UNIQUE OPTIMAL PARTITIONS OF DISTRIBUTIONS AND CONNECTIONS TO HAZARD RATES AND STOCHASTIC ORDERING

David Mease and Vijayan N. Nair

San Jose State University and University of Michigan

Abstract: Optimal partitioning of a distribution arises in many contexts, including quantization in information theory, piecewise constant approximation of a function, stratified sampling, goodness-of-fit tests, principal points and clustering, and selective assembly in manufacturing. This article studies the behavior of optimal partitions, develops conditions under which the optimal partitioning of a distribution is unique, and establishes connections to hazard rate and likelihood ratio orderings of the distribution. An earlier proof which gives a slightly weaker condition than the sufficient condition in this article is shown to be incorrect by means of a counter-example. Optimal partitioning is compared with some heuristic partitioning strategies that are commonly used in applications and is shown to lead to substantial improvements in efficiency.

Key words and phrases: Likelihood ratio ordering, piecewise constant approximation, quantization, strongly unimodal.

1. Optimal Partitioning

The problem of optimally partitioning a distribution arises in many areas of application, some of which are described in Sections 2 and 5. The problem can be mathematically formulated as follows. Let X be a continuous random variable distributed according to the CDF F. The goal is to find an optimal partition of the support of F according to some criterion. Specifically let $d(X)$ be an M-level discretized version of X such that $d(X)$ takes the value d_m when X is in the interval $(x_{m-1}, x_m]$. Further, let $L(\cdot) \geq 0$ be a convex loss function such that $L(z) = 0$ if and only if $z = 0$. Consider the expected loss resulting from approximating X by its discretized version $d(X)$ given by

$$\mathrm{E}L(X - d(X)). \tag{1}$$

The problem is to find the values $x_1 < \cdots < x_{M-1}$ and $d_1, \ldots, d_M$ that minimize (1) for a given integer $M \geq 2$. The values x_0 and x_M are fixed to be the upper and lower endpoints of the support of X.

Let us first characterize the partitions that minimize (1). We assume throughout that X is a continuous random variable with density $f(x)$ that is strictly positive on (x_0, x_M) and zero elsewhere. For simplicity, we take x_0 and x_M to be finite although the case of $x_0 = -\infty$ and $x_M = \infty$ can be treated in exactly the same way. Define $p_m = F(x_m) - F(x_{m-1})$ and rewrite (1) as

$$\sum_{m=1}^{M} p_m \mathrm{E}L(X_m - d_m), \tag{2}$$

where X_m is the random variable X truncated to be in the $m-$th cell $(x_m, x_{m+1}]$. The following argument gives two sets of equations for the values of $(x_1, \ldots, x_{M-1})$ and $(d_1, \ldots, d_M)$ that minimize (2).

For a fixed set of partitions $x_1 < \cdots < x_{M-1}$, (2) is minimized by taking

$$d_m = U_m \equiv \arg\min_u \mathrm{E}L(X_m - u). \tag{3}$$

This is uniquely defined since L is convex. For example, for squared loss, U_m is the conditional mean while for absolute error loss, it is the conditional median. Note that U_m is a function of the partitions limits x_{m-1} and x_m as well as the distribution F. We suppress the dependence on F and denote this function by $u(\cdot, \cdot)$, i.e., $U_m = u(x_{m-1}, x_m)$.

Now for a fixed set of values $d_1 < \cdots < d_M$, the partitions $x_1, \ldots, x_{M-1}$ that minimize (2) can be obtained as the solutions of

$$L(x_m - d_m) = L(x_m - d_{m+1}). \tag{4}$$

For a symmetric L, this is just

$$x_m = \frac{d_m + d_{m+1}}{2}. \tag{5}$$

It follows that if $x_1 < \cdots < x_{M-1}$ and $d_1, \ldots, d_M$ minimize (2), they must satisfy (3) and (4) simultaneously.

One way of obtaining the solutions is to use a fixed-point iteration algorithm which starts with an initial set of values of $x_1 < \cdots < x_{M-1}$ and iterates between (3) and (4) until convergence. Lloyd's method (Trushkin (1982)) is such a technique used in the quantization literature.

