Validation of K-means Clustering: Why is Bootstrapping Better Than Subsampling?

Hans-Joachim Mucha and Hans-Georg Bartel

Abstract In simulation studies based on many synthetic and real datasets, we found out that subsampling has a weaker behavior in finding of the true number of clusters K than bootstrapping (Mucha and Bartel 2014, 2015, Mucha 2016). But why? Based on further investigations, here especially concerning the K-means clustering with the comparison of bootstrapping and a special version of subsampling named "Boot2Sub", we try to answer this question. In subsampling, usually a parameter H, the cardinality of the drawn subsample, has to be pre-specified. Its specification means an additional serious problem. The way out would be to take the bootstrap sample but discard multiple points. We call such a special subsampling scheme "Boot2Sub". Then, bootstrapping and subsampling "Boot2Sub" result exactly in the same subset of drawn observations. This way allows us to make fair direct comparisons of the performance of bootstrapping and subsampling. As a result of the assessment of applications to generated and real datasets, the conjecture arises that multiple points play

Hans-Joachim Mucha
Weierstrass Institute for Applied Analysis and Stochastics (WIAS), 10117 Berlin, Mohrenstraße 39, Germany,
✉ mucha@wias-berlin.de

Hans-Georg Bartel
Department of Chemistry, Humboldt University, Berlin, Brook-Taylor-Straße 2, 12489 Berlin, Germany,
✉ hg.bartel@yahoo.de

ARCHIVES OF DATA SCIENCE, SERIES A
(ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. 2, No. 1, 2017

DOI 10.5445/KSP/1000058749/27
ISSN 2363-9881

an important role for the validation of the true number of clusters in K-means clustering.

1 Introduction

Cluster analysis aims to find a partition of a set of I observations $\mathscr{C} = \{1,2,\ldots,I\}$ into K non-empty clusters $\mathscr{C}_k$, $k = 1,2,\ldots,K$. Often, the starting point is a data matrix $\mathbf{X} = (x_{ij})$, $i = 1,2,\ldots,I$, $j = 1,2,\ldots,J$ of I observations and J variables or a $I \times I$ distance matrix $\mathbf{D} = (d_{il})$. The clusters should be stable, i.e. they should be confirmed and reproduced to a high degree if the data set is changed in a non-essential way (Hennig 2007). For example, clustering of a randomly drawn sample should lead to similar results.

Nonparametric bootstrapping is resampling with replacement from the original sample of size I. This very simple technique allows the estimation of the sampling distribution of almost any statistic. An alternative well-known resampling method is subsampling: draw a subsample of a smaller size $H < I$ without replacement. It requires the pre-specification of the parameter H which is a serious drawback. For instance, bootstrapping of the adjusted Rand index (ARI, see Hubert and Arabie 1985) assesses the stability of the original clustering by comparison with each of B bootstrap clustering results. In this paper, always $B = 250$ is used in order to be at the safe side. For the purpose of direct and verifiable comparisons of the performance of bootstrapping and subsampling, exactly the same subsets of drawn observations are investigated. To do so, we take a bootstrap sample but discard multiple points. As a result, the cardinality H^b of the drawn subsample will vary around 63% of the total sample size I. Further, here we investigate only the stability of the well-known and very popular K-means clustering.

The proposed resampling simulation scheme is applied to synthetic datasets (no-structure data and three class data) and to a real dataset.

2 K-means clustering and the proposed resampling scheme

The simplest Gaussian model-based clustering minimizes the well-known sum of squares clustering criterion (Banfield and Raftery 1993) that can be formulated alternatively as the criterion (Späth 1985)

$$V_K = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in \mathscr{C}_k} \sum_{l \in \mathscr{C}_k, l>i} d_{il}, \tag{1}$$

that has to be minimized concerning a partition of a set of I observations $\mathscr{C} = \{1,2,\ldots,I\}$ into K non-empty clusters $\mathscr{C}_k$, $k = 1,2,\ldots,K$, where n_k is the cardinality of the cluster $\mathscr{C}_k$. Here d_{il} is the pair-wise squared Euclidean distance between two observations i and l:

$$d_{il} = d(\mathbf{x}_i, \mathbf{x}_l) = (\mathbf{x}_i - \mathbf{x}_l)^T (\mathbf{x}_i - \mathbf{x}_l), \tag{2}$$

where $\mathbf{x}_i$ and $\mathbf{x}_l$ are the vectors of measurements of the corresponding observations i and l. There are two well-known methods of minimizing (1), the partitional K-means clustering which is considered in the following, and the hierarchical Ward's clustering method (see, for instance, Mucha 2007, Mucha and Bartel 2014, 2015).

