LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Sara Dolnicar & Friedrich Leisch

Evaluation of Structure and Reproducibility of Cluster Solutions Using the Bootstrap

Technical Report Number 063, 2009
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Evaluation of Structure and Reproducibility of Cluster Solutions Using the Bootstrap

Sara Dolnicar

Marketing Research Innovation Centre (MRIC), School of Management and Marketing, University of Wollongong
Northfields Ave, NSW 2522
Australia

Friedrich Leisch

Department of Statistics, Ludwig-Maximilians-Universität München
Ludwigstrasse 33, 80539 Munich
Germany

This is a preprint of an article that has been accepted for publication in **Marketing Letters**. *The original publication is available at www.springerlink.com. Please use the journal version for citation, see* http://dx.doi.org/10.1007/s11002-009-9083-4 *for details.*

Abstract

Segmentation results derived using cluster analysis depend on (1) the structure of the data and (2) algorithm parameters. Typically neither the data structure is assessed in advance of clustering nor is the sensitivity of the analysis to changes in algorithm parameters. We propose a benchmarking framework based on bootstrapping techniques that accounts for sample and algorithm randomness. This provides much needed guidance both to data analysts and users of clustering solutions regarding the choice of the final clusters from computations which are exploratory in nature.

Keywords: cluster analysis, mixture models, bootstrap

1 Introduction

Market segmentation aims at "dividing a market into smaller groups of buyers with distinct needs, characteristics or behaviors who might require separate products or marketing mixes" (Kotler and Armstrong, 2006). A good market segmentation strategy can lead to substantial competitive advantages for an organization. But the quality of the market segmentation strategy depends on the quality of the segmentation solution informing it.

A wide range of general quality criteria, such as measurability, accessibility, substantiality, differentiability and actionability, have been proposed to assess segmentation solutions (Kotler, 1997; Kotler and Armstrong, 2006; Wedel and Kamakura, 1998; Evans and Berman, 1997; Morritt, 2007). Yet, very little practical guidance is available to help users to choose the best segmentation solution. This lack of guidance is particularly concerning because (1) the sample and the algorithm used introduce a significant amount of randomness into the final segmentation solution, (2) no clear distinction is currently made between solutions that reveal naturally existing clusters (density clusters) and solutions that construct clusters, and (3) managers have significant knowledge gaps in relation to segmentation methodology and cannot be expected to make an informed decision about the final market segmentation solution that should be used as a basis for strategic decisions (Dolnicar and Lazarevski, 2009).

1.1 Sample and algorithm randomness

Two sources of randomness affect the final segmentation solution: the sample - which is a random subset of the population - and the algorithm - which is known to impose structure on data

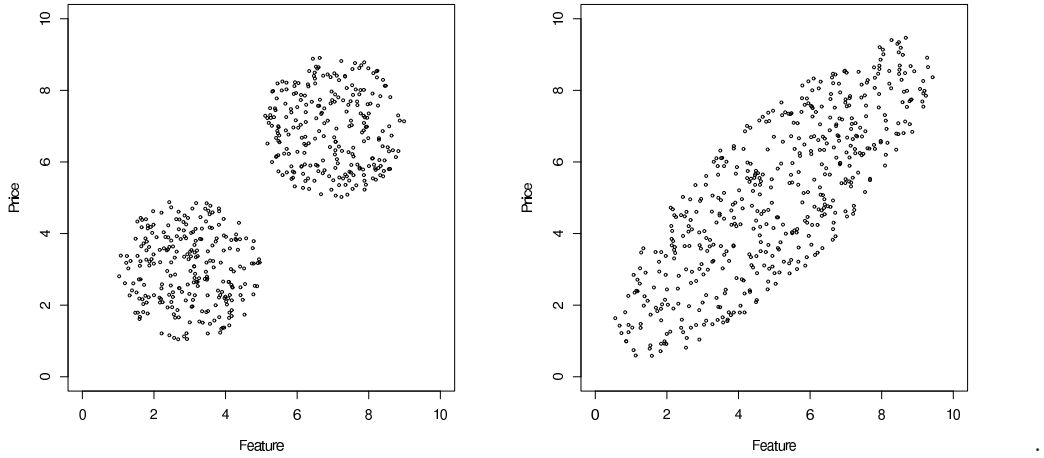


Figure 1: Artificial data for the mobile phone example: two clearly separated circles (left) and one big elliptical cluster (right). Both will be used as running examples to demonstrate theory and methods for evaluating cluster stability.

(Aldenderfer and Blashfield, 1984), especially if the data are not well structured. Using one single segmentation solution that results from one single computation of one single algorithm therefore puts the manager in danger of using a random solution, rather than a reliable solution. Let us assume we cluster a sample with 500 respondents. If we survey another 500 people and cluster the data, we can either get a very similar or a very different solution, depending on the variability of answers in the (unknown) total population. This is sample randomness. Another source of randomness is that many cluster algorithms are stochastic in nature, e.g., by using random starting points. This algorithm randomness can be reduced by repeatedly starting the algorithm and keeping only the best solution, but it usually remains unknown whether the global optimum has actually been found.

1.2 Natural, reproducible or constructive clustering

Optimally, density clusters exist in consumer data. If this is the case, the clustering algorithm has to reveal these density clusters correctly and reliably over repeated computations. This is the way market segmentation has been originally conceptualized by the pioneers of market segmentation (Frank et al., 1972; Myers and Tauber, 1977). We refer to this situation as (*natural clustering / segmentation*

Density clusters rarely exist in consumer data. In the worst case, data are entirely unstructured. Although it may technically seem "foolish to impose a clustering structure on data known to be random" (Dubes and Jain, 1979, p.42) it is now acknowledged that market segmentation is still of managerial value in such situations. While the cluster algorithm cannot reveal true groups of consumers, it can still help to create managerially useful subgroups of customers (Mazanec et al., 1997; Wedel and Kamakura, 1998). Artificially creating segments and customizing products to parts of the market is often managerially preferable to accepting that no natural segments exist and pursuing a mass marketing strategy. We refer to the situation where market segments are created on the basis of an unstructured data set as *constructive clustering / segmentation*.

