQ is square integrable (i.e. T(X) has finite variance) and is not piecewise constant for any k there exists at least one optimal $U \in D_k$ and, hence, at least one optimal partition for $T_U(X)$.

Theorem 1 (Chow (1982)) If $TQ \in C^1[0,1] \cap L_2[0,1]$ with $(TQ)' > 0$ on [0,1] a necessary condition for U to minimize $V(T(X) - T_U(X))$ is that

$$S_i(U) = 2TQ(u_i) - m_i - m_{i+1} = 0, \quad i = 1,\ldots,k . \tag{2.9}$$

If, in addition, $\log(TQ)'$ is concave on (0,1) the solution to (2.9) is unique and, hence, the optimal spacings for problem (2.7) are unique for each positive integer k.

Equations (2.9) provide a method for computing optimal spacing candidates since, under the stated necessary conditions, one may use Newton's method to search for zeros of the mapping $S(U) = (S_1(U),\ldots,S_k(U))$. This is particularly simple, in this case, since the Jacobian matrix is tridiagonal with nonzero elements

$$\frac{\partial S_i}{\partial u_{i-1}} = (u_i - u_{i-1})^{-1}[TQ(u_{i-1}) - m_i] , \quad 2 \le i \le k , \tag{2.10}$$

$$\frac{\partial S_i}{\partial u_i} = 2(TQ)'(u_i) - TQ(u_i)[(u_i-u_{i-1})^{-1}-(u_{i+1}-u_i)^{-1}] - (u_i-u_{i-1})^{-1}m_i - (u_{i+1} - u_i)^{-1}m_{i+1} , \quad 1 \le i \le k , \tag{2.11}$$

and

$$\frac{\partial S_i}{\partial u_{i+1}} = -(u_{i+1}-u_i)^{-1}[TQ(u_{i+1}) - m_{i+1}],\ 1 \le i \le k-1. \qquad (2.12)$$

When log (TQ)' is concave, it follows from the proof of Theorem 1 that the Jacobian is diagonally dominant and positive definite at the optimal spacing so that, with a good initial guess, Newton's method will find the optimal solution. A discussion of uniqueness conditions such as those in Theorem 1, as well as the algorithm implied by (2.9)-(2.12) that is phrased in a regression design setting can be found in Eubank, Smith, and Smith (1981, 1982). See also Barrow, et al (1978) for related work.

Frequently for complicated TQ functions it will be convenient to use the approximate (asymptotic) solution provided by the next theorem whose proof is an application of Theorem 1.1 of Burchard and Hale (1975) and Theorem 4.4 of Pence and Smith (1982).

Theorem 2. Assume that $TQ \in L_2[0,1] \cap C(0,1)$ and that either i) $(TQ)' \in C[0,1]$ or ii) $|(TQ)'|$ is integrable over $[\alpha,\beta]$ for any $0 < \alpha < \beta < 1$ and monotone almost everywhere with $|(TQ)'(u)|^{2/3}$ integrable. Define the density

$$h(u) = |(TQ)'(u)|^{2/3} / \int_0^1 |(TQ)'(s)|^{2/3} ds \qquad (2.13)$$

with corresponding q.f., H^{-1}, assumed to be in $C^1[0,1]$ and let $\{U_k\}$ denote the spacing sequence whose kth element is $U_k = \{0, H^{-1}(\frac{1}{k+1}), \ldots, H^{-1}(\frac{k}{k+1}), 1\}$. Under these assumption

$$\lim_{k\to\infty} k^2 V(T(X) - T_{U_k}(X)) = \lim_{k\to\infty} k^2 \inf_{U \in D_k} V(T(X) - T_U(X))$$

$$= [\int_0^1 |(TQ)'(u)|^{2/3} du]^3/12. \qquad (2.14)$$

Theorem 2 has the interpretation that the within group variances corresponding to optimal spacings and spacings chosen as the (k+1)-

tiles of h have identical asymptotic (as $k\to\infty$) behaviour which suggests that a computationally expedient solution may be obtained by using the partition $x_i = Q(H^{-1}(i/k+1))$ for k sufficiently large. Alternative conditions on h (rather than H^{-1}) that, under assumption i), also imply Theorem 2 are given in Theorem 3.1 of Sacks and Ylvisaker (1968).

In the remainder of this paper it will be seen that a variety of statistical problems can be formulated as in this section and, hence are all variable knot piecewise constant approximation problems. Consequently, Theorems 1 and 2 furnish a unified approach that, in many cases, provides new results for the problem areas we consider. Connections with the work of others will be discussed in the appropriate sections. However, we note at the outset that the conditions imposed here appear to be weaker than those employed by others to obtain comparable results. In addition, the uniqueness conditions in Theorem 1 are essentially the first of their kind for most of the problems we examine. This is of particular importance in view of their implications for the computational algorithm that follows from equations (2.9)-(2.12).

To conclude this section it should be noted that in some cases, which arise subsequently, T(X) will involve unknown parameters. In such instances values that may be used for these parameters may be available from previous or pilot studies, prior knowledge or, perhaps, from a null hypothesis that is to be tested. Of course if the parameters are of a "location-scale" variety, i.e., $TQ(u)=c+dW(u)$ for some known function W, an optimal value for U can still be determined since $||TQ-P_U(TQ)|| = |d|\ ||W-P_UW||$. Although the computation of the x_i's may still require knowledge of c and d, the optimal U's will

still be useful in analyzing the robustness of (2.5) to incorrect guesses for the parameter values (c.f. Kulldorff (1961, Sections 2.7, 8.5 and 9.4)).

3. Optimal stratification and grouping. In this section several problems of optimal stratification and grouping are considered that are related in the sense that all can be formulated as piecewise constant approximation problems for the quantile function. In each case the results of Section 2 provide techniques for both exact and approximate solutions. We begin with an optimal stratification problem.