How to find out (A) if bootstrapping is better than subsampling, and if so, (B) why? And, with the annoying parameter H of subsampling in mind, how to do it? So, with a really fair comparison of the resampling techniques in mind, let us reformulate (1) with consideration of weights of the observations $m_i \geq 0$, $i = 1,2,\ldots,I$,:

$$V_K = \sum_{k=1}^{K} \frac{1}{M_k} \sum_{i \in \mathscr{C}_k} m_i \sum_{l \in \mathscr{C}_k, l>i} m_l d_{il}, \tag{3}$$

where $M_k = \sum_{i \in \mathscr{C}_k} m_i$ denotes the weight of cluster $\mathscr{C}_k$, and where, of course, $M_k > 0$ has to be ensured. With the implicit understanding that usually the standard weights $m_i = 1$, $i = 1,2,\ldots,I$, are used in clustering, standard weights are not stated explicitly in (1). Standard weights in (3) mean that M_k becomes simply the cardinality of the cluster $\mathscr{C}_k$.

Bootstrapping, i.e., resampling with replacement, is nothing else than playing with the standard weights (masses) of the I observations at random to generate the following "bootstrap weights":

$$m_i^b = \begin{cases} n \text{ if observation } i \text{ is drawn } n \text{ times} \\ 0 \text{ otherwise .} \end{cases} \tag{4}$$

Here, in bootstrapping, clearly $I = \sum_i m_i^b$ holds. Obviously, one can compare bootstrapping with a special subsampling scheme in a really fair way simply by using the following "subsampling weights":

$$m_i^s = \min(m_i^b, 1). \tag{5}$$

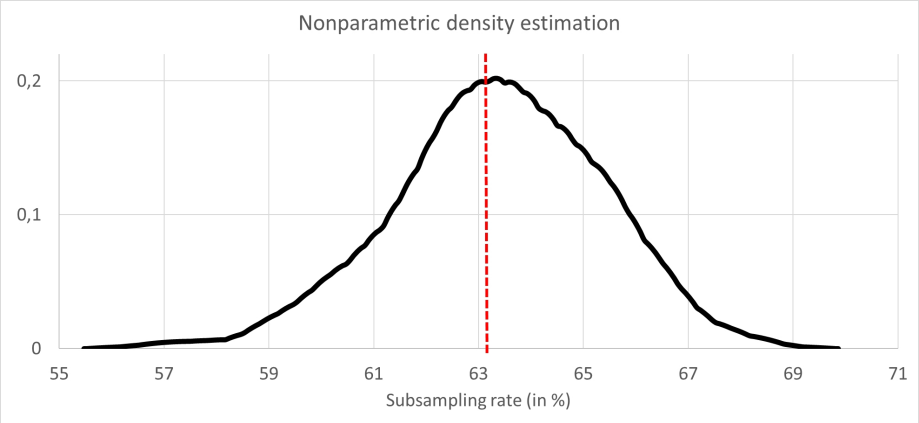

Fig. 1 Visualization of the density of 250 resampling rates H^b/I (in percent) obtained via (5) from the corresponding 250 bootstrap samples.