If the equations (3) and (4) have a *unique* solution, then the algorithm is guaranteed to converge to the global minimum. However, this is not true in general, and it is easy to construct examples with multiple local minima (Tarpey (1994) and Mease, Nair and Sudjianto (2004)). Furthermore, the global minimum can exhibit unexpected behavior. For instance, the optimum partitions for symmetric distributions and a symmetric loss function can be asymmetric

(Tarpey (1994) and Mease, Nair and Sudjianto (2004)). In these cases, there exist symmetric partitions that also solve equations (3) and (4) but these yield local minima or maxima.

In this article, we study the nature of optimal partitions, develop conditions under which the optimum partitions are unique, and establish connections to two notions of stochastic ordering of distributions: hazard rate and likelihood ratio orderings. Specifically, we establish conditions when equations (3) and (4) have a unique solution and show that this is satisfied by distributions that have the likelihood ratio ordering property. As shown by Shaked and Shanthikumar (1994), likelihood ratio ordering is equivalent to the log-concavity of the density, but the former is precisely the representation needed to establish the uniqueness result. A practical implication of uniqueness is that any solution of (3) and (4) determines the global minimum for (1). In particular, optimal partitions of symmetric distributions with likelihood ratio ordering (strongly unimodal density) under symmetric loss can only be symmetric. In Section 6, we compare the efficiency of optimal paritions to two commonly used heuristic algorithms (equal widths and equal probabilities) under several loss functions and distributions.

2. Some Applications

The optimal partitioning problem considered here arises in many contexts, some of which are described below (see also Section 5). Most of the literature, however, focuses on squared loss.

A. Quantization in information theory

In information theory, the problem of finding $d_1, \ldots, d_M$ and $x_1, \ldots, x_{M-1}$ to minimize (1) arises as a problem in (lossy) data compression known as quantization. Among the earliest examples of this is analog-to-digital (A/D) conversion (Shannon (1948)). In this case, the random variable X would represent the analog signal requiring an infinite number of bits that is transformed into a digital signal $d(X)$ which requires a finite number of bits. The function $d(\cdot)$ is referred to as the *quantizer* and the loss function L is referred to as the *distortion measure*.

A number of popular current uses of quantization involve compression of visual images and other data types for more economical storage in computer memory and faster transmission over digital communication networks. In these applications, the input signal itself is often also digital, but with a much finer resolution than that of the output. An example would be compressing the colors in an image represented by floating point values in $(0, 1)$ to integers between 0 and 255 inclusive. The colors of the compressed image would require only $\log_2 256 = 8$ bits. The popular color image format developed by the Joint Photographic Experts Group (JPEG) employs quantization.

In this article we deal with the one-dimensional problem, i.e., the case where X is a random variable. This is known as *scalar quantization* in information theory. When X is instead a random vector, scalar quantization on each of the components is less efficient than what is known as *vector quantization*, although often the former is adequate. There is no known condition to ensure uniqueness in this multidimensional case analogous to the condition for one dimension that we present in this article. An extensive review of the literature on quantization can be found in Gray and Neuhoff (1998).

B. Piecewise constant function approximation

The problem can also be viewed as one of optimal breakpoint selection for piecewise constant function approximation in $L_2[0,1]$ (Eubank (1988)). Specifically, given a function g on $[0,1]$, the goal is to find constants $c_1, \ldots, c_M$ and breakpoints $0 = g_0 < \cdots < g_M = 1$ to minimize

$$\sum_{m=1}^{M} \int_{g_{i-1}}^{g_i} (g(u) - c_i)^2 du. \tag{6}$$

If the function $g(u)$ is strictly increasing, we can see that this is the same problem by letting $g = F^{-1}$ and minimizing (1) with squared loss (i.e. $L(z) = z^2$). Under squared loss, each U_m is equal to the conditional mean of X conditioned on the interval $(x_{m-1}, x_m]$ so that we can write (4) as

$$x_m = \frac{\mathrm{E}\,(X_m) + \mathrm{E}\,(X_{m+1})}{2}. \tag{7}$$

Eubank (1988) provides an excellent review of a number of statistical problems that are equivalent to finding the optimal piecewise constant approximation to g defined by minimizing (6). By taking g to be the inverse of a cumulative distribution function, the problem becomes equivalent to optimal selection of strata in stratified sampling under proportional allocation. The solution to this problem was first derived in Dalenius (1950), although the uniqueness issue was not discussed. Eubank (1988) also reviews problems of minimizing loss of information (Cox (1957)), discriminating between two p-variate normal distributions using qualitative variables (Cochran and Hopkins (1961)), determining optimal grouping for chi-squared tests, and identifying groupings for bivariate distributions.