Typically, consumer data contain some structure, but not density clusters. For an example see the right hand plot in Figure 1. In such situations true density clusters cannot be revealed but data structure can be used to derive stable, reproducible market segments. We refer to this situation as (*reproducible clustering / segmentation*.

Imagine a two dimensional data set resulting from customers of mobile phones being asked about their aspiration level concerning the feature sophistication of the phone and the price (see Figure 1). The data set collected could take one of three forms: (1) true natural clusters exist

in the data set (left chart), (2) structure exists in the data, but the structure does not represent true, natural clusters (right chart), or (3) there is no structure in the data at all. An example for the latter is shown in Figure 6 and will be explained in detail below. Natural clusters or segments are strongly homogeneous internally and distinctly different from other segments. In our example they could consist of consumers who prefer the maximum feature sophistication and are willing to pay a high price for such a hi-tech mobile phone (segment 1). If such a segment existed, it would obviously represent a highly attractive target market. Not only because customers in this group are willing to pay a lot of money. A naturally existing segment of consumers not requiring many technical-features but wanting the price to be low (segment 2) would also represent an interesting target segment as long as the members of this group are homogeneous in their preferences. If the natural segments are clearly separated as in the left panel of Figure 1, probably any appropriate clustering technique will correctly identify the true number and nature of segments.

This is not, however, the case if the data does not form two distinct groups but rather a continuum between the low-end and high-end segments as shown in the right panel of Figure 1. The "statistically correct" solution in this case is to identify one big segment. Such a solution would, however, not be of much managerial use; it has to be split into several segments. In this case the clustering method chosen will have major structure-imposing effects and, consequently, will have to be selected very carefully in order to arrive to the managerially most useful solution. Yet, such a data situation enables clustering algorithms to repeatedly identify similar segmentation solutions (reproducible segmentation), whereas a total lack of structure in the data requires constructive clustering. In the case of constructive clustering the algorithm chosen and the sample of the population has the highest impact on the final solution. Or, as Aldenderfer and Blashfield (1984, p.16) put it: *"Although the strategy of clustering may be structure-seeking, its operation is one that is structure-imposing. ... The key to using cluster analysis is knowing when these groups are 'real' and not merely imposed on the data by the method."*

1.3 Managers' need for methodological guidance

Managers are struggling to fully understand market segmentation solutions. A survey we conducted among 198 marketing managers in 2007 indicates that more than two thirds believe that clustering of survey data only leads to segments if clear, distinct segments exist in the data, that "market segmentation reveals naturally occurring segments". Futhermore managars expressed that the way segmentation solutions presented to them are derived is like a black box to them (data goes in and the solution comes out, it is not clear what happens in between). Sixty-five percent admit to having difficulties interpreting segmentation solutions. Similar conclusions emerged from earlier conceptual and qualitative studies (e.g., Greenberg and McDonald, 1989; Dibb and Simkin, 1997). These findings confirm that there is not only a need for more guidance regarding the choice of segmentation solutions among academic researchers, but that users of segmentation solutions in industry suffer from a knowledge deficit, which makes it even more important to provide them with guidance to avoid the use of random and sup-optimal market segmentation solutions as the basis of their strategic decisions.

In this paper we propose a benchmarking framework that will inform data analysts and users of market segmentation solutions about (1) the data structure they are facing and consequently whether natural or reproducible clustering is possible or constructive clustering is required, and (2) how critical methodological decisions (such as the choice of the algorithm, the number of clusters etc.) are given this data situation. This will be achieved through the use of bootstrapping, a method that accounts for the two sources of randomness in clustering solutions: randomness of samples and randomness of algorithms. The remainder of this paper is organized as follows: Section 2 introduces a general framework for bootstrapping cluster algorithms, Section 3 demonstrates the framework using the simple artificial example from above, and Section 4 gives two case studies on real world data. All examples are taken from the context of market segmentation, however the underlying principles can be applied in any field of application where clustering or related techniques are used.

2 Bootstrapping Segmentation Algorithms

In the following we extend the benchmarking framework by Hothorn et al. (2005) for regression and classification problems ("supervised learning") to the case of cluster analysis ("unsupervised learning"). Let $\mathcal{X}_N = \{x_1, \ldots x_N\}$ denote a data set of size N. A partition of this data set into K segments assigns each x_n a vector $C(x_n) = (p_{n1}, \ldots, p_{nk})'$, where $p_{nk} \geq 0$, $\sum_{k=1}^{K} p_{nk} \leq 1$, and p_{nk} measures the "membership" of x_n in segment k. For partitioning algorithms like k-means, where each observation is assigned to one class, exactly one of the $p_{nk} = 1 \forall n$. For finite mixture models the memberships are the posterior probabilities of the respective components; for fuzzy clustering these are fuzzy logic memberships. In the following we will refer to all of those as *segmentation algorithms*.

As segmentation algorithms are unsupervised learning methods, membership values are identified only up to permutations of the labels. If Π denotes any permutation of the numbers $\{1, 2, \ldots, K\}$, then $\Pi C(\cdot)$ is equivalent to $C(\cdot)$, only the segment labels have changed. Most segmentation algorithms use random starting values and/or stochastic optimization techniques resulting in a local minimum of their respective optimization problem. Running the same segmentation algorithm twice on the same data set we obtain two membership functions $C_1(\cdot)$ and $C_2(\cdot)$. These may be almost equivalent up to label permutations Π such that $P\{C_1(x) = \Pi C_2(x)\} = 1 - \epsilon$ with ϵ small and the probability measure P is with respect to new observations x. However, especially in higher-dimensional spaces or if the data contains no natural cluster structure, C_1 and C_2 may belong to different local minima.

In order to avoid "obvious" local minima it is recommended to run the segmentation algorithm several times with different starting conditions and use the best solution. Below we assume this procedure as part of model fitting. Nevertheless several local minima can remain, as demonstrated with artificial data in Section 3. E.g., "fitting a mixture model with the EM algorithm" assumes that the EM algorithm is run several times and the solution with the maximum likelihood is returned.