In difference to (4), this is nothing else than the well-known subsampling, i.e., resampling taken without replacement from the original data. Here, what is special is that we do not need to specify a parameter $H < I$ that defines the cardinality of the drawn subsample. Here, implicitly, the cardinality H^s of the drawn subsample will be around 63% of the total sample size I. Figure 1 shows an example of the density of 250 subsampling rates obtained via (5). Here the average subsampling rate (= 63.37%) is marked by the vertical broken line. Subsampling by (5) based on (4) means simply discarding multiple observations of the corresponding bootstrap sample. At the end, the investigation of stability of K-means clustering by bootstrapping and subsampling is based exactly on the same subsample consisting of the corresponding observations with $m_i^b > 0$ and $m_i^s = 1$, respectively. In the following, let us name this special subsampling scheme "Boot2Sub". In this way, without any doubt, a fair and objective comparison between bootstrapping and subsampling can be realized in an easy manner.

Moreover, the stability of exactly the same original K-means clustering result is investigated by both, bootstrapping and subsampling "Boot2Sub". That is important to say because K-means clustering depends on the initial (usually random) partition to start with, and it ends usually in one of many possible local minima. This is different to hierarchical clustering where one gets usually unique (nested) partitions into K clusters $K = 2, 3, \ldots$ in a parallel fashion

(Mucha 2007). Concerning more general comparisons between bootstrapping and subsampling with quite different values of cardinality H ($H = 90\%$, 75%, and 60%) in validation studies of K-means clustering, the reader is referred to the previous publication of Mucha (2016). It comes out that the higher the subsampling rate the poorer is the ability to find the true number of clusters. Especially, validation with a subsampling rate of $H = 90\%$ (this corresponds in some sense to tenfold cross-validation which is a standard validation technique in supervised classification) performs very bad in finding the true number of clusters.

3 Stability of K-means clustering of no-structure data

First, let us look at no-structure data as shown in Fig. 2 in order to get reference values for the adjusted Rand index R which is our favorite measure of the assessment of stability of clustering (Hubert and Arabie 1985). There are two well-known methods of minimizing (1) and (3), the partitional K-means clustering which is considered in the following, and the hierarchical Ward's method (see, for instance, Mucha and Bartel 2015). Figure 2 shows a no-structure data of $I = 300$ observations in R^2 and a partition of K-means clustering into $K = 3$ clusters. Obviously, there is no structure in the two-dimensional data. It is likely that one gets a quite different K-means clustering result when repeating by starting with another random initial partition. And it is almost sure that K-means clustering of a bootstrap sample of such data results in a quite different partition. Concretely, in this case of no-structure data, the slopes (or angles) of the border lines between the clusters will be completely different to the ones in Fig. 2 (see also Fig. 4 concerning the zones of variation of locations of cluster centroids).

Figure 3 shows the result of the investigation of the stability of K-means clustering by resampling based on the ARI R. Here the average values of five randomly generated no-structure datasets like the one of Fig. 2 are shown. A R-value near 1 (= maximum) means most stable partition. It seems that the K-means clustering of no-structure data is very instable because the ARI values are far from the maximum value 1. That is true for the partitions into all considered numbers of clusters. Obviously, this is because the borderlines between the clusters vary to a high degree, see Fig. 2 and also Fig. 4 concerning $K = 3$. Generally, the adjusted Rand index R seems to be an appropriate mea-

