HOW INFORMATIVE IS YOUR SEGMENTATION?

A simple new metric yields surprising results.

BY PIETER SHETH-VOSS AND ISMAEL E. CARRERAS

Segmentation is a common tool that helps organizations discern and describe key patterns in complex markets. In his article, "Rediscovering Market Segmentation" (*Harvard Business Review*, 2006), Daniel Yankelovich cites a finding that 59 percent of executives across industries reported a major segmentation exercise in the past 24 months. Yet segmentation is often perceived as among the most enigmatic of quantitative market research analyses. The same article later presents a case study stating, "Although the solution was mathematically sound, management did not trust its findings." Sentiments like this are commonly heard in practice.

Segmentation is distinct from analyses such as conjoint analysis, where different experienced practitioners would likely collect similar data, estimate similar models and yield similar solutions. By contrast, segmentations are often based on diverse information sources, complex algorithms and much iterative analysis. Practitioners advocate a variety of approaches that can be bewildering to clients. These include a diverse array of algorithms, such as latent class analysis (LCA), K-means clustering, CART/CHAID, hierarchical clustering, factor-cluster and ensemble models, among others. Furthermore, there are many philosophies about goals, bases and inputs to segmentation, including psychographic, behavioral, attitudinal, "human centric" and others. Different business goals entail different results. An "unmet-needs-based" segmentation for marketing communication will almost certainly yield a different solution than an "opportunity-based" segmentation for sales force targeting. A multivariate solution based on multiple data sources will differ from a 2x2 matrix posed by an industry veteran.

Regardless of the segmentation algorithm used, how do we know that a particular solution is good for the intended purpose? What do the mathematics actually do? Without a clear understanding of these algorithms, and absent objective criteria for evaluating and comparing solutions, segmentation retains its aura of mystery. The goal here is not to advocate any algorithm or approach, but rather merely to explore how different solutions can be compared.

A simple practical metric—net mutual information—can be used to compare segmentations "apples to apples," based on how much information they convey about a set of attributes. This can be applied to compare alternative segmentations on common terms.

Importantly, the comparison is independent of the means by which the solutions were derived, whether using different bases, by different algorithms or even by people versus computers. We show that latent class analysis (and K-means clus-

Executive Summary

Alternative segmentations must be compared on an "apples-to-apples" basis, by the amount of information they convey about a set of customer attributes. This article presents a simple and effective new metric to compare different segmentations on a common basis—the information they convey about a particular set of relevant attributes. This turns out to be the same metric that latent class analysis maximizes, yielding new insights into how segmentation algorithms work and how they can be better harnessed in practice.

tering) can be viewed as maximizing this metric. This yields a simple new understanding of how these two algorithms work and can support more effective application of these two algorithms in practice.

Any discrete attribute is a segmentation. Any customer attribute measured as a discrete (or discretized) variable can be a segmentation. Customers can be divided by age, gender, region, product interest, geography, attitudes, blood pressure or any other dimension. Segmentations "exist" as soon as we define them. But not all segmentations are equally good for a given purpose.

One key decision in segmentation research is choosing an algorithm to define segments. Figure 1(a) on page 11 shows how K-means clustering analysis segmented an actual market data set. Interestingly, K-means clustering will find such partitions even if there is no clear natural separation, and the boundaries are linear (in much the same way that boundaries between soap bubbles are planes). Figure 1(b) shows another reasonable way this data set might reasonably be linearly partitioned—as a 2x2 matrix divided at the medians.

How can we compare two segmentations? And, importantly for quantitative segmentation analysis, how can we measure the difference in meaningful terms?

Good segmentations convey information. In our view, segmentation is information compression. A segmentation is not useful unless it conveys information about important customer attributes. Ideally the converse is also true; observable customer attributes convey information about segment membership.

Fortunately, information is measurable. Information theory was pioneered by Claude Shannon in the 1940s to define the requirements for telecommunication bandwidth. Today, information theory is the basis for technology including modern cryptography, cell phones and ZIP files.

Shannon's work provides a framework to measure the information a segmentation conveys about relevant customer attributes. As a result, the "best" segmentation is one that conveys the most information about those attributes. This is the main principle behind LCA—to generate segments that maximize as much information on the basis variables.

To be clear, information is only one reasonable metric of segmentation quality. Other aspects also matter in practice, including established (if less readily quantifiable) considerations of identifiability (the ability to find these segments in a larger population), substantiality (the relative size of the segments), accessibility (the ease with which segments can be reached), responsiveness (the extent to which segments respond to marketing interventions or strategies), stability (the repeatability of the segmentation solution) and actionability (the ability to execute marketing strategies to the segments).