C. Principal points and clustering

Tarpey and Flury (1996) consider the problem of minimizing (1) under squared loss when X is a p-dimensional random vector. Specifically, the problem is to

find M p-dimensional vectors $\mathbf{d}_1, \ldots, \mathbf{d}_M$ to minimize

$$\mathrm{E}\|\mathbf{X} - \mathbf{d}(\mathbf{X})\|^2, \tag{8}$$

where $\mathbf{X}$ is a random vector and the vector-valued function $\mathbf{d}(\mathbf{x})$ is defined by

$$\mathbf{d}(\mathbf{x}) \equiv \arg\min_{\mathbf{d}_i} \|\mathbf{x} - \mathbf{d}_i\|^2. \tag{9}$$

Vectors $\mathbf{d}_1, \ldots, \mathbf{d}_M$ that minimize (8) are referred to as *principal points*. Applications that involve minimizing (8) include optimal stratification, selection of shapes for masks to fit human faces, standardizing clothing, and $k-$means clustering. The minimizer of (8) also corresponds to finding the optimal vector quantizer for the random vector $\mathbf{X}$ in the information theory problem described earlier.

If we define the sets $S_i = \{\mathbf{x} : \mathbf{d}(\mathbf{x}) = \mathbf{d}_i\}$, $i = 1, \ldots, M$, then to minimize (8), we can restrict the vectors $\mathbf{d}_i$ to be

$$\mathbf{d}_i = \mathrm{E}\left[\mathbf{X}|\mathbf{X} \in S_i\right]. \tag{10}$$

This is analogous to the means conditioned on the intervals in the one-dimensional case. Tarpey and Flury (1996) refer to a set of vectors $\mathbf{d}_1, \ldots, \mathbf{d}_M$ for which both (9) and (10) hold as *self-consistent points*. While log-concavity is sufficient in the one dimensional case, for $p > 1$ there is no known condition to ensure that there is only one set of self-consistent points. (A unique set of self-consistent points must be the principal points.) Patterns of self-consistent points for symmetric multivariate distributions are discussed in Tarpey and Flury (1996).

3. Hazard Rate and Increasing Likelihood Ratio Orderings

We now review some concepts from reliability theory which will be helpful for proving the uniqueness result for the optimal partitioning. A proof of the uniqueness result will be given in the following section.

Given the random variable of interest X with CDF F and any constant A, define the truncated random variable $X_A \equiv [X - A|X \geq A]$. Suppose X (or equivalently F) satisfies the condition that X_A is stochastically larger than X_B (denoted as $X_A \succeq X_B$) for any $A < B$. Then X (or F) is said to have *hazard rate ordering*. This is a type of "stochastic dominance to the left" of the distribution F. It is known that F has hazard rate ordering if and only if F has an increasing (more specifically nondecreasing) hazard rate (IHR) (Shaked and Shantikumar (1994, Chap. 1)). Recall that the hazard rate for F with density f is given by $\lambda(x) = f(x)/[1-F(x)]$. While hazard rates are usually discussed for non-negative random variables only, we do not make that restriction here.

Consider now the doubly or interval-truncated random variable $X_A(h) \equiv [X-A|A < X \leq A+h]$ for given constants A and $h > 0$. Suppose $X_A(h) \succeq X_B(h)$ whenever $A < B$ for every $h > 0$. Then, X (or F) is said to have the *increasing likelihood ratio* (ILR) property (Shaked and Shantikumar (1994, Chap. 1)). This is clearly a stronger condition than hazard rate ordering which corresponds to the special case of $h = \infty$. The ILR property can be viewed as a notion of uniform stochastic dominance. It is known that a distribution has the ILR property if and only if its density is log-concave (Shaked and Shanthikumar (1994)).

Next consider the (doubly) conditional hazard rate of F given by

$$\lambda_b(x) = \lim_{\triangle x \downarrow 0} \frac{P(X \in (x, x + \triangle x] \mid x < X \leq x + b)}{\triangle x} = \frac{f(x)}{F(x + b) - F(x)}.$$

It can be shown that F satisfies the ILR property if and only if the conditional hazard rate is increasing (nondecreasing) in x for every $b > 0$. This conditional hazard rate characterizes the propensity for instantaneous failure given that the unit has survived until x and will fail in the interval $(x, x + b]$.