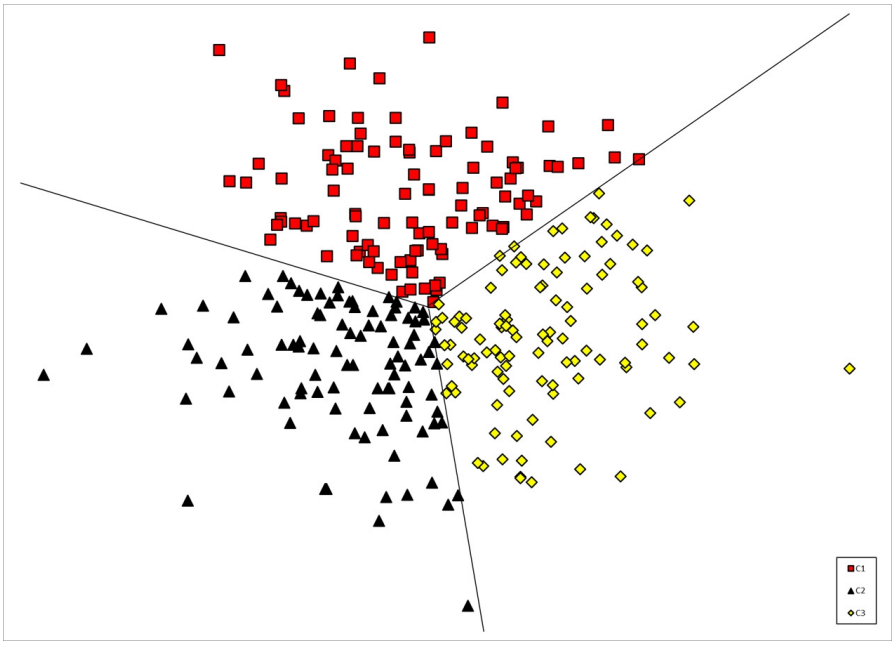

Fig. 2 *K*-means clustering of 300 randomly generated points coming from a standard normal in R^2. The borders of the three clusters are lines.

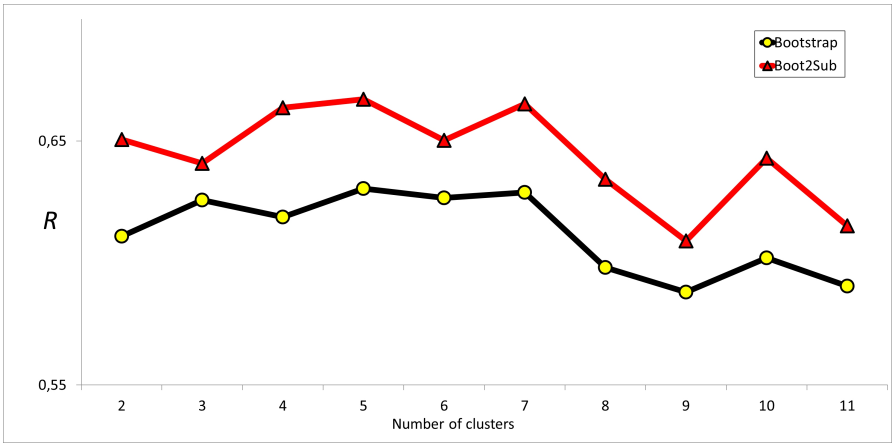

Fig. 3 Investigation of the stability of *K*-means clustering of no-structure data by bootstrapping and "Boot2Sub" based on the adjusted Rand index *R*.