For a random variable X with continuous p.d.f., f, Dalenius (1950) considered the problem of dividing the range of X into strata, with boundaries $a = x_0 < \ldots < x_{k+1} = b$, so as to minimize the variance of the usual estimate of the mean from a stratified random sample of size N, $\bar{X} = \Sigma_{i=1}^{k+1}(F(x_i)-F(x_{i-1}))\bar{X}_i$ where $\bar{X}_i$ is the sample mean for the ith stratum. Using the notation of Section 2 the mean and variance of the ith stratum $(x_{i-1},x_i]$ can be written as

$$m_i = (u_i-u_{i-1})^{-1}\int_{u_{i-1}}^{u_i} Q(u)du \tag{3.1}$$

and

$$\sigma^2 = (u_i - u_{i-1})^{-1}\int_{u_{i-1}}^{u_i} (Q(u)-m_i)^2du. \tag{3.2}$$

For proportional allocation, where the number of elements taken from the ith stratum is $N(u_i-u_{i-1})$, it follows from Dalenius (1950) and the previous section that the variance of $\bar{X}$ is

$$\begin{aligned} V(\bar{X}) &= N^{-1}\Sigma_{i=1}^{k+1}(u_i-u_{i-1})\sigma_i^2 = N^{-1}[\int_0^1 Q(u)^2du - \Sigma_{i=1}^{k+1}(u_i-u_{i-1})m_i^2] \\ &= N^{-1}||Q - P_UQ||^2 \quad . \end{aligned} \tag{3.3}$$

Thus, selecting strata to minimize $V(\bar{X})$, under proportional allocation, is equivalent to finding the best set of knots for $L_2[0,1]$ piecewise constant approximation of the q.f. and we are now in a position to apply results from Section 2. Consequently, for $Q \in C^1[0,1] \cap L_2[0,1]$ with $Q' > 0$ on $[0,1]$ optimal spacing candidates may be found as solutions to

$$S_i(U) = 2Q(u_i) - m_i - m_{i+1} = 0, \quad i = 1,\ldots,k, \tag{3.4}$$

using (2.10) - (2.12), with $TQ = Q$, and Newton's method. Equations (3.4) were first considered in the context of optimal stratification by Daleniun (1950). An approximate solution to these equations is provided by $u_i = H^{-1}(i/k+1)$, the $(k+1)$-tiles of the density

$$h(u) = \{Q'(u)\}^{2/3} / \int_0^1 \{Q'(s)\}^{2/3} ds. \tag{3.5}$$

Examples of this approximate solution are $u_i = i/k+1$ for the uniform distribution $(Q(u)=u)$, $u_i = 1-(1-\frac{i}{k+1})^3$ for the exponential distribution $(Q(u) = -\log(1-u))$ and $u_i = (i/k+1)^{3/2}$ for $F(x) = x^2$ on $[0,1]$ $(Q(u)=u^{1/2})$. Whereas all three distributions satisfy the hypotheses of Theorem 2 the latter two do not satisfy the continuity conditions on Q' imposed by Theorem 1, meaning we are not immediately justified in using (3.4) to compute optimal strata for random variables with these distributions. This problem will now be considered in more detail.

From Theorem 1, equations (3.4) will have a unique solution if Q and $Q' = 1/fQ$ are continuous with $1/fQ > 0$ on $[0,1]$ and $-\log fQ$ concave on $(0,1)$. The latter two conditions are usually satisfied. However, for most laws $fQ(0) = fQ(1) = 0$ and Q is finite at 0 and 1 only for laws having a finite range. As a result, Q and Q' frequently will not satisfy the continuity conditions at 0 and 1. We now illustrate

how the approach of Section 2 can be modified to deal with such difficulties for certain types of laws. The arguments follow those by Barrow, et al (1978) and Chow (1982). The basic approach will be indicated here with the interested reader referred to either of these two papers for further details. Although the discussion which follows will be phrased in terms of approximation of Q the results will, of course, apply to TQ in general upon appropriate modification.

Now assume that Q is not piecewise constant for any k and is an element of $C^1(0,1)\cap L_2[0,1]$. It then follows from Chow (1982), or may be verified directly, that

$$Q(u_i) - (P_U Q)(u_i-) = (u_i-u_{i-1})\int_0^1 sQ'(s(u_i-u_{i-1})+u_{i-1})ds,\ i=2,\ldots,k \qquad (3.6)$$

and

$$Q(u_i)-(P_U Q)(u_i+) = -(u_{i+1}-u_i)\int_0^1 (1-s)Q'(s(u_{i+1}-u_i)+u_i)ds, i=1,\ldots,k-1, \qquad (3.7)$$

It is now assumed that (3.6) and (3.7) are well defined at i = 1 and k respectively. Note that this allows for cases such as $Q(u) = u^{1/2}$ and $-\log(1-u)$. Using the local nature of piecewise constant approximation, it is easily shown by differentiating the error functional on the subintervals (u_{i-1},u_{i+1}), that for optimal U

$$|Q(u_i) - (P_U Q)(u_i-)| = |Q(u_i) - (P_U Q)(u_i+)|,\ i = 1,\ldots,k. \qquad (3.8)$$