An additional source of randomness, which is ignored in many applications, is the actual sample $\mathcal{X}_N$ of observations. Any partition $C(\cdot) = C(\cdot|\mathcal{X}_N)$ is a random variable (in function space) depending on the learning sample $\mathcal{X}_N$, and stochastic components of the segmentation algorithm. If we use one realization, i.e., one particular segmentation of the data, it is interesting to know how much variation this distribution has. If replications of the algorithm on different samples $\mathcal{X}_N$ from the same data generating process (DGP) return similar results, we call the corresponding partitions *reproducible*, otherwise *non-reproducible*. This is closely connected to the main idea behind cluster ensembles, where many partitions are combined into average (or meta-)partitions (Strehl and Gosh, 2002; Dudoit and Fridlyand, 2003; Dolnicar and Leisch, 2000).

2.1 Evaluating Reproducibility

In order to evaluate the reproducibility of a segmentation algorithm we need to integrate out all sources of randomness in the partition, hence we need independent

- replications of the sample $\mathcal{X}_N$, and
- replications of the algorithm.

It is easy to get replications of the algorithm, but usually we are given only one sample $\mathcal{X}_N$ of fixed size N. If N is very large, we can split it randomly into several smaller sets and use those, cf. function `clara` in Kaufman and Rousseeuw (1990). If N is not very large such that we cannot afford to split it into several independent samples, we have to rely on approximations. The simplest and most widely used approximation of the unknown distribution F of the DGP is the empirical distribution $\hat{F}_N$ of $\mathcal{X}_N$. Drawing samples from $\hat{F}_N$ is the same as sampling with replacement from $\mathcal{X}_N$, i.e., bootstrapping (Efron and Tibshirani, 1993). Note that it is the usual convention to draw bootstrap samples which have the same size N as the original data set.

Running the segmentation algorithm on B bootstrap samples $\mathcal{X}_N^b$ $(b = 1, \ldots, B)$ gives us B replications $C_1, \ldots, C_B$ of $C(\cdot)$ which are *independent* random variables in function space, where independence is with respect to training sample and segmentation algorithm. To assess how often we obtain the "same" solution we have to define a similarity (or distance) measure $s(C_1(\cdot), C_2(\cdot))$ on partitions. Possible measures include

- Kullback-Leibler divergence between the mixture densities for model-based clustering,
- Euclidean distance between centers for centroid-based cluster algorithms like k-means, and
- agreement measures for partitions like the Rand index or adjusted Rand index (Hubert and Arabie, 1985).

In the following we will use the Rand index adjusted for agreement by chance as it makes no further assumptions about the segmentation algorithm, see, e.g., Brusco et al. (2003) for an example of evaluating market segmentation algorithms. However, all analyzes could as well be done using any other measure for similarity or distance of partitions.

Given two clusterings C_1 and C_2 of two different bootstrap samples, we first predict cluster membership in these two partitions for all observations in the original data set by assigning each to the closest centroid in each partition, respectively. For model-based clustering, observations are assigned to the cluster with the highest a-posteriori probability. Given these two partitions of the original data set, we then compute the Rand index, which is defined as the percentage of pairs of points which are either assigned to the same cluster twice, or assigned to different clusters twice (both cases suggest agreement of partitions).

Using $2B$ bootstrap partitions of the original data set we get B independent and identically distributed (iid) replications

$$s_1 = s(C_1, C_2), \quad \ldots \quad s_B = s(C_{2B-1}, C_{2B}) \tag{1}$$

of partition similarity. Note that we could compute all $2B(2B-1)/2$ pairwise similarity coefficients, however those would no longer be independent. As we will use standard statistical techniques on the sample $\mathcal{S} = \{s_1, \ldots, s_B\}$, independence is of great importance, see Hothorn et al. (2005). Many questions about stability of the segmentation algorithm can now be formulated in terms of standard statistical inference on $\mathcal{S}$ as demonstrated below on several examples.

2.2 The Benchmarking Framework

Does the data set contain natural clusters? This question is similar to the typical "number of clusters" question (Thorndike, 1953). A huge number of indices has been proposed in the literature for assessing the number of clusters (see, e.g., Milligan and Cooper, 1985; Dimitriadou et al., 2002). For model-based clustering information criteria like the AIC or BIC can be used (e.g., Fraley and Raftery, 1998). The modification to common practice we propose is to compute the index of choice (or several of them) on all bootstrap replicates, not only on the original data (or one single split into two halves) to assess stability of the index estimate, see also Tibshirani and Walther (2005).

Does the data set allow reproduction of similar segmentation solutions? This question has received considerably less attention in the literature. Given the framework described above, we have several iid samples $\mathcal{S}^K$ of algorithm stability. In many cases exploratory analysis of these data sets, e.g., using kernel density estimates as shown in the examples below, gives sufficient insight to answer the question. If the (adjusted) Rand index is used, then "reproducibility" is equivalent to the fact that the distribution of the s_b^k should have most mass close to 1. If exploratory analysis does not give sufficient insight, then standard significance tests can be used (e.g., that the mean or median is above a certain threshold).

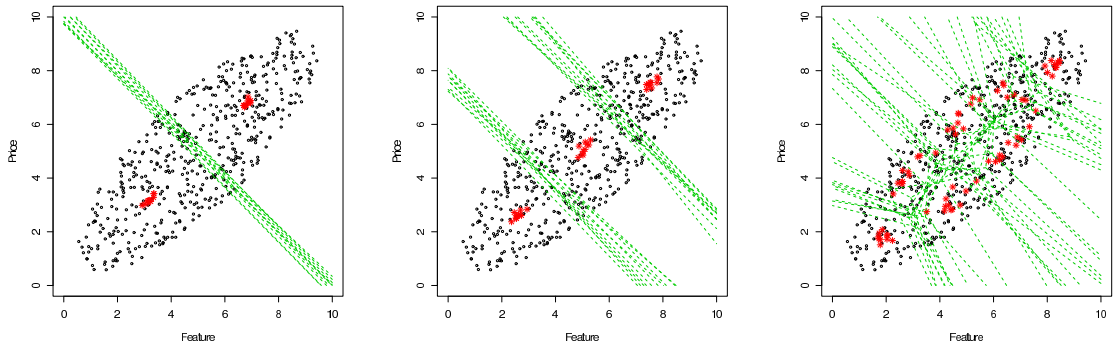


Figure 2: Both two cluster (left) and three cluster (middle) partitions are stable for the ellipse data; while seven cluster partitions (right) are mostly random, only the extreme low-end and high-end segments are stable.