A Brief Primer on Information Theory

Computer science and statistics have derived similar concepts independently using different terms and assumptions. In particular, computer science has made extensive use of a metric called Shannon information (a different concept than Fisher information in statistics). Here we provide a brief primer on information theory, describing three key concepts: surprise, entropy and information. Here, surprise and entropy are mainly stepping stones on the way to understanding information.

Surprise! Surprise, $S(x)$, is a measure of improbability of an outcome $X=x$, defined as the negative logarithm of probability, $p = \Pr\{X = x\}$:

$$S(x) = -\log_2 p$$

The logarithm is usually base 2 as computers represent data in binary form. Therefore, the units are "bits" (as in computer bits). Surprise is also known as the Shannon information content of an outcome.

For example, getting heads on one fair coin flip would be exactly 1 bit surprising: $S(X) = -\log_2 (0.50) = 1.0$. Also, getting five heads on five consecutive coin flips would be 5.0 bits surprising, $S(X) = -\log_2 (2^{-5}) = 5.0$. On the other hand, getting heads on a double-headed coin would be 100 percent likely, and thus not at all surprising: $S(X) = -\log_2 (1) = 0$.

What should be evident from these examples is that the surprise of an event is inversely related to its likelihood. In other words, the lower the likelihood of an event, the greater its surprise if it occurs. Surprise has a lower bound of 0 for certainties and an upper bound of positive infinity for events that are "impossible" to occur.

Entropy. Entropy, $H(X)$, is a measure of uncertainty, analogous to variance for discrete variables. It is formally defined as the expected surprise over all possible values of X:

$$H(X) = -\sum p_i \log_2 p_i$$

Here i is an index over all the possible values that X may assume. Entropy is also defined for continuous variables, replacing the discrete summation with a continuous integral, in which case it is called the differential, or Boltzmann, entropy.

Entropy may seem at first to be an abstract concept, but it has a very practical interpretation: It is the number of computer bits that are theoretically required, on average, to store the observed values of a random variable in an optimally compressed data file (under basic assumptions).

As an example, we might measure product interest and partition customers into two segments: "high" if product in-

terest is above the population median and "low" if product interest is below the median. Any one customer has a 50 percent chance of being high (which is 1.0 bits surprising) and a 50 percent chance of being low (which is also 1.0 bits surprising). So the expected surprise, or entropy, is 1.0 bits:

$H(X) = 0.50 \log_2 (0.50) + 0.50 \log_2 (0.50) = 1.0$

Now, imagine a bet is placed on getting two heads in a row. Assuming the coin is fair, there is a 25 percent chance of winning that bet, which would be 2.0 bits surprising. Conversely, there is a 75 percent chance of losing the bet, which would be 0.41 bits surprising. Therefore, the expected surprise or entropy $H(X) = 0.81$ bits:

$H(X) = 0.25 \log_2 (0.25) + 0.75 \log_2 (0.75) = 0.81$

Entropy is maximized when the data are uniformly distributed. As probability mass is concentrated in fewer states, entropy decreases. Said another way, a variable's entropy becomes smaller as the distribution of responses gets concentrated into fewer levels. For example, if income is measured with six categories, but most respondents in a sample fall into only three of the categories, its entropy will be smaller than if the responses were more broadly distributed across all of the categories. And if all the probability is concentrated in a single state (i.e., the variable is certain), the entropy is zero.

For binary variables, the maximum entropy is 1.0, where the maximum possible entropy is the logarithm of the number of states. For example, a variable with four possible states has a maximum possible entropy of 2.0. This makes intuitive sense, as one could represent any four-state variable as two binary variables.

Mutual information. Mutual information, $I(X;S)$, between two variables $(X$ and $S)$ is the reduction in uncertainty (entropy) in X we would expect from knowing the value of S:

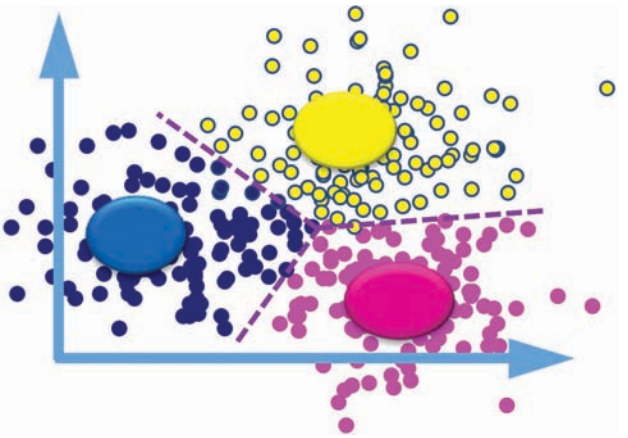
$I(X;S) = H(X) - H(X \mid S)$

Conceptually, segmentations are useful if they convey information about attributes of interest. Here, information is a measure of correlation for discrete variables. Let's continue with our previous example of product interest where, in the overall sample, 50 percent of customers are above the median and 50 percent are below. Now let's imagine that product interest varies by segment, as shown in Figure 2, on page 13. Within the "Quinn" segment, 75 percent of customers have high interest, 60 percent in the "Welby" segment and 15 percent in the "Becker" segment. Overall, the entropy of product interest is 1.0, as shown previously. Note that within each segment, however, entropy is lower. Assuming that each segment is equally likely, the expected entropy within each segment is $H(X|S) = 0.80$. Therefore, knowing the value of segment reduces our expected uncertainty by 0.20 bits, or said in another way, segment conveys 0.20 bits of information about product interest.

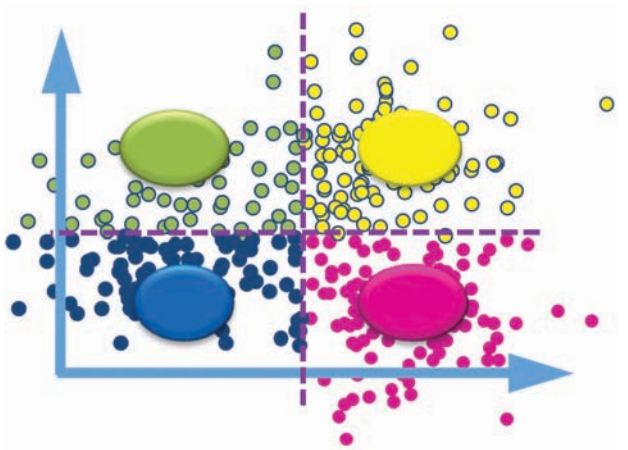
Mutual information is symmetric, that is $I(S;X) = I(X;S)$. Conversely, knowing product interest would convey 0.20 bits of information about segment membership.

Figure 1: Two segmentations of the same data

(a) K-means segmentation



(b) Simple 2x2 matrix



Total Mutual Information

We now introduce a new metric of segmentation quality we call the total mutual information, $I(S,X)$, that a segmentation S conveys about a set of basis variables, $X = \{X_i\}$:

$$I(\mathbf{X};S) = \sum_i I(X_i;S)$$

This metric can easily be calculated for any segmentation (even one defined by managerial insight) and evaluated with respect to any set of basis variables (even if different from that used to derive or define the segmentation). It can be calculated for discrete and/or continuous basis variables.

For example, let's say the variable segment is associated with product interest (I=0.20), market TRx (I=0.22) and a certain attitude statement (I=0.11). Thus the total mutual information that segment conveys about these three variables is 0.53 bits.

$I(X;S) = 0.20 + 0.22 + 0.11 = 0.53$

If the segmentation S is itself defined in terms of the basis variables, as is often the case in "post-hoc" segmentation analyses and classification algorithms, a more appropriate

measure is the net mutual information, I'(X;S), which subtracts the entropy of the segmentation itself:

$\mathbf{I'(X;}S) = \sum I(X_i;S) - H(S)$

To get an intuitive sense of why this correction is useful, consider an extreme example of a segmentation defined as one basis variable itself, $S \equiv X_1$. Clearly, S conveys all the possible information about X_1, as $I(X;S) = H(X)$, but it conveys no net information, $I'(\mathbf{X},S) = I(X_1;S) - H(X_1) = 0$.

Subtracting the entropy of the segmentation itself, $H(S)$, effectively penalizes the total mutual information score for more complex segmentations. Generally, one can almost always convey more total information by allowing more segments. Net mutual information corrects for that effect, as adding more segments eventually increases entropy $H(S)$ as fast (or faster) than the total mutual information $I(X; S)$.

Latent class analysis maximizes net mutual information. It turns out we can show that net mutual information is actually the very metric that LCA implicitly uses to compare alternative segmentations. In fact, LCA actually yields the segmentation with the maximum net mutual information with respect to a given set of basis variables, for a given number of segments.