If Q' > 0 on (0,1) we may use (3.6) and (3.7) to rewrite (3.8) as

$$S_i(U) = (u_i-u_{i-1})\int_0^1 sQ'(s(u_i-u_{i-1}) + u_{i-1})ds \qquad (3.9)$$

$$-(u_{i+1}-u_i)\int_0^1 (1-s)Q'(s(u_{i+1}-u_i)+u_i)ds = 0, i=1,\ldots,k,$$

which is precisely (3.4). Consequently, the necessary conditions (3.4) still hold under these weaker assumptions. If, in addition, the Jacobian

matrix of $S(U) = (S_1(U),\ldots,S_k(U))$ is positive definite then arguments in Section 4 of Barrow, et al (1978) may be used to show that (3.9) has a unique solution. In particular, it follows from results in Section 3 of Barrow, et al (1978) that <u>if logQ' is concave the solution to</u> (3.9) <u>is unique provided</u> $\Sigma_{i=1}^{k} \partial S_1/\partial u_j$ <u>and</u> $\Sigma_{j=1}^{k} \partial S_k/\partial u_j$ <u>are positive</u>. To illustrate the use of this result let $Q(u) = u^{1/2}$ or logu and observe that we need only show $\Sigma_{j=1}^{k} \partial S_1/\partial u_j > 0$ as both functions are continuously differentiable at 1. From (3.9) we have

$$\Sigma_{j=1}^{k} \partial S_1/\partial u_j = \int_0^1 sQ'(su_1)ds + u_1\int_0^1 s^2Q''(su_1)ds$$
$$- (u_2-u_1)\int_0^1 (1-s)Q''(s(u_2-u_1) + u_1)ds$$

which is found to be positive for both $u^{1/2}$ and logu. Thus, there exists unique optimal strata boundaries for the distribution $F(x)=x^2$ and, from symmetry considerations, for the exponential distribution as well.

The approximate solutions obtained from h in (3.5) (and others that are asymptotically equivalent) have also been studied by Ekman (1960, 1963, 1969) and Särndal (1961, 1962). Ekman (1963, pg. 78) imposes the conditions that f' and f'' exist and are continuous over any finite interval and that $f_1(z) = z^k f(z^{-1})$ exists for some k>3 for which f_1' and f_1'' also exist and are finite for some z in a neighborhood of 0 and $0 < f_1 < \infty$. Although comparison is somewhat difficult these conditions seem more restrictive and more difficult to check, for most laws, than the conditions on Q required in Theorem 2. More immediate comparisons can be made with Särndal (1961) who requires Q to have four bounded continuous derivatives. It is also of interest to note that Ekman (1963) shows that, for optimal strata boundaries, $\lim_{k\to\infty} k^2 \inf_{U\in D_k} NV(\bar{X}) = [\int_0^1 Q'(u)^{2/3}du]^3/12$, provided f is ccntinuous and Q

is square integrable. Quantile based conditions for this result to hold, such as $Q \in L_2[0,1]$ with Q' integrable, can be obtained from Theorem 1.1 of Burchard and Hale (1975).

Under optimal (or Neyman) allocation the variance of $\bar{X}$ is not (3.3) but rather $\Sigma_{i=1}^{k+1} (u_i - u_{i-1})\sigma_i$. Approximate solutions to the optimal stratification problem in this case, similar to those discussed previously, have been studied by Dalenius and Hodges (1957, 1959), Ekman (1959a, b, c, 1960, 1963), Sethi (1963) and others. They use stratification points that are selected (or are asymptotically equivalent to those selected) from the density proportional to $f(x)^{1/2}$. Making the change of variable $X=Q(u)$, this is recognized as equivalent to selecting spacings according to $h(u) = Q'(u)^{1/2} / \int_0^1 Q'(s)^{1/2} ds$ which is the same density one would use in knot selection for piecewise constant $L_1[0,1]$ approximation of Q (c.f. Pence and Smith (1982)). This has the interesting consequence of establishing an asymptotic equivalence between variable knot $L_1[0,1]$ piecewise constant approximation of Q and optimal strata selection under Neyman allocation.

Several other authors have considered problems that are formally equivalent to the problem of optimal stratification with proportional allocation. A problem of grouping to "minimize loss of information" considered by Cox (1957) utilizes a loss function whose expectation is proportional to (3.3) and a "mixing problem" considered by Ekman (1969) can also be formulated as minimization of $||Q-P_U Q||^2$ in a particular instance. Under certain restrictions (see Chan and Adatia (1981)) a three group regression problem discussed by Gibson and Jowett (1975) provides an estimate of a regression coefficient whose variance is proportional to $[\Sigma_{i=1}^{3} (u_i - u_{i-1}) m_i^2]^{-1} = [||Q||^2 - ||Q-P_U Q||^2]^{-1}$.

Consequently, the results presented in this section are directly applicable to all these problems.

A grouping and combining problem posed by Rade (1963) can also be formulated as piecewise constant approximation of Q. Given an additive quality variable X with zero mean and symmetric density f, and a grouping $-\infty = x_{-(k+1)} < \ldots < x_{-1} < x_0 < x_1 < \ldots < x_{k+1} = \infty$, where $x_0 = 0$ and $x_{-i} = -x_i$, an observation on X that falls between x_{-i} and $x_{-(i-1)}$ is paired with one from the interval (x_{i-1}, x_i). The objective is to choose a grouping that maximizes the proportional increase in variability from pairing values at random over that for the grouped pairing scheme. This proportional increase in variability is shown to be $\Sigma_{i=1}^{k+1}(u_i - u_{i-1})m_i^2$ from which we see that the problem is equivalent to optimal knot selection for piecewise constant $L_2[.5,1]$ approximation of Q. The results of this section are now applicable after the obvious modifications..

4. Optimal grouping and spacing. The problems considered in this section can all be formulated as piecewise constant approximation of fQ or the product of fQ and Q, fQ·Q. We begin by considering a problem of optimal quantile selection for location or scale parameter estimation.