In reliability and survival analysis, the random variable $[X - x|X > x]$ corresponds to residual life given that the unit has survived up to time x. It is well known that the mean of the residual life, i.e., $E[X - x|X > x]$, is decreasing under the IHR condition. This property also holds for the median and other quantiles under IHR or, equivalently, the hazard rate ordering property. (Recall that the hazard rate ordering is defined as the stochastic ordering $X_A \succeq X_B$ for all $A < B$). The following proposition and the hazard rate ordering property show that an analogous result holds more generally for any convex loss function. In this proposition Y and Z have the roles of X_A and X_B respectively.

Proposition 1. *Given two random variables Y and Z, let $c_Y = \arg\min_c E[L(Y - c)]$ and $c_Z = \arg\min_c E[L(Z - c)]$. If $Y \succeq Z$, then $c_Y \geq c_Z$.*

Proof. Let F_Y^{-1} and F_Z^{-1} denote the inverse CDF's of Y and Z respectively. Note that c_Y and c_Z are unique since the functions $R_Y(c) = E[L(Y - c)]$ and $R_Z(c) = E[L(Z - c)]$ are strictly convex on $(F_Y^{-1}(0), F_Y^{-1}(1))$ and $(F_Z^{-1}(0), F_Z^{-1}(1))$ respectively, as a result of the assumptions (Trushkin (1982)). Fix $\epsilon > 0$. We have

$$\begin{aligned}
0 < R_Y(c_Y + \epsilon) - R_Y(c_Y) &= E[L(Y - c_Y - \epsilon) - L(Y - c_Y)] \\
&= \int_0^1 [L(F_Y^{-1}(u) - c_Y - \epsilon) - L(F_Y^{-1}(u) - c_Y)]du \\
&\leq \int_0^1 [L(F_Z^{-1}(u) - c_Y - \epsilon) - L(F_Z^{-1}(u) - c_Y)]du \\
&= E[L(Z - c_Y - \epsilon) - L(Z - c_Y)] = R_Z(c_Y + \epsilon) - R_Z(c_Y).
\end{aligned}$$

The second inequality follows from the fact that L is convex and that $F_Y^{-1}(u) \geq F_Z^{-1}(u)$ for all $u \in [0,1]$. Thus, $c_Z \leq c_Y + \epsilon$ for any $\epsilon > 0$, and hence Proposition 1 holds.

4. Uniqueness of Optimal Partitions

We are now ready to prove the uniqueness result using the concepts developed in the previous section. First we need the following lemma. Recall the definition of $u(x, y)$ following (3).

Lemma 1.
(I) $u(t, t+x)$ is increasing in x for all t.
(II) $u(t-x, t)$ is decreasing in x for all t.
(III) $u(t, t+h) - t$ is non-increasing in t for all $0 < h \leq \infty$ provided f is log-concave.

Proof. Results (I) and (II) follow from the fact that the density f is strictly positive and that the function $L(\cdot)$ is convex with $L(z) = 0$ if and only if $z = 0$. To show (III) define $X_A(h) = [X - A | A < X \leq A + h]$ and $X_B(h) = [X - B | B < X \leq B + h]$. Recall that log-concavity of f is equivalent to the ILR property which implies that $X_A(h) \succeq X_B(h)$ if $A < B$. Now (III) follows by applying Proposition 1 to $X_A(h)$ and $X_B(h)$.

We are now ready to establish the uniqueness result. Let F be the distribution of X with support (x_0, x_M), and f denote the density.
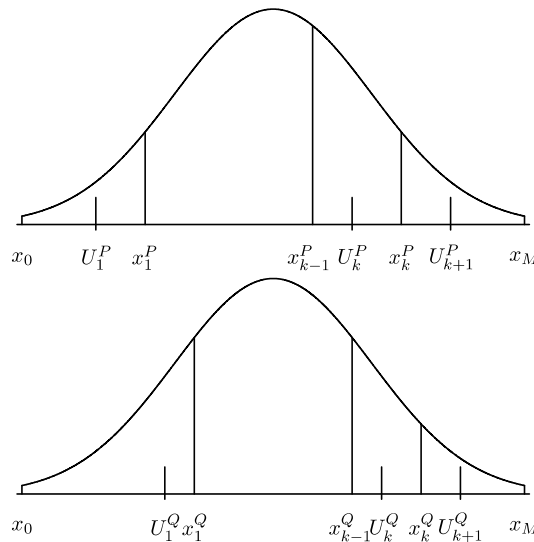

Figure 1. An Illustration for the Proof of Proposition 2 with $k = M - 1$.