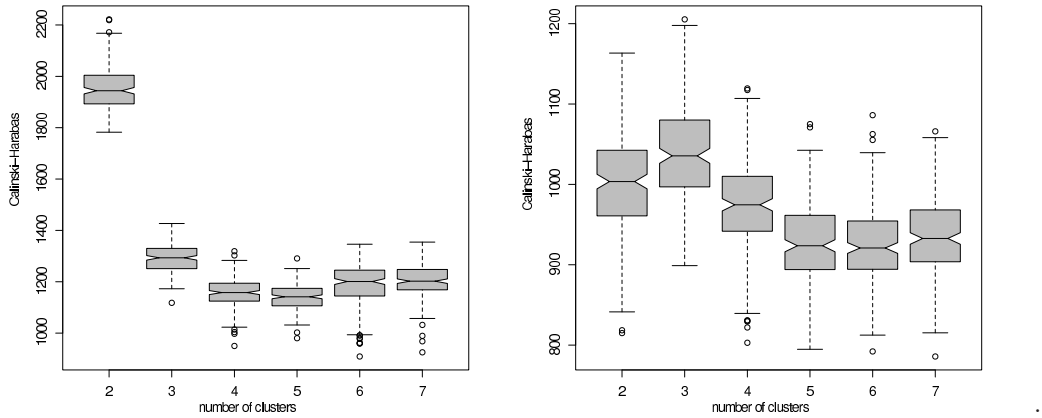


Figure 3: Boxplots of the Calinski-Harabas index for the two circles and the ellipse data for 200 partitions each.

3 Demonstration Using Artificial Data

3.1 Natural structure versus reproducible structure

Figure 2 shows the partitions induced by running the k-means algorithm repeatedly on new data sets from the data generating process of the simple example used in Section 1. The exact choice of the segmentation algorithm is not critical in the case of this simple data set. Experiments using model-based clustering gave almost the same results and are omitted for brevity. Note that in our simulations "running k-means" is equivalent to running the algorithm five times and using only the best solution (smallest within-cluster sum of squares).

It can clearly be seen that the data set – although consisting only of one big cluster – can reliably be split into two or three clusters, whereas a seven cluster solution will split the data rather randomly (with respect to reproducibility). Note that for the two-dimensional data the maximum number of stable groups can be inferred from plots of the data, for real data with more than two dimensions this is not an option. Consequently, indices need to be used to indicate the most appropriate number of segments in a data. We compute the Calinski-Harabas index (variance between cluster centers divided by sum of within cluster variances). Other indices can be used. We chose Calinski-Harabas because it is widely used and the best performing index in the classic paper by Milligan and Cooper (1985).

Figure 3 shows the Calinski-Harabas index values for two to seven clusters, indicating that two segments represent the best solution for the two circle data and that three segment should be

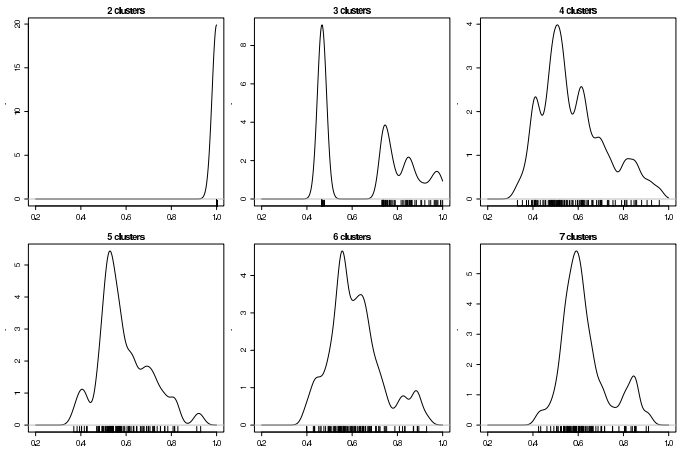


Figure 4: Kernel density estimates of 100 bootstrapped adjusted Rand indices for the two circles data with two to seven clusters each.

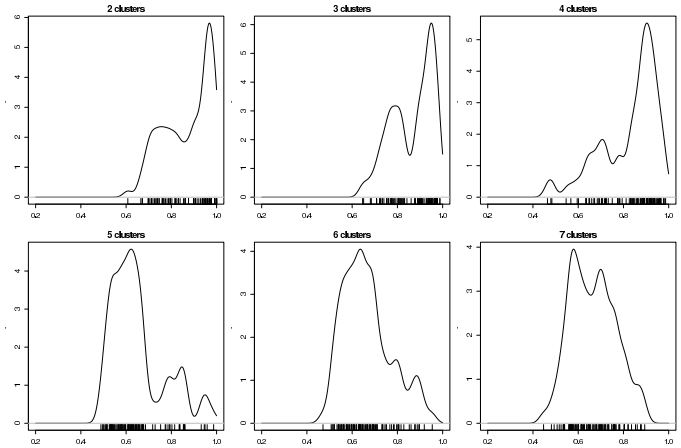
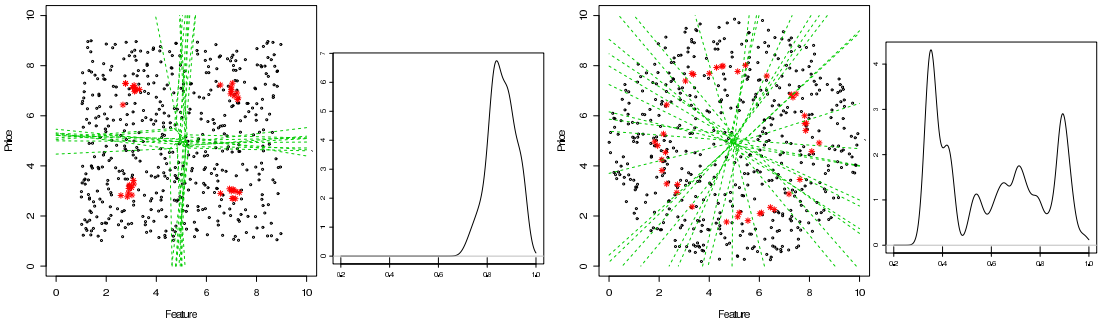