Let $X_1, \ldots, X_N$ denote a random sample from a distribution of the form $F(\frac{x-\mu}{\sigma})$ where μ and σ are respectively location and scale parameters and F is a known distributional form with associated p.d.f. f and q.f. Q. Define the sample quantile function by

$$\tilde{Q}(u) = X_{(j)}, \quad \frac{j-1}{N} < u \leq \frac{j}{N}, \; j = 1, \ldots, N, \tag{4.1}$$

where $X_{(j)}$ is the jth sample order statistic. It is frequently convenient to estimate μ or σ by linear functions of $k < N$ sample

quantiles. For a given $U \varepsilon D_k$ such estimators have the form $b_0 + \Sigma_{i=1}^{k} b_i \tilde{Q}(u_i)$ where explicit formulae for asymptotically (as $N \to \infty$) optimal weights have been given by Ogawa (1951). The estimators of μ and σ that result from Ogawa's weights are called the asymptotically best linear unbiased estimators (ABLUE's) and will be denoted here by $\mu(U)$ and $\sigma(U)$. When σ is known, $\mu(U)$ has asymptotic relative Fisher efficiency (ARE)

$$ARE(\mu(U)) = I(\mu)^{-1} \Sigma_{i=1}^{k+1} [fQ(u_i) - fQ(u_{i-1})]^2/(u_i - u_{i-1}) \qquad (4.2)$$

where $I(\mu) = \int_0^1 [(fQ)'(u)]^2 du$ and we assume that $fQ(0) = fQ(1) = 0$. Similarly, when μ is known and $fQ(0)Q(0) = fQ(1)Q(1) = 0$,

$$ARE(\sigma(U)) = I(\sigma)^{-1} \Sigma_{i=1}^{k+1} [fQ(u_i)Q(u_i) - fQ(u_{i-1})Q(u_{i-1})]^2/(u_i - u_{i-1}) \qquad (4.3)$$

where $I(\sigma) = \int_0^1 [(fQ \cdot Q)'(u)]^2 du$. The ARE's of both estimators are functions of U and, consequently, U should be chosen to maximize one of (4.2) or (4.3) thereby obtaining a best k-quantile subset for estimating the parameter of interest. This problem of optimal spacing selection has received considerable attention in the literature (see Cheng (1975) and Eubank (1981) for references).

Maximizing (4.2) (or, equivalently, minimizing 1-ARE($\mu(U)$)) is seen to follow the pattern in Section 2 by taking

$$m_i = (u_i - u_{i-1})^{-1} \int_{u_{i-1}}^{u_i} (fQ)'(u) du$$

so that $TQ = (fQ)'$. For scale parameter estimation the analogous result follows with $TQ = (fQ \cdot Q)'$. Therefore, the problem of optimal spacing selection for $\mu(U)$ and $\sigma(U)$ is equivalent to optimal knot selection for piecewise constant $L_2[0,1]$ approximation of $(fQ)'$ and $(fQ \cdot Q)'$ respectively.

Equations (2.9), in this setting, have been utilized to compute

optimal spacings for a variety of distributions (c.f. Chan and Kabir (1969) and Cheng (1975)). A general approach to this problem, including a computational algorithm using (2.9)-(2.12), is discussed in Eubank, Smith and Smith (1982). The conditions for uniqueness of optimal spacings provided by Theorem 1 were given previously in Eubank (1981) and, for $\mu(U)$, require that $(fQ)'$ and $(fQ)''$ be continuous with $(fQ)''$ of one sign on $[0,1]$ and $\log(fQ)''$ (or $\log-(fQ)''$ as appropriate) concave on $(0,1)$. Results for $\sigma(U)$ follow similarly. We note that these restrictions can be weakened, as in Section 3, to deal with distributions such as the Weibull, $F(x) = 1 - \exp\{-x^{\nu}\}$, $x, \nu > 0$, for which $(fQ\cdot Q)''(u) = \nu(1-u)^{-1}$ and, hence, does not satisfy the stated continuity conditions. These uniqueness conditions are to be compared with those imposed by Rhodin (1976) who requires that fQ and $fQ\cdot Q$ have three continuous derivatives and also satisfy a concavity condition. As an approximate solution one may use spacings selected according to the densities

$$h(u) = \begin{cases} |(fQ)''(u)|^{2/3} / \int_0^1 |(fQ)''(s)|^{2/3} ds \ , & \sigma \text{ known}, \\ |(fQ\cdot Q)''(u)|^{2/3} / \int_0^1 |(fQ\cdot Q)''(s)|^{2/3} ds, & \mu \text{ known}, \end{cases} \tag{4.4}$$

examples of which can be found in Eubank (1981). These densities were also proposed by Särndal (1961, 1962) under the condition that fQ and $fQ\cdot Q$ have four continuous derivatives.

Now suppose that one has two random samples $Z_1,\ldots,Z_n$ and $Y_1,\ldots,Y_m$, with d.f.'s F and G respectively, and wishes to test the hypothesis $G(x) = F(x)$ against the alternative $G(x) = F(x-\mu)$.

If $X_{(1)},\ldots,X_{(N)}$ $(N = n+m)$ denotes the combined ordered sample and $\tilde{Q}$ is defined as in (4.1), this hypothesis may be tested using a rank test based on a statistic of the form

$$R_N = \int_0^1 J(u)\delta(\tilde{Q}(u))du$$

where $J(u) = c_{jN}$, $\frac{j-1}{N} < u \leq \frac{j}{N}$, and $\delta(\tilde{Q}(u)) = 1$, if $\tilde{Q}(u)$ is a Z observation and is zero otherwise. Gastwirth (1966) shows how J may be chosen to obtain the asymptotically most powerful rank test (a.m.p.r.t.) and, given a spacing $U\epsilon D_k$, also considers group rank tests based on statistics of the form

$$R_N(U) = \Sigma_{j=1}^{k+1} c_j \int_{u_{j-1}}^{u_j} \delta(\tilde{Q}(u))du.$$

It is then shown that, for optimal c_j, the ARE of the resulting asymptotically most powerful group rank test (a.m.p.g.r.t.) to the a.m.p.r.t. is precisely (4.2). For testing $G(x) = F(x)$ against the alternative $G(x) = F(x/\sigma)$ the analogous result is that the asymptotic efficiency of the a.m.p.g.r.t. relative to the a.m.p.r.t. is (4.3). Thus previous comments on optimal spacing selection for $\mu(U)$ and $\sigma(U)$ including conditions for uniqueness, the computational algorithm in Eubank, Smith and Smith (1982) and the densities (4.4) apply to the problem of optimal group selection for the a.m.p.g.r.t. as well.