Proposition 2. *If F has the ILR property (i.e., f is log-concave), then there is at most one solution to equations* (3) *and* (4).

Proof. Suppose there exist two solutions to (3) and (4) given by $(x_1^P, \ldots, x_{M-1}^P)$ and $(x_1^Q, \ldots, x_{M-1}^Q)$. Let the values of the U_m's corresponding to these two sets of partitions be denoted by $U_1^P, \ldots, U_M^P$ and $U_1^Q, \ldots, U_M^Q$, respectively. Now let k be the largest i such that $x_i^P \neq x_i^Q$. Without loss of generality we take $x_k^P < x_k^Q$. (It is useful to refer to Figure 1 for notation and the relative orderings of the various quantities discussed below). Then we have $U_{k+1}^Q - x_k^Q < U_{k+1}^P - x_k^P$ as a consequence of (III) and (I) of Lemma 1. From (4) we have $L(x_k^Q - U_{k+1}^Q) = L(x_k^Q - U_k^Q)$ and $L(x_k^P - U_{k+1}^P) = L(x_k^P - U_k^P)$, so that $x_k^Q - U_k^Q < x_k^P - U_k^P$. From this we can infer $x_{k-1}^P < x_{k-1}^Q$ since (III) and (II) of Lemma 1 give $x_k^P - x_{k-1}^P > x_k^Q - x_{k-1}^Q$. From this last inequality, we also have $U_k^Q - x_{k-1}^Q < U_k^P - x_{k-1}^P$ using (III) and (I). Now we can repeat the entire argument with k replaced by $k-1$ to obtain $x_{k-2}^P < x_{k-2}^Q$ and $U_{k-1}^Q - x_{k-2}^Q < U_{k-1}^P - x_{k-2}^P$. Continuing inductively, we eventually have $x_1^P < x_1^Q$ and $U_2^Q - x_1^Q < U_2^P - x_1^P$. Finally, (4) implies $L(x_1^Q - U_2^Q) = L(x_1^Q - U_1^Q)$ and $L(x_1^P - U_2^P) = L(x_1^P - U_1^P)$ so that $x_1^Q - U_1^Q < x_1^P - U_1^P$ which, along with $x_1^P < x_1^Q$, is a contradiction of (II) and (III).

As noted earlier, the uniqueness issue has been discussed previously in the literature. Tarpey (1994) established optimality results for the very specialized case of symmetric distributions with two partitions. Eubank (1988) argued that log-concavity of $f(F^{-1}(u))$ is a sufficient condition for uniqueness in the context of function approximation with piecewise constants. While this condition is weaker than log-concavity of $f(x)$, we construct a counter-example below to show that it is not sufficient. These authors considered only squared-error loss functions. Much earlier than these papers, Trushkin (1982, 1984) had studied uniqueness for symmetric convex loss functions. This work does not seem to be known in the statistical literature. Trushkin's proof, however, is much more involved than the one here, and it does not make the interesting connections to the stochastic ordering properties of the distribution. Our result also covers asymmetric loss functions. For instance, loss functions that yield quantiles as the minimizers are asymmetric (except for the median). Asymmetric loss functions arise naturally in some applications, such as selective assembly which will be discussed in Section 5.

Returning to the log-concavity condition for $f(F^{-1}(u))$ in Eubank (1988),

consider the density $f(x)$ proportional to the function

$$g(x) = \begin{cases} 0, & x < -10, \\ e^{\frac{x-0.5}{2.06}}, & -10 \le x < 0.5, \\ \frac{w(x)}{w(0.5)}, & 0.5 \le x \le 10, \\ 0, & x > 10. \end{cases}$$