Figure 5: Kernel density estimates of 100 bootstrapped adjusted Rand indices for the ellipse data with two to seven clusters each.

chosen for the ellipse data. Note that, not surprisingly, the recommendation for the three segment solution for the ellipse data is weaker than the two segment recommendation for the two circle data. This reflects the difference between a natural and a reproducible segmentation solution.

Even more informative diagnostics of segment stability can be obtained by looking at the bootstrapped Rand indices s_b as shown in Figures 4 and 5. The kernel density estimates can be seen as smoothed versions of histograms, e.g., see Silverman (1986) for details. Like a histogram, the density estimate is high in regions with many observations, and low else. The two circle data is always correctly split into two groups, hence the strong peak at one for the two cluster solution. When looking for three clusters the peak moves to 0.5 because one segment is always split randomly into two segments. There is still a lot of density mass above 0.75 which corresponds to s_b where the same circle got split into two segments. In summary: the data cannot be split into more than two segments in a stable reproducible manner.

Although the ellipse data consists of only one natural segment, the data can be split into up to four segments in a stable manner. We refer to this as pseudo cluster structure: although no natural clusters exist in the data, segmenting the data several times repeatedly leads to the same results. Data containing pseudo cluster structure enables stable segmentation (consumer segments can be reproduced with different algorithms over multiple repeated computations), which gives



(a) circle data

(b) square data

Figure 6: A four cluster partition is stable on a square, but completely unstable on a circle.

management more confidence in their target segments than if constructed and not reproducible segments are selected. Managerially it makes a lot of sense to treat consumers located in the top right hand corner (*price-inelastic techno buffs*) differently than consumers in the bottom left hand corner (*basic mobile bargain hunters*). Management may even consider a third segment in the middle as a potentially useful target market.

3.2 Reproducible structure versus constructive structure

As another example consider the square and circle data shown in Figure 6. The dotted lines indicate the partitions derived from simple k-means clustering. It is obvious, that the square data set leads to the detection of a reproducible structure for four segments, while the solutions for the circle are completely random, depending on the starting points of the algorithm rather than data structure. This can also be seen from the density plots of the bootstrapped Rand indices s_b for four segments. High Rand-Index values are repeatedly achieved for the square data, whereas the compliance of repeated runs of the partitioning algorithm varies between 0.3 and 1, depending on the partitions compared, thus not indicating any structure in the data that would enable either natural or reproducible segmentation.

To continue the marketing management interpretation of the artificial example, we would suggest the following consequences for the two data sets: the square data would be classified as stable segmentation case with four market segments. In addition to the two segments mentioned above, this data set would include also *highly price-elastic techno buffs* and – every manager's dream segment – the *modest price-inelastic mobile phone users* in the top right-hand corner who seemingly do not mind to pay a lot of money for a basic mobile phone with few features only. In the case of the circular data set, management has to be informed that natural segments do not exist and that the data set does not enable the researcher to determine stable segments of consumers. In this constructive segmentation case a number of solutions have to be computed, visualized, described and presented to management for comparative evaluation.

Bootstrapping the partition similarity (or any other measure of interest) not only allows for exploratory analysis of partition stability, but gives iid samples that can be analyzed using inferential statistics. Consider we want to select the most stable number of clusters for the ellipse data. Sloppily speaking we want to select the sample in Figure 5 which has most density mass towards the right side (Rand index very often close to 1, high agreement between partitions). If we assume that stability decreases with an increasing number of clusters (as the density plots strongly suggest for this example), we want to test whether a shift to the left is significant between neighboring numbers of clusters.

The simplest method would be to do pairwise t-tests comparing the solutions with two and three clusters, three and four clusters, four and five clusters, etc. and look for significant changes.

# clusters	Estimate	Std. Error	t-value	p-value
3 - 2	-0.0040	0.0154	-0.264	0.999
4 - 3	-0.0386	0.0154	-2.501	0.057
5 - 4	-0.1775	0.0154	-11.504	<0.001
6 - 5	0.0152	0.0154	0.989	0.808
7 - 6	0.0004	0.0154	0.026	1.000

Table 1: Multiple comparison test for successive differences of bootstrapped Rand index values for the ellipse data. The significant jump in stability is between four and five clusters.

However, the tests are not independent from each other, such that correction for multiple testing is not straightforward. A better approach is to do an ANOVA-type analysis controlling the family-wise error rates of the tests. Table 1 shows sequential t-tests corrected for multiple comparisons (e.g., Searle, 1971), also known as Tukey's "honest significant difference" method. The two and three segment solutions do not differ significantly with respect to stability, while the four segment solution is slightly less stable. A large gap in stability is detected between the four and five segment solution (the Rand index of the five segment solution is on average 0.178 smaller, $p < 0.001$), but the five, six and seven segment solutions do not differ significantly in stability. Similarly, we can compare different cluster algorithms on the same data set, see Section 4.1.