LCA maximizes the expected log-likelihood function of a formal statistical model that makes a number of statistical assumptions, including that the basis variables are conditionally independent given segment:

$$E[\log L(\partial,\grave{e} \mid \mathbf{x})] = \sum_{n=1}^{N}\sum_{s=1}^{S} p_{ns} \log f(\mathbf{x}_n|\grave{e}_s) + \sum_{n=1}^{N}\sum_{s=1}^{S} p_{ns} \log \pi_s$$

Here $\mathbf{p}\ \mathbf{p} = \{ p\ p_s\}$ are the population proportions of each segment, $s \epsilon S$, x_n are the observed attributes for customer n, $f(\mathbf{x}_n|\grave{e}_s)$ is the probability distribution function for x, qq_s are the parameters for the distribution for segment s and p_{ns} is the probability that customer n is in segment s.

The log-likelihood function can be rewritten in terms of the net mutual information, which parallels the prior expression:

$$E\left[\log L(\partial,\grave{e} \mid \mathbf{x})\right] = N\left[I(\mathbf{X},S) - H(\mathbf{X}) - H(S)\right]$$

In other words, LCA can be re-derived as maximizing net mutual information:

$$E\left[\log L(\partial,\grave{e} \mid \mathbf{x})\right] = N\left[I'(\mathbf{X},S) - H(\mathbf{X})\right]$$

Note that $H(\mathbf{X})$ is constant for given set of attributes $\mathbf{X}$ and a given set of observed data and thus does not affect the maximization.

Unlike traditional presentations of LCA, this derivation does not assume that attributes are "conditionally independent" given segment (i.e., that segment membership alone explains all the associations among the basis attributes). These results also apply to K-means clustering, which can be regarded as a special case of latent class analysis when all variables are assumed continuous (normally distributed with equal variance).

Relation to other metrics. The Bayes information criterion (BIC) is another metric that compares alternative segmentation solutions. The BIC penalizes the log-likelihood function based on the number of model parameters k and sample size N:

$$BIC = -2\log L(\partial,\grave{e}|\mathbf{x}) + k \log N$$

Given the relationship between net mutual information and the expected log-likelihood, one could similarly penalize net mutual information analogous to the BIC:

$$\mathbf{I''(X:}S) = \mathbf{I'(X:}S) - \frac{k \log N}{2N}$$

The penalty term favors the selection of solutions with fewer segments, particularly with smaller data sets. Theoretically, the penalty term estimates the difference between the estimated model and an assumed true but unobservable model. The BIC is typically used to compare hierarchically nested solutions (e.g., a three-segment vs. a four-segment solution) where the solutions are derived from the same data set and model parameterization. In such a case, choosing the model that minimizes the BIC will yield the "true" number of segments with probability $p \to 1$ as $N \to \infty$.

Indeed, much literature on segmentation methods focuses on how well algorithms can recover "true" segment membership in artificial examples. However, in practice, segmentations are applied to real-world populations where segmentation is merely a useful simplification. In our view, a more practical question is how much information alternative segmentations convey about key attributes.

IN PRACTICE, SEGMENTATIONS ARE APPLIED TO REAL-WORLD POPULATIONS WHERE SEGMENTATION IS A USEFUL SIMPLIFICATION.

One advantage of net mutual information $I'(\mathbf{X},S)$ is that it is not explicitly dependent on either the number of parameters estimated in the model used to derive the solution S, on the sample size or even predicated on the assumption that a segmentation exists *a priori*. Thus it may be used to compare alternative solutions without regard to how they were derived. For instance, even within LCA, one might consider different bases (e.g., assumptions of nominal vs. ordinal vs. continuous etc.), yielding a number of solutions. Net mutual information with respect to a common basis can be compared across solutions, whereas the BIC values originally calculated for each solution cannot.

Net mutual information also differs from total correlation, another generalization of mutual information, not least in that the metric proposed here does not attempt to consider other correlations among the attributes $\{X_i\}$ beyond their conditional dependence on S. This is similar to the assumption in the LCA model that each attribute is a conditionally

independent given segment, but different in that the metric posed here does not explicitly assume that the data actually fit the conditional independence model.

Example. Figure 3 shows the net mutual information for an actual recent project. For any given number of segments, latent class analysis yields the segmentation with the maximum total mutual information as well as the maximum net mutual information. However, this maximum can increase with the number of segments. Thus a five-segment solution is potentially more informative than a four-segment solution, with diminishing returns to increasing numbers of segments. We further used mutual information to identify the specific attributes that are highlighted in each segmentation. One segmentation was more focused on "hard" attributes such as prescription volumes, while another better captured "soft" attributes such as treatment attitudes.