Here $w(x)$ is the Weibull density function with a scale parameter of 1 and a shape parameter of $1/2$. While $f(F^{-1}(\cdot))$ is in fact log-concave, the solution is not unique. The expected squared loss for $M = 2$ is plotted in Figure 2 as a function of x_1. In this figure it can be seen that there is one local minimum and one local maximum of (1) in addition to the one global minimum. It can be verified that all of these correspond to solutions to (3) and (4), giving a total of three solutions and thus a counter-example for uniqueness.
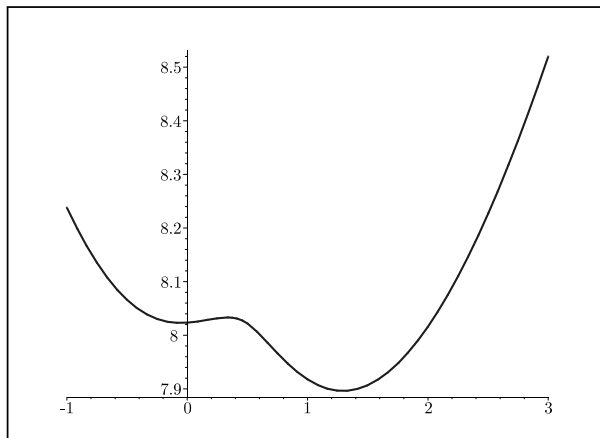

Figure 2. Expected Squared Loss as a Function of x_1.

5. Application to Selective Assembly In Manufacturing

Our interest in this problem was motivated by an application to selective assembly in manufacturing. This is a cost-effective approach that is commonly used to reduce variation (Mease, Nair and Sudjianto (2004), Kwon, Kim and Chandra (1999) and Pugh (1986)). The basic idea behind selective assembly is as follows. Let X and Y be the (random) dimensions of two mating components. For example, in a piston and cylinder assembly X could be the inside radius of the cylinder and Y could be the radius of the piston which fits inside the cylinder.

The clearance between the inside wall of the cylinder and outer surface of the piston is critical. Too large of a clearance leads to excessive vibration while too small of a clearance can cause excessive friction. If the quality loss is taken to be proportional to a loss function $L(\cdot)$ evaluated at the difference between the true clearance and a target clearance τ, then with random (i.e., non-selective) assembly the expected loss is $\mathrm{E}L(X - Y - \tau)$. Selective assembly reduces this expected loss by measuring and grouping one or both of the individual components (X and Y) prior to assembly. The final product is assembled by matching the two components from appropriate groups to minimize loss. In other words, small cylinders are matched with small pistons, and large cylinders are matched with large pistons.

In "one-sided" selective assembly, one of the components, say Y, is capable of being produced with negligible variation relative to the X component. Then, the procedure partitions the X distribution into M parts while the Y component is manufactured at (or close to) M corresponding nominal values. The problem then reduces to finding the optimal M nominal values $d_1, \ldots, d_M$ for Y (or more correctly for $Y + \tau$) and cut points $x_1 < \cdots < x_{M-1}$ that determine the partitions for X. Thus, this selective assembly problem is equivalent to the problem we are considering in minimizing (1).

With selective assembly, asymmetric loss functions are often appropriate. For instance, in the piston and cylinder example, a clearance below the target value results in different problems than a clearance which exceeds the target value. An asymmetric loss function can capture the relative quality loss associated with these two types of deviations from the target clearance.

While we described selective assembly only for dealing with differences $X - Y$, it can also be applied to fits in which the total of the two dimensions is the relevant measurement. For instance, if two pieces are stacked on top of one another, the quality would be related to $\mathrm{E}L(X + Y - \tau)$ instead of $\mathrm{E}\,L(X - Y - \tau)$ where τ is now the target height. Such problems are handled in our formulation by replacing Y with $-Y$.

In this discussion we have focussed our attention on the one-sided selective assembly problem, in which one of the two components is manufactured with negligible variation relative to the other. However, a two-sided problem in which the X and Y components are manufactured with a similar degree of variation is also quite common (Mease, Nair and Sudjianto (2004)) and Kwon, Kim and Chandra (1999)). In this problem, both the X and Y distributions are partitioned. While in general this problem is more complex, in the special case of squared loss there is an equivalence, as discussed in Mease, Nair and Sudjianto (2004)).

6. Comparisons of Optimal Partitions to Two Heuristic Partitioning Strategies

In this section, we compare the efficiency of optimal partitions with two heuristic partitioning schemes that are commonly used in engineering applications: equal probability and equal width. The comparisons are done for three different loss functions and two different distributions. For equal probability partitioning, which is also called equal area partitioning in selective assembly, the partitions are chosen such that all the p_i are equal. Equal width partitioning partitions the (bounded) support of the X distribution into intervals of equal width so that $x_{m+1} - x_m = (x_M - x_0)/M$ for all m. These two heuristic strategies are commonly used in the selective assembly application (Pugh (1986, 1992)).