4 Case Studies Using Empirical Data

4.1 Guest Survey Data

We use guest survey data collected by the Austria National Tourism Organization in the summer seasons of 1994 and 1997 as our first empirical example. We choose this data because the sample size is large (14,571 respondents), which is a pre-requisite for market segmentation based on a high-dimensional segmentation base. We will investigate which segmentation concept is appropriate if destination management is interested in targeting behavioral segments. The behavioral information available consists of 22 vacation activities. Each respondent was asked to state whether he or she undertook each one of the activities during the stay or not. We included 12,273 respondents with complete answer profiles. The average participation rate in those activities ranged between 3% percent (horse riding) and 93% (going for walks). In addition to this segmentation base, a number of descriptive pieces of information of demographic, socioeconomic, psychographic, attitudinal or behavioral nature were available for each respondent, see, e.g., Dolnicar and Leisch (2000) for a more detailed description and analysis of the data.

We compare the k-means and neural gas (Martinetz and Schulten, 1994) clustering algorithms, and a binomial mixture model estimated using the EM algorithm, on the guest survey data. See, e.g., Brusco (2004), Wedel and Kamakura (1998) and references therein for market segmentation using binary data. The boxplots of the Calinski-Harabas index – the traditional way of assessing the optimal number of clusters in a data set – indicate no natural segment structure (not shown). The kernel density plot, however, clearly identifies that only the two segment solution emerges as really stable (Figure 7).

Unfortunately, the two segment solution is not managerially useful as it merely splits respondents in people who tend to agree and those who tedn to disagree with survey questions. Managers require more differentiated profiles of segments than that. We therefore are looking for a more differentiated grouping which still provides a relatively high level of stability or reproducibility.

For three and four segments two alternative segmentation solutions are recommended. If the data is split into five or more segments, all resulting solutions are created through data analysis and can be used in the sense of constructive segmentation only.

The neural gas algorithm generally results in more stable solutions. Pseudo cluster structure is identified for up to five segments, enabling stable segmentation. The binomial mixture model is not necessarily identified as we have no repeated measurements per respondent (e.g.,

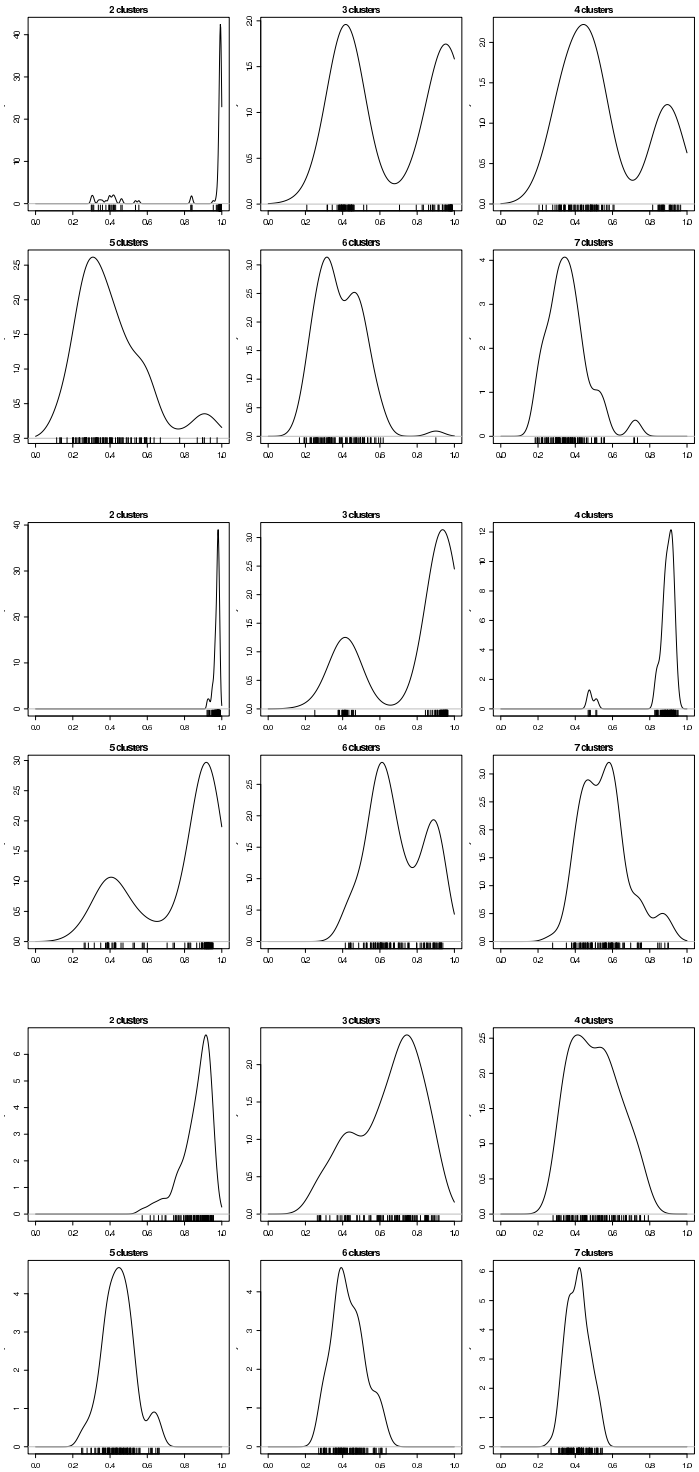


Figure 7: Kernel density estimates of 100 bootstrapped adjusted Rand indices for the GSA data clustered by k-means (top panel), the neural gas algorithm (middle panel) and a binomial mixture model (bottom panel).

Algorithms	Estimate	Std. Error	t-value	p-value
kmeans - binmix	-0.0431	0.0254	-1.694	0.209
neuralgas - binmix	0.3070	0.0254	12.061	<0.0001
neuralgas - kmeans	0.3501	0.0254	13.755	<0.0001

Table 2: Multiple comparison test for pairwise comparisons of bootstrapped Rand index values for the GSA data. The significant jump in stability is between four and five clusters.

Titterington et al., 1985). Stability results are similar to the stability results for the k-means algorithm on this particular data set. Table 2 shows pairwise comparisons of the stability (measured again by the corrected Rand index) of the three algorithms for a five segment solution. There is no significant difference between k-means and a binary mixture model (p-value 0.209), but neural gas clearly outperforms the other two (both p-values < 0.0001).