Given a grouping $a = x_0 < x_1 < \ldots < x_{k+1} = b$, Kulldorff (1958a, b, 1961) considered the problem of maximum likelihood estimation of a parameter, θ, when the available information from a random sample of size N consists only of the number of values falling in each interval $(x_{i-1}, x_i]$, $i = 1,\ldots,k+1$. Let $F(x;\theta)$, denote the common

d.f. for the sample elements with associated p.d.f., q.f. and density-quantile function $f(x;\theta)$, $Q(u;\theta)$ and $fQ(u;\theta) = f(Q(u;\theta);\theta)$. Then, under regularity conditions, it is shown that the asymptotic (as $N\to\infty$) variance of the maximum likelihood estimator (MLE) is

$$V(\hat{\theta}) = N^{-1}\{\Sigma_{i=1}^{k+1} (u_i-u_{i-1})(\frac{\partial}{\partial\theta} \log(u_i-u_{i-1}))^2\}^{-1} \tag{4.5}$$

where $u_i = F(x_i;\theta)$. Now

$$\frac{\partial}{\partial\theta}\log(u_i-u_{i-1}) = -(u_i-u_{i-1})^{-1}\left[fQ(u_i;\theta)\frac{\partial Q(u_i;\theta)}{\partial\theta} - fQ(u_{i-1};\theta)\frac{\partial Q(u_{i-1};\theta)}{\partial\theta}\right],$$

which follows from the identity $\left.\frac{\partial F(x;\theta)}{\partial\theta}\right|_{x=Q(u;\theta)} = -fQ(u;\theta)\frac{\partial Q(u;\theta)}{\partial\theta}$,

so Theorems 1 and 2 are applicable with $TQ(u) = \frac{\partial}{\partial u}[fQ(u;\theta)\frac{\partial Q(u;\theta)}{\partial\theta}]$. When θ is a location or scale parameter (4.5) is, apart from constant multiples, identical to (4.2) and (4.3) respectively so that selecting optimal spacings for the ABLUE's and MLE's of μ and σ are equivalent problems. However, in the latter case the x_i's must also be determined, which requires knowledge of θ. Kulldorff (1958a, b, 1961) has investigated the solutions to equations (2.9) for the normal and exponential distributions and found that, in these cases, $V(\hat{\theta})$ behaves somewhat robustly with respect to incorrect guesses for θ.

An insightful paper by Adatia and Chan (1981) investigates the question of when the problems of optimal quantile selection for the ABLUE, optimal stratification with proportional allocation and optimal grouping for the MLE's of μ and σ are equivalent. It now follows from the work in Section 2 that these problems are equivalent if we are approximating, in each case, linear combinations of

the same function. For instance, for location parameter estimation these three problems are equivalent if

$$(fQ)'(u) = c + dQ(u). \tag{4.6}$$

For scale parameter estimation the analogous condition is

$$(fQ \cdot Q)'(u) = c + dQ(u). \tag{4.7}$$

Conditions (4.6) and (4.7) are the quantile domain version of the principal condition in Theorem 5 of Adatia and Chan (1981) (they also provide conditions under which (4.6) and (4.7) are both necessary and sufficient). If one considers location parameter estimation for distributions having support on the entire real line (4.6) gives a differential equation in f (namely $f' - (c + dx) f = 0$) for which the normal distribution is the only solution. Similarly, for scale parameter estimation and distributions having support on $(0,\infty)$ the only solution to (4.7) is the gamma family of distributions. In particular, it follows from this that all the problems considered, up to this point, are equivalent in the special case of a normal or gamma distribution. This result will also be found to hold in the remaining section.

Other problems, related to those in this section, have been considered by Ogawa (1952), McClure (1980a,b), Koutrouvellis (1981) and Saleh (1981). There is also a relationship between the problem of optimal quantile selection for the ABLUE's and regression design for time series with Brownian motion or Brownian bridge errors that is explored in Eubank (1981) and Eubank, Smith and Smith (1982).

5. Other applications. In this final section several other problems are considered some of which have a bivariate nature. We begin with another stratification problem.

5.1 Optimal stratification. In sampling, the variable that is used for the purpose of stratification usually differs from the response variable. Let X denote the variable, having p.d.f. and q.f. denoted f and Q, upon which we intend to stratify. Assuming that X is related to the response variable Y by

$$Y = \mu_Y(X) + \varepsilon \,, \tag{5.1}$$

where ε is a zero mean random variable that is independent of X, the problem we now consider is how to select strata boundaries, $a = x_0 < x_1 < \ldots < x_{k+1} = b$, which minimize the variance of $\bar{Y}$, the mean response from a sample of size N selected with proportional allocation.