Table 1 compares the expected squared loss ($L(z) = z^2$) resulting from the optimal partitions for the standard normal distribution truncated at $x_0 = -3$ and $x_M = 3$ to the expected squared loss using equal probability and equal width partitioning. Table 2 gives the same comparison for absolute error loss ($L(z) = |z|$). The improvement is quite substantial, especially for squared loss. For example, using the optimal partitioning for $M = 14$ or $M = 15$ results in an expected squared loss that is less than half of that using equal probability partitioning.

Table 1. Optimal partitions for standard normal distribution truncated at −3 and 3 under squared loss.

M	Optimal Partition Limits (Only the nonnegative values are given since all partitions are symmetric.)	Expected Squared Difference from Target	Percentage Savings Over Equal Probability Partitioning	Percentage Savings Over Equal Width Partitioning
1	-	0.973	-	-
2	0.000	0.347	0%	0%
3	0.604	0.179	7.9%	29.6%
4	0.000 0.964	0.109	15.5%	31.0%
5	0.375 1.215	0.073	22.0%	31.7%
6	0.000 0.643 1.405	0.052	27.4%	31.8%
7	0.273 0.850 1.555	0.039	32.1%	31.7%
8	0.000 0.486 1.017 1.677	0.031	36.0%	31.5%
9	0.215 0.659 1.154 1.779	0.025	39.4%	31.4%
10	0.000 0.391 0.804 1.271 1.866	0.020	42.4%	31.2%
11	0.177 0.539 0.928 1.372 1.940	0.017	45.0%	31.0%
12	0.000 0.327 0.667 1.035 1.460 2.005	0.014	47.3%	30.9%
13	0.151 0.457 0.778 1.130 1.538 2.063	0.012	49.4%	30.8%
14	0.000 0.281 0.570 0.877 1.214 1.607 2.113	0.010	51.2%	30.7%
15	0.131 0.396 0.671 0.965 1.289 1.669 2.159	0.009	52.9%	30.6%

Table 2. Optimal partitions for standard normal distribution truncated at
-3 and 3 under absolute error loss.

M	Optimal Partition Limits (Only the nonnegative values are given since all partitions are symmetric.)	Expected Absolute Difference from Target	Percentage Savings Over Equal Probability Partitioning	Percentage Savings Over Equal Width Partitioning
1	-	0.791	-	-
2	0.000	0.467	0%	0%
3	0.512	0.335	0.9%	19.8%
4	0.000 0.817	0.261	2.0%	21.6%
5	0.316 1.032	0.214	3.0%	22.7%
6	0.000 0.544 1.196	0.182	4.0%	23.1%
7	0.231 0.720 1.328	0.158	4.9%	23.2%
8	0.000 0.411 0.862 1.437	0.139	5.7%	23.3%
9	0.182 0.558 0.982 1.530	0.125	6.4%	23.2%
10	0.000 0.331 0.682 1.084 1.611	0.113	7.1%	23.1%
11	0.150 0.457 0.789 1.173 1.681	0.103	7.7%	23.0%
12	0.000 0.278 0.566 0.882 1.251 1.744	0.095	8.2%	22.9%
13	0.128 0.388 0.662 0.965 1.321 1.800	0.088	8.7%	22.9%
14	0.000 0.239 0.485 0.747 1.039 1.385 1.851	0.082	9.2%	22.8%
15	0.111 0.337 0.572 0.824 1.105 1.442 1.897	0.077	9.7%	22.7%

For these two examples, the intuition as to why the optimal partitions perform better than equal width is as follows. With a uniform distribution, equal width is equivalent to optimal partitioning. However, for a unimodal distribution like the normal the optimal partitions give a finer partitioning near the mode and less fine in the tails where fewer observations will fall. For any fixed number of partitions, this is a superior strategy. Equal probability partitioning also has finer partitions near the mode than in the tails, but actually too much so, at least in the case of the normal distribution.