From this analysis we have to conclude that no natural behavioral market segments exist among tourists visiting Austria. Based on the results depicted in Table 2 we can, however, take a reproducible segmentation approach in segmenting this data set, in which case the five segment solution computed with the neural gas algorithm should be used.

4.2 Cake Conjoint Data

The second real world example is a conjoint data set used as test file in the Glimmix software package (Wedel and Boer, 2002). The data are from a study conducted for a bakery using the following four explanatory variables: price (Fl 1.5, 2.0, 2.5, 3.0), taste (neutral, raisins, ananas, mixed), number of pieces per package (4, 6), and label (bakery, luxury). A fractional factorial design with 16 combinations of the explanatory variables was used, and each combination was rated by 68 customers on a 9-point scale (1=low, 9=high). The data analyst has the choice of using a metric or an ordinal version of the price variable.

The Glimmix user manual (Wedel and Boer, 2002, p. 94) recommends the following latent class mixture regression model for the data: *"Thus one may specify a normal distribution, and an identity link. Running the analysis from one to eight segments may reveal that around seven segments provides the best description of the data as indicated by the information criteria, but one may wish to use a lower number of segments."* We will use the proposed benchmarking framework to answer the following two questions: 1. How many segments should the selected solution contain? 2. Should price be entered into the model as a metric or ordinal variable?

We first fit mixture models with two to nine components by running the EM algorithm ten times for each number of components and keeping the best solution with respect to model likelihood. If price is entered as a categorical variable, three coefficients are estimated for price alone in each mixture component, while only one coefficient is needed for a metric variable. The AIC recommends eight or nine segment solutions, while the BIC recommends six segments when the metric price variable is used and four if the ordinal price variable is used. The BIC penalizes stronger than the AIC for additional parameters, hence the preference to more parsimonious models. The two different recommendation of four and six clusters, however, poses the serious managerial question on how many segments to use. More diagnostic information about the structure and stability of alternative segmentation solutions is required to make a final decision on which segmentation solution to retain.

To get a better picture of the stability of the model selection criteria with respect to changes in the data set, we bootstrap the whole procedure $B = 100$ times, and compare AIC, BIC and corrected Rand indices as above. The recommendations resulting from the AIC do not change, and results are therefore not shown here. Boxplots of the bootstrap replica of the BIC are shown in Figure 8. Taking the variability of the BIC with respect to changes in the data set into account, the minimum median BIC is obtained for six segments, independent from the question whether the price is used as a metric or ordinal variable. Using the bootstrap makes a difference when evaluating the BIC.

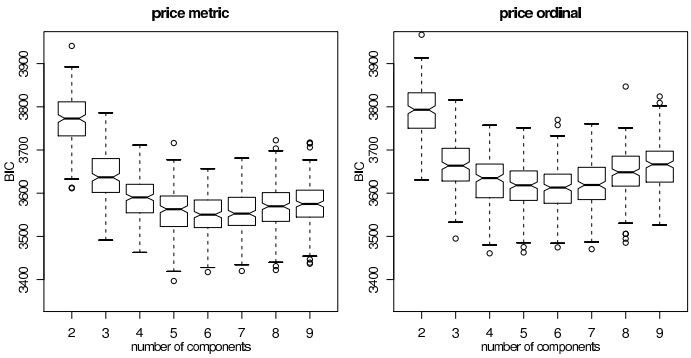


Figure 8: Bootstrap replica of the BIC of latent class regression models for the cakes conjoint data with two to nine components.

	price metric				price ordinal			
# clusters	Estimate	Std. Error	t-value	p-value	Estimate	Std. Error	t-value	p-value
3 - 2	0.0310	0.0179	1.735	0.407	0.0058	0.0181	0.321	1.000
4 - 3	-0.1973	0.0179	-11.020	<0.001	-0.1504	0.0181	-8.299	<0.001
5 - 4	-0.0027	0.0179	-0.152	1.000	-0.0068	0.0181	-0.376	0.999
6 - 5	0.0221	0.0179	1.238	0.759	-0.0105	0.0181	-0.580	0.993
7 - 6	0.0074	0.0179	0.415	0.999	-0.0052	0.0181	-0.289	1.000
8 - 7	-0.0263	0.0179	-1.471	0.595	-0.0132	0.0181	-0.729	0.975
9 - 8	-0.0002	0.0179	-0.016	1.000	-0.0097	0.0181	-0.536	0.995

Table 3: Multiple comparison for successive differences of bootstrapped Rand index values for the cakes data.

To assess if stable segmentation can be undertaken for this data set, we conduct multiple sequential comparisons of corrected Rand indices for two to nine segments (see Table 3). Significant drops in stability can be identified between three and four segments, but the differences in all other comparisons are insignificant. We can conclude that the data set can either be split into two or three segments with the same high level of stability, or into between four and nine segments at a significantly lower constant level of stability.

In fact, the partitions of the maximum likelihood models for three to six segments using price either as metric or ordinal variable are exactly the same, this is not the case for lower or higher numbers of clusters. Using the bootstrap, we know that both BIC and stability also behave similarly in both cases. Summing up, question 1 (how many segments to choose) can be answered independently from the more technical question 2 (which measurement scale to use for the price variable).

The data analyst benefits from this analysis because she or he can be confident that the scale of the price variable will not affect the segmentation solution. Furthermore it is clear that some structure exists in the data. The data analyst has the choice of either preferring a reproducible segmentation approach (in which case the reproducibility of segment across repeated computations is the main quality criterion) or opting for constructive segmentation. For reproducible segmentation the three cluster solution is recommended. If, however, constructive clustering is preferred because three clusters possibly to not deliver small enough market niches, the six cluster solution is recommended by the model based criteria.

5 Conclusions and Future Research

In this paper we demonstrate how bootstrap techniques can be used to assess whether clusters in any given data set represent the natural data structure of the data or whether they are constructed groupings of entities, based on either pseudo-structure or no structure in the data. The fundamental idea is to conduct large numbers of repeated computations with any chosen segmentation algorithm and to assess the similarity of the resulting cluster solutions. If the resulting clusters are the same across repeated computations using many bootstrap samples it can be assumed that natural clusters exist in the data. If, however, repeated computations do not lead to reproducible cluster results, clusters can be assumed to be constructed.