Let $\mu_Y Q(u) = \mu_Y(Q(u))$ and define

$$m_i = (u_i - u_{i-1})^{-1} \int_{u_{i-1}}^{u_i} \mu_Y Q(u)\,du \tag{5.2}$$

where u_i is given by (2.1). The variance of $\bar{Y}$ is then readily verified to be

$$\begin{aligned} V(\bar{Y}) &= N^{-1}[\sigma_\varepsilon^2 + \int_0^1 \mu_Y Q(u)^2 du - \Sigma_{i=1}^{k+1}(u_i - u_{i-1})m_i^2] \\ &= N^{-1}[\sigma_\varepsilon^2 + ||\mu_Y Q - P_U(\mu_Y Q)||^2] \end{aligned} \tag{5.3}$$

where σ_ε^2 is the variance of ε. Consequently, the problem of optimal strata selection under model (5.1) is equivalent to free knot piecewise constant approximation of $TQ = \mu_Y Q$. Under the conditions of Theorem 1 a U satisfying a necessary condition for optimality can be obtained as a solution to the equations

$$2\mu_Y Q(u_i) - m_i - m_{i+1} = 0 \,, \quad i = 1,\ldots,k, \tag{5.4}$$

which have also been considered by Dalenius and Gurney (1951) and Herlekar (1967). We now observe that their solution is unique if $\log(\mu_Y Q)'$ is concave. An approximate solution is provided by the

density proportional to $|(\mu_Y Q)'(u)|^{2/3}$.

In the event that $\mu_Y Q(u) = c + dQ(u)$, i.e., Y has a linear regression on X, it follows immediately from comments in Section 2 that the problem of strata selection reduces to the problem of approximating Q treated in Section 3.

A similar problem that concerns optimal grouping and combining has been considered by Rade (1963). The problem is essentially the same as the one discussed in Section 3 except that now the grouping is to be performed on an auxiliary variable X which is correlated with the quality variable Y. The selection of optimal groupings, in this case, is found to be a best $L_2[.5, 1]$ approximation problem for the "conditional mean", $\mu_Y Q$, which parallels the result obtained in Section 3 for the one variable case. We note in passing that the grouping problem of Cox (1957) and the "mixing problem" considered by Ekman (1969) have bivariate extensions that can also be analyzed using the techniques presented here.

5.2 Optimal chi-squared test for homogeneity. Let X be a random variable having p.d.f. f and q.f. Q. For a continuous density, g, Pearson's ϕ^2 is defined by

$$1 + \phi^2 = \int \left(\frac{g(x)}{f(x)}\right)^2 f(x)\,dx = \int_0^1 \left(\frac{gQ(u)}{fQ(u)}\right)^2 du \qquad (5.5)$$

where integration is over the range of X and $gQ(u) = g(Q(u))$. We assume that (5.5) is finite and note that ϕ^2 provides a measure of the distance between f and g. If the range of X is now partitioned into contiguous subintervals having boundaries $a = x_0 < x_1 < \ldots < x_{k+1} = b$ it then follows from Lancaster (1969, pg. 86) or Bofinger (1975) that the resulting grouped ϕ^2 can be written as

$$1 + \phi_U^2 = \Sigma_{i=1}^{k+1} (u_i - u_{i-1}) m_i^2 \tag{5.6}$$

where the u_i are defined by (2.1) and

$$m_i = (u_i - u_{i-1})^{-1} \int_{u_{i-1}}^{u_i} \frac{gQ(u)}{fQ(u)}\, du. \tag{5.7}$$

To obtain an optimal grouped distance measure the spacing, U, should be chosen to minimize

$$\phi^2 - \phi_U^2 = \int_0^1 \left(\frac{gQ(u)}{fQ(u)}\right)^2 du - \Sigma_{i=1}^{k+1} (u_i - u_{i-1}) m_i^2 \tag{5.8}$$

$$= \left|\left| \frac{gQ}{fQ} - P_U\left(\frac{gQ}{fQ}\right) \right|\right|^2 .$$

Bofinger (1975) also notes that a spacing selected to minimize (5.8) will, under certain conditions, maximize the non-centrality parameter of a chi-squared test for the equality of the distributions corresponding to f and g, thereby providing a best chi-squared homogeneity test. By taking TQ = gQ/fQ, optimal and asymptotically (as $k \to \infty$) optimal groupings for ϕ_U^2 can now be obtained using the results in Section 2.

If $g(x) = f(x;\theta)$ and $f(x) = f(x;\theta_0)$ for θ close to θ_0 then, with notation as in Section 4, we may use the approximation (see Lancaster (1969, pg. 89) or Bofinger (1975))

$$gQ(u;\theta)/fQ(u;\theta_0) \doteq 1 - (\theta - \theta_0)\frac{\partial}{\partial u}\left[fQ(u;\theta_0) \left.\frac{\partial Q(u;\theta)}{\partial \theta}\right|_{\theta=\theta_0}\right] . \tag{5.9}$$

The minimization of $\phi^2 - \phi_U^2$ then reverts to the problem of approximating $\frac{\partial}{\partial u}[fQ(u;\theta) \frac{\partial Q(u;\theta)}{\partial \theta}]$ for $\theta = \theta_0$ that was previously considered in Section 4. In the case of θ a location or scale parameter and f a normal or gamma density, previous comments regarding problem equivalences now also extend, approximately, to this setting.

5.3 Optimal grouping for bivariate distributions. Let (X,Y) denote a continuous bivariate random variable with joint p.d.f. $\ell(x,y)$

and marginal densities f and g, for X and Y respectively, that are assumed to satisfy $\iint[\ell(x,y)/f(x)g(y)]\,dxdy < \infty$. Also, let $\xi(X)$ and $\eta(Y)$ denote the first canonical variables of the X and Y space (c.f. Lancaster (1969, Chap. VI)) that correspond to the first (i.e., largest) canonical correlation, ρ. If X is grouped as in previous sections, the resulting first canonical variable for the new grouped X space was shown by Bofinger (1970) to be

$$\xi_U(X) = m_i/[\Sigma_{j=1}^{k+1}(u_j-u_{j-1})m_j^2] \tag{5.10}$$

where

$$m_i = (u_i-u_{i-1})^{-1}\int_{u_{i-1}}^{u_i}(\xi Q)(u)du \tag{5.11}$$

and Q is the q.f. for X. The correlation between $\xi_U(X)$ and $\eta(Y)$ was then shown to be