As mentioned earlier, asymmetric loss functions are often appropriate for certain applications such as selective assembly. Table 3 gives the optimal partitions for this same truncated normal distribution but now using the asymmetric loss function

$$L(z) = \begin{cases} -z & z < 0 \\ \frac{z}{4} & z \geq 0. \end{cases}$$

Note that with this loss function the d_m are equal to the conditional 0.2 quantiles.

Finally Table 4 gives the optimal partition limits under squared loss for the Weibull distribution with a shape parameter of 2 and a scale parameter of 1 truncated at 6 (and 0). This provides an example of partitioning for an asymmetric distribution. Here the improvement over the two heuristic strategies of equal probability and equal width is even more pronounced.

Table 3. Optimal partitions for standard normal distribution truncated at
-3 and 3 under the asymmetric loss function $L(z) = -z$ for $z < 0$ and $z/4$
for $z \geq 0$.

M	Optimal Partition Limits	Expected Loss	Percentage Savings Over Equal Probability Partitioning	Percentage Savings Over Equal Width Partitioning
1	-	0.346	-	-
2	-0.204	0.198	1.4%	1.4%
3	-0.715 0.333	0.139	2.9%	19.7%
4	-1.014 -0.174 0.656	0.108	4.3%	21.6%
5	-1.222 -0.483 0.162 0.883	0.088	5.4%	22.6%
6	-1.379 -0.704 -0.147 0.403 1.056	0.074	6.5%	22.9%
7	-1.504 -0.874 -0.371 0.097 0.589 1.196	0.064	7.4%	23.0%

Table 4. Optimal partitions for a truncated Weibull distribution under
squared loss.

M	Optimal Partition Limits	Expected Squared Difference from Target	Percentage Savings Over Equal Probability Partitioning	Percentage Savings Over Equal Width Partitioning
1	-	0.215	-	-
2	0.972	0.073	5.9%	65.8%
3	0.716 1.297	0.037	15.5%	79.6%
4	0.581 1.004 1.504	0.022	23.8%	81.2%
5	0.495 0.838 1.199 1.655	0.015	30.5%	81.3%
6	0.434 0.726 1.019 1.346 1.773	0.011	36.1%	81.9%
7	0.389 0.646 0.895 1.159 1.464 1.870	0.008	40.7%	82.6%

References

Cochran, W. G. and Hopkins, C. E. (1961). Some classification problems with multivariate qualitative data. *Biometrics* **17**, 10-32.

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

Dalenius, T. (1950). The problem of optimal stratification. *Skand. Aktuarietids.* **33**, 203-212.

Eubank, R. L. (1988). Optimal grouping, spacing, stratification, and piecewise constant approximation. *SIAM Rev.* **30**, 404-420.

Gray, R. M. and Neuhoff, D. L. (1998). Quantization. *IEEE Trans. Inform. Theory* **44**, 2325 -2383.

Kwon, H.M., Kim, K.J. and Chandra, M.J. (1999). Economic selective assembly procedure. *Naval Res. Logist.* **46**, 809-821.

Mease, D., Nair, V. N. and Sudjianto, A. (2004). Selective assembly in manufacturing: Statistical issues and optimal binning strategies. *Technometrics* **46**, 165-175.

Pugh, G. A. (1986). Partitioning for selective assembly. *Computers and Industrial Engineering* **11**, 175-179.

Pugh, G. A. (1992). Selective assembly with components of dissimilar variance. *Comput. Indust. Engineering* **23**, 487-491.

Shaked, M. and Shanthikumar, J. G. (1994). *Stochastic Orders and Their Applications*. Academic Press, Boston.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379-423 and 623-656.

Tarpey, T. (1994). Two principal points of symmetric, strongly unimodal distributions. *Statist. Probab. Lett.* **20**, 253-257.

Tarpey, T. and Flury, B. (1996). Self-consistency: A fundamental concept in statistics. *Statist. Sci.* **11**, 229-243.

Trushkin, A. (1982). Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions. *IEEE Trans. Inform. Theory* **28**, 187-198.

Trushkin, A. (1984). Monotony of Lloyd's method II for log-concave density and convex error weighting function. *IEEE Trans. Inform. Theory* **30**, 380-383.

Department of Marketing and Decision Sciences, College of Business, San Jose State University, San Jose, CA 95192-0069, U.S.A.

E-mail: mease_d@cob.sjsu.edu

Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1092, U.S.A.

E-mail: vnn@umich.edu