Of course, one is not obligated to choose the latent class solution. Even in this case, the client ultimately chose a solution that had a slightly lower information score, but which was more practical in other aspects, not least simplicity.

Figure 2: Product interest varies by specialty

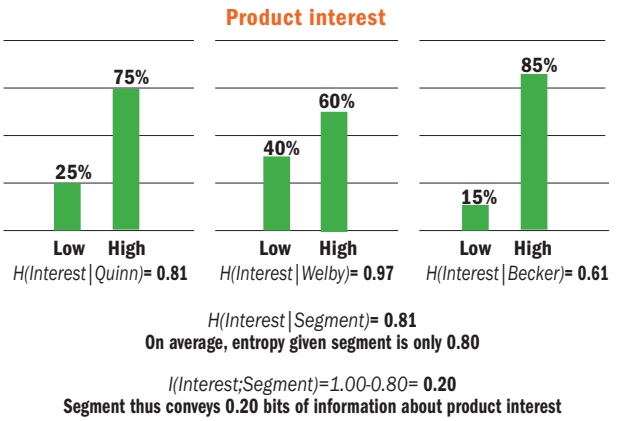
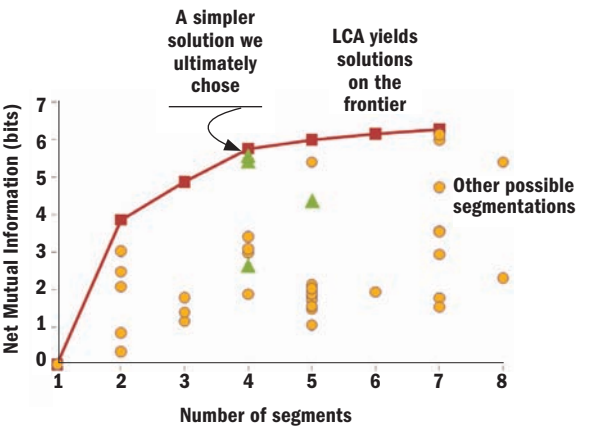


Figure 3: LCA yields total segment info



The Art of Segmentation

In practice, analysts often face the challenge of comparing alternative segmentations. Within one project, alternative solutions may be derived using a variety of approaches and iterative judgments (e.g., "What if we treat these variables as ordinal rather than nominal?" "What if we include current product shares as attributes?" "What if we reduce the attitudinal questions via factor analysis and use the factors in the segmentation analysis?") There may be prior segmentations currently in use, perhaps based on statistical analysis or simple breaks by market volume. Experienced brand managers may postulate segmentations based on their experience. Relevant questions would be "What attributes do these segmentations convey information about?" and "How much information do these segmentations convey about customer attributes relevant to our current goals?" Approaches such as ensemble models address this challenge by aggregating the alternative solutions. Net mutual information provides a means to compare them.

We present a simple, effective, new metric to compare different segmentations on a common basis—the information they convey about a particular set of relevant attributes. This can be compared for any segmentation, including ones defined via computer algorithms such as latent class, existing segmentations already in use and even segmentations posed qualitatively based on managerial insight. It is not dependent on the algorithms or samples used to derive the segmentations. One can also use mutual information to compare the attributes that are highlighted in each segmentation.

Further, we can define a "best" segmentation as one that mathematically optimizes this simple metric to find the most informative segmentation. This is exactly what algorithms such as latent class analysis (and its special case, K-means clustering) do. So we can compare any two segmentations on the same metric implicitly used by latent class analysis, a leading segmentation algorithm. And we now have a simple way to understand these algorithms without needing statistical assumptions such as conditional independence or model fit.

Of course, there are many other practical aspects to a good segmentation beyond information (and more specifically, this narrow definition of information). The art of segmentation includes identifying those customer attributes that are important, establishing the relative importance of simplicity and identifiability versus nuance, measuring them accurately and communicating how these can inform decision-making. Understanding information (and how segmentation algorithms work) helps focus the science in support of the art. ●

Pieter Sheth-Voss, PhD, is director of product innovation and **Ismael Carreras,** PhD, is a senior research director in market intelligence at Quintiles. They may be reached at Pieter.ShethVoss@quintiles.com and Ismael.Carreras@quintiles.com.