$$\rho_U = \rho[\Sigma_{i=1}^{k+1}(u_i-u_{i-1})m_i^2]^{1/2} \quad . \tag{5.12}$$

One method of optimally grouping one of the variables in a bivariate distribution, considered by Bofinger (1970), is to choose a spacing that maximizes ρ_U. In view of (5.12), this problem is now recognized as equivalent to optimal knot selection for piecewise constant approximation of $TQ = \xi Q$. Consequently, for ξQ and $(\xi Q)'$ continuous with $(\xi Q)'>0$ on [0,1] a $U\varepsilon D_k$ satisfying a necessary condition for optimality can be computed by solving the system of equations

$$2\xi Q(u_i) - m_i - m_{i+1} = 0, \quad i = 1,\ldots,k,$$

that was also derived by Bofinger (1970). As an approximate solution one may instead use spacings selected according to

$$h(u) = |(\xi Q)'(u)|^{2/3}/\int_0^1|(\xi Q)'(s)|^{2/3}ds. \tag{5.13}$$

For a standardized bivariate distribution the canonical variables are Hermite-Chebycheff polynomials (Eagleson (1964)) with $\xi(X) = X$

so that $\xi Q = Q$. Consequently, for the normal distribution the problems of optimal startification, optimal quantile selection (for $\mu(U)$), optimal grouping for the MLE of μ, optimal grouping for ρ_{UV}, etc. are all equivalent. As a result the optimal groupings and spacings for all these problems can be found in Kulldorff (1963) for $k = 1(1)10$. The asymptotically optimal spacing given by (5.13) is found to be $u_i = \Phi(\sqrt{3}\ \Phi^{-1}(i/k+1))$, where Φ is the standard normal d.f., with corresponding grouping $\sqrt{3}\ \Phi^{-1}(i/k+1)$, both of which are easily computed from tables of the standard normal. There are also bivariate gamma distributions having polynomial canonical variables (Kibble (1941), Eagleson (1964)) so that similar comments regarding the equivalence of previous problems obtain for these laws. In this instance optimal spacings have been computed by Rhodin (1975) for $k = 1(1)10$ and shape parameter values $\nu = 2(1)10$. Asymptotically optimal spacings obtained using (5.13) have been given by Särndal (1964) for $k = 1(1)10$ and $\nu = 2(1)5$.

In the case when both margins (i.e. both X and Y) are grouped, Bofinger (1970, 1975) proposed an approximate solution that, in our formulation, is equivalent to finding best free knot approximants to ξQ and ηQ_Y separately where Q_Y is the q.f. for Y. This problem is, therefore, also amenable to analysis by the techniques presented in this section.

Acknowledgement. The author would like to express his appreciation to Professors W. R. Schucany and P. W. Smith for several helpful discussions during the preparation of this manuscript.

REFERENCES

[1] Adatia, A. and Chan, L. K. (1981). Relations between stratified, grouped and selected order statistics samples. Scand. Acatuarial J., 193-202.

[2] Barrar, R.B. and Loeb, H.L. (1970). Existence of best spline approximations with free knots. J. Math. Anal. Appl. 31, 383-390.

[3] Barrow, D. L., Chui, C.K., Smith, P.W. and Ward, J.D. (1978). Unicity of best mean approximation by second order splines with variable knots. Math. Comp. 32, 1131-1143.

[4] Bofinger, E. (1970). Maximizing the correlation of grouped observations. J. Amer. Statist. Assoc. 65, 1632-1638.

[5] Bofinger, E. (1975). Optimal condensation of distributions and optimal spacing of order statistics. J. Amer. Statist. Assoc. 70, 151-154.

[6] Buhler, W. and Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. Metrika 22, 161-175.

[7] Burchard, H.G. and Hale, D.F. (1975). Piecewise polynomial approximation on optimal meshes. J. Aprox. Theor. 14, 128-147.

[8] Chan, L. K. and Kabir, A.B.M.L. (1969). Optimum quantiles for the linear estimation of the parameters of the extreme value distribution in complete and censored samples. Naval Res. Logistics Quart. 16, 381-404.

[9] Cheng, S.W. (1975). A unified approach to choosing optimum quantiles for the ABLE's. J. Amer. Statist. Assoc. 70, 155-159.

[10] Chow, J. (1982). Uniqueness of best $L_2[0,1]$ approximation by piecewise polynomials with variable breakpoints. Math. Comp.,to appear.

[11] Cox, D.R. (1957). Note on grouping. J. Amer. Statist. Assoc. 52, 543-547.

[12] Dalenius, T. (1950). The problem of optimal stratification. Skand. Aktuarietids. 33, 203-213.

[13] Dalenius, T. and Gurney, M. (1951). The problem of optimum startification II. Skand. Aktuarietids. 34, 133-148.

[14] Dalenius, T. and Hodges, J. L. (1957). The choice of stratification points. Skand. Aktuarietids. 40, 198-203.

[15] Dalenius, T. and Hodges, J.L. (1959). Minimum variance stratification. J. Amer. Statist. Assoc. 54, 88-101.

[16] Eagleson, G.K. (1964). Polynomial expansions of bivariate distributions. Ann. Math. Statist. 35, 1208-1215.

[17] Ekman, G. (1959a). An approximation useful in univariate stratification. Ann. Math. Statist. 30, 219-229.

[18] Ekman, G. (1959b). Approximate expressions for the conditional mean and variance over small intervals of a continuous distribution. Ann. Math. Statist. 30, 1131-1134.