All computations in this paper have been done using the open source statistical computing environment R (R Development Core Team, 2008) using extension packages flexmix (Leisch, 2004) and flexclust (Leisch, 2006). Function `bootFlexclust()` automates the bootstrapping and can be used to easily create plots like the ones shown throughhout the manuscript. Runtimes for the experiments shown above are between 5 and 60 minutes on a current laptop (Intel 2.2 GHz CPU) running Debian Linux, making them feasible for everyday usage. Using the bootstrap approach enables the data analyst to assess segmentation solutions without the risk of over-interpreting a single random computation which is affected both by the sample at hand and the random determinants of the algorithm (such as randomly picked starting points).

We recommend the following set of steps:

- Draw bootstrap samples of the sample of respondents including as many cases as there were respondents in the original survey (200 bootstrap samples worked well in the examples reported in this paper).
- Compute the Rand index or other cluster indices across all replications.
- Inspect box plots or kernel density estimates to assess the reproducability of clustering solutions. For the Rand index, many replications close to 1 indicates the existence of reproducible clusters, while many replications close to 0 indicate no structure in the data, consequently requiring the construction of clusters.
- If the inspection of plots does not provide sufficient clarification, conduct standard significance tests.
- Describe resulting segments and report on the nature of the segments (natural, reproducible or constructive)

The proposed procedure provides guidance to users of cluster analytic and related techniques with respect to the choice of the clustering algorithm and the number of clusters. If natural clusters exist in the data, the choice of the clustering algorithm is less critical and a reliable recommendation regarding the true number of clusters emerges from the bootstrapping procedure. If clusters are constructed, the choice of the clustering algorithms is critical, because each algorithm will impose structure on the data in a different way and, in so doing, affect the resulting cluster solution. Also, the optimal number of clusters cannot easily be determined. In such a situation nothing can safeguard a manager from interpreting an arbitrary segmentation solution because every partition of the unstructured data set will be different and therefore arbitrary if selected. In such a situation the best protection from overinterpreting results is the awareness that one is working with one of many possible splits of the data. It is important for managers to understand that this is probably still preferable to treating the entire market as a homogeneous mass of consumers.

In many cases data structure (other than density clusters) can cause some number of clusters to produce more reproducible results than others. Such comparative reproducibility information can validly be used to guide the data analysts choice of the number of segments. This safeguards managers from interpreting arbitrary segmentation solutions.

References

Aldenderfer, M. S. and Blashfield, R. K. (1984). *Cluster analysis*. Sage Publications, Beverly Hills, USA.

Brusco, M. J. (2004). Clustering binary data in the presence of masking variables. *Psychological Methods*, 9(4):510–523.

Brusco, M. J., Cradit, J. D., and Tashchian, A. (2003). Multicriterion clusterwise regression for joint segmentation settings: An application to customer value. *Journal of Marketing Research*, 40:225–234.

Dibb, S. and Simkin, L. (1997). A program for implementing market segmentation. *Journal of Business and Industrial Marketing*, 12:51–65.

Dimitriadou, E., Dolnicar, S., and Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67(1):137–160.

Dolnicar, S. and Lazarevski, K. (2009). Methodological reasons for the theory/practice divide in market segmentation. *Journal of Marketing Management*. In press.

Dolnicar, S. and Leisch, F. (2000). Behavioral market segmentation using the bagged clustering approach based on binary guest survey data: Exploring and visualizing unobserved heterogeneity. *Tourism Analysis*, 5(2–4):163–170.

Dubes, R. and Jain, A. K. (1979). Validity studies in clustering methodologies. *Pattern Recognition*, 11:235–254.

Dudoit, S. and Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Monographs on Statistics and Applied Probability. Chapman & Hall, New York, USA.

Evans, J. R. and Berman, B. (1997). *Marketing*. Prentice Hall, USA.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41:578–588.

Frank, R. E., Massy, W. F., and Wind, Y. (1972). *Market Segmentation*. Prentice Hall, Englewood Cliffs.

Greenberg, M. and McDonald, S. (1989). Successful needs/benefits segmentation: A user's guide. *The Journal of Consumer Marketing*, 6:29.

Hothorn, T., Leisch, F., Zeileis, A., and Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data*. John Wiley & Sons, Inc., New York, USA.

Kotler, P. (1997). *Marketing Management: Analysis, Planning, Implementation and Control*. Prentice Hall.

Kotler, P. and Armstrong, G. (2006). *Principles of Marketing*. Prentice Hall, Upper Saddle River.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8):1–18.

Leisch, F. (2006). A toolbox for k-centroids cluster analysis. *Computational Statistics and Data Analysis*, 51(2):526–544.

Martinetz, T. and Schulten, K. (1994). Topology representing networks. *Neural Networks*, 7(3):507–522.

Mazanec, J. A., Grabler, K., and Maier, G. (1997). *International City Tourism: Analysis and Strategy.* Pinter/Cassell, London, UK.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

Morritt, R. M. (2007). *Segmentation Strategies for Hospitality Managers: Target Marketing for Competitive Advantage*. Haworth Press, Binghamton, USA.

Myers, J. H. and Tauber, E. (1977). *Market Structure Analysis*. American Marketing Association, Chicago.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Searle, S. R. (1971). *Linear Models*. John Wiley & Sons, New York, USA.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, New York, USA.

Strehl, A. and Gosh, J. (2002). Cluster ensembles — a knowldege reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18:267–276.

Tibshirani, R. and Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.

Wedel, M. and Boer, P. (2002). *Glimmix: A Program for Estimation of Latent Class Mixture and Mixture Regression Models, Version 3.0*. ProGAMMA, Groningen, The Netherlands.

Wedel, M. and Kamakura, W. A. (1998). *Market Segmentation - Conceptual and Methodological Foundations*. Kluwer Academic Publishers, Boston.