[19] Ekman, G. (1959c). A limit theorem in connection with stratified sampling, part I. Skand. Aktuarietids. 42, 208-223.

[20] Ekman, G. (1960). A limit theorem in connection with stratified sampling, part II. Skand. Aktuarietids. 43, 1-26.

[21] Ekman, G. (1963). On the sum $\sum^{n} P_h^i \sigma_h^j$. Review Internat. Statist. Institute 31, 67-80.

[22] Ekman, G. (1969). A graphical solution for the determination of an optimal stratification or grouping, with an example concerning a problem of maximizing sales revenues. Review Internat. Statist. Institute 37, 186-193.

[23] Eubank, R. L. (1981). A density-quantile function approach to optimal spacing selection. Ann. Statist. 9, 494-500.

[24] Eubank, R. L., Smith, P. L. and Smith, P. W. (1981). Uniqueness and eventual uniqueness of optimal designs in some time series models. Ann. Statist. 9, 486-493.

[25] Eubank, R. L., Smith, P. L. and Smith, P. W. (1982). On the computation of optimal designs for certain time series models with applications to optimal quantile selection for location or scale parameter estimation. SIAM J. Sci. Statist. Comput. 3, 238-249.

[26] Gastwirth, J.L. (1966). On robust procedures. J. Amer. Statist. Assoc. 61, 929-948.

[27] Gibson, W.M. and Jowett, G.H. (1957). Three group regression analysis. Appl. Statist. 6, 114-122.

[28] Herlekar, R. K. (1967). The problem of optimum stratification, part I. Investigation into some two-stage stratified sampling procedures for populations represented by probability density functions. Skand. Aktuarietids. 47, 1-18.

[29] Kibble, W. F. (1941). A two-variate gamma type distribution. Sankhyā 5, 137-150.

[30] Koutrouvelis, I.A. (1981). Large-sample quantile estimation in Pareto laws. Commun. Statist. - Theor. Meth. A10(2), 189-201.

[31] Kulldorff,G. (1958a). Maximum likelihood estimation of the mean of a normal random variable when the sample is grouped. Skand. Aktuarietids. 41, 1-17.

[32] Kulldorff,G. (1958b). Maximum likelihood estimation of the standard deviation of a normal random variable when the sample is grouped. Skand. Aktuarietids. 41, 18-36.

[33] Kulldorff,G. (1961). Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples. Almqvist and Wiksell: Stockholm.

[34] Kulldorff,G. (1963). On the optimum spacing of sample quantiles for a normal distribution, part I. Skand. Aktuarietids 46, 143-156.

[35] Lancaster, H.O. (1969). The Chi-Squared Distribution. John Wiley: New York.

[36] McClure, D.E. (1980a). Optimized grouping methods, part I. Statistisk Tidskrift 18, 101-110.

[37] McClure, D.E. (1980b). Optimized grouping methods, part II. Statistisk Tidskrift 18, 189-198.

[38] Ogawa, J. (1951). Contributions to the theory of systematic statistics, I. Osaka Math. J. 3, 175-213.

[39] Ogawa, J. (1952). Contributions to the theory of systematic statistics, II. Osaka Math. J. 4, 41-68.

[40] Pence, D.D. and Smith, P. W. (1982). Asymptotic properties of best $L_2[0,1]$ approximation by splines. SIAM J. Math. Anal. 13, 409-420.

[41] Rade, L. (1963). Grouping and combining: an operations research problem. Skand. Aktuarietids. 46, 56-69.

[42] Rhodin, L. (1975). On the optimum spacing of sample quantiles for estimating the location and scale parameters of a gamma distribution. Statist. Res. Rep. No. 1975-11, Institute of Math. and Statist., Univ. of Umeå, Sweden.

[43] Rhodin, L. (1976). On the optimum spacing of sample quantiles for the ABLUE. Statist. Res. Rep. No. 1976-1, Institute of Math. and Statist., Univ. of Umeå, Sweden.

[44] Sacks. J. and Ylvisaker, D. (1968). Designs for regression problems with correlated errors; many parameters. Ann. Math. Statist. 39, 40-69.

[45] Saleh, A.K. Md. E. (1981). Estimating quantiles of exponential distribution. In Statistics and Related Topics, Csörgó, M., Dawson, D.A., Rao, J.N.K. and Saleh, A.K. Md.E. (Eds.), North Holland Publishing Co.

[46] Särndal, C.E. (1961). On a maximizing problem with several applications in statistical theory. Ark. Mat. 4, 385-392.

[47] Särndal, C.E. (1962). Information from Censored Samples. Almqvist and Wiksell: Stockholm.

[48] Särndal, C.E. (1964). Estimation of the parameters of the gamma distribution by sample quantiles. Technometrics 6, 405-414.

[49] Sethi, V.K. (1963). A note on optimum stratification of populations for estimating the population means. Austral. J. Statist. 5, 20-32.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 164	2. GOVT ACCESSION NO. AD-A117023	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) OPTIMAL GROUPING, SPACING, STRATIFICATION AND PIECEWISE CONSTANT APPROXIMATION		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER 164
7. AUTHOR(s) R. L. Eubank		8. CONTRACT OR GRANT NUMBER(s) N00014-82-K-0207
9. PERFORMING ORGANIZATION NAME AND ADDRESS Southern Methodist University Dallas, Texas 75275		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-479
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA 22217		12. REPORT DATE June 1982 13. NUMBER OF PAGES 26
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purposes of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Grouping; Spacing; Splines; Stratification

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A variety of statistical problems of optimal grouping, spacing and stratification are seen to be best $L_2[0,1]$ free knot piecewise constant approximation problems. This allows for the development of conditions that insure the existence and uniqueness of solutions, a computational algorithm and simple approximate solutions. In addition, this approach is seen to provide insight into the relationships between the various problems considered.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)