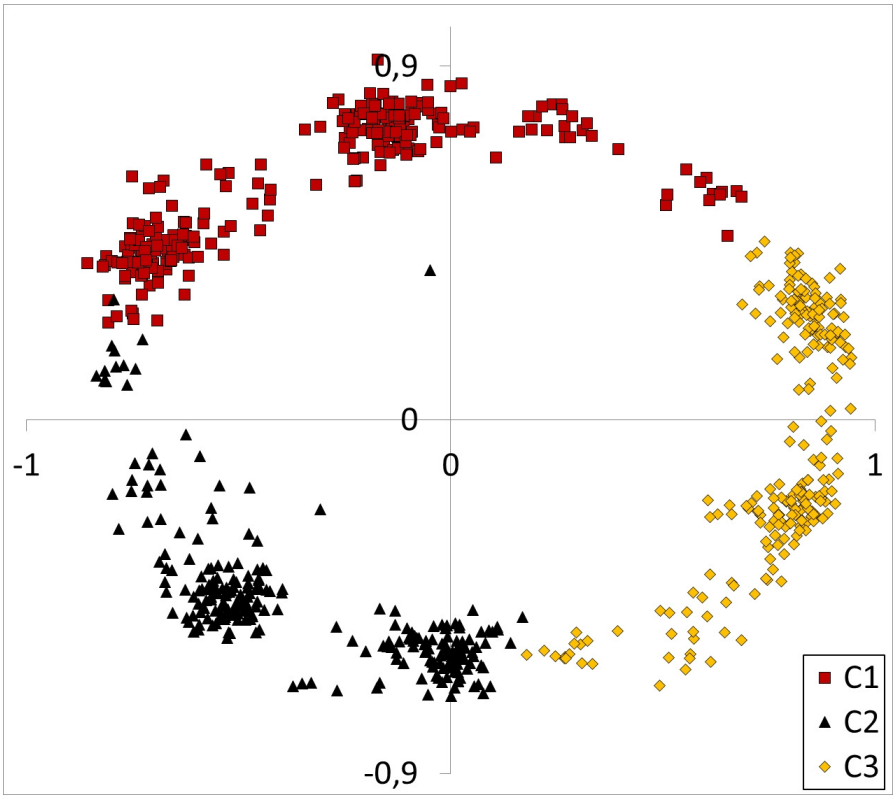

Fig. 4 Visualization of the centroids of K-means clustering of 250 bootstrap samples of the no-structure data of Fig. 2. Here three clusters were investigated.

sure for a decision about the number of clusters because both reference lines of the ARI-values show no trend with the number of clusters as the Rand index does (Hubert and Arabie 1985, Rand 1971).

However, “Boot2Sub” computes clearly higher stability values R for almost all numbers of clusters $K = 2,3,\ldots,11$. Obviously, it overestimates the stability in the case of no-structure data. The "baseline" or reference curve has almost everywhere greater values than the corresponding reference curve of bootstrapping.

Resampling methods can also be used to investigate the variations of the cluster means (centroids, expected values). As expected in the case of the no-structure data, the centroids of K-means clustering of the bootstrap samples

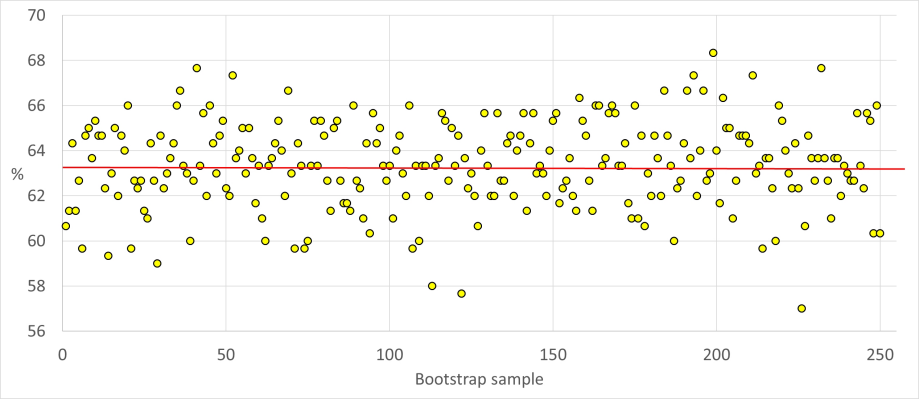

Fig. 5 Visualization of the 250 resampling rates H/I of "Boot2Sub" in percent coming from 250 bootstrap samples. The horizontal line is the average resampling rate. The density estimation of these rates is shown in Fig. 1

look quite instable for $K = 3$ in Fig. 4. All together there are 750 estimates in the case of $K = 3$ clusters and 250 bootstrap samples (750 = 3 * 250). In K-means clustering, the numbering of clusters is arbitrary. Therefore, in order to make the visualization in Fig. 4 possible, one has find out the maximum corresponding bootstrap cluster for each original cluster. Here the maximum Jaccard similarity coefficient between clusters (sets) is used (Hennig 2007).

Figure 5 visualizes the resampling rates of "Boot2Sub" that vary around the rate of approximately 63% (= 63.37%) with a corresponding variance of approximately 3.75.

4 Validation of K-means clustering of synthetic data

Now let us come to our real interest: structured data. As above in the case of no-structure data, several randomly generated three class datasets were investigated. The three Gaussian sub-populations of cardinalities 80, 130, and 90 are generated by pre-specification of the following different mean values (-3, 3), (0, 0), and (3, 3), and the different standard deviations (1, 1), (0.7, 0.7), (1.2, 1.2). Fig. 6 shows the result of the investigation of the stability of K-means clustering by resampling based on the ARI R. As usual in the case of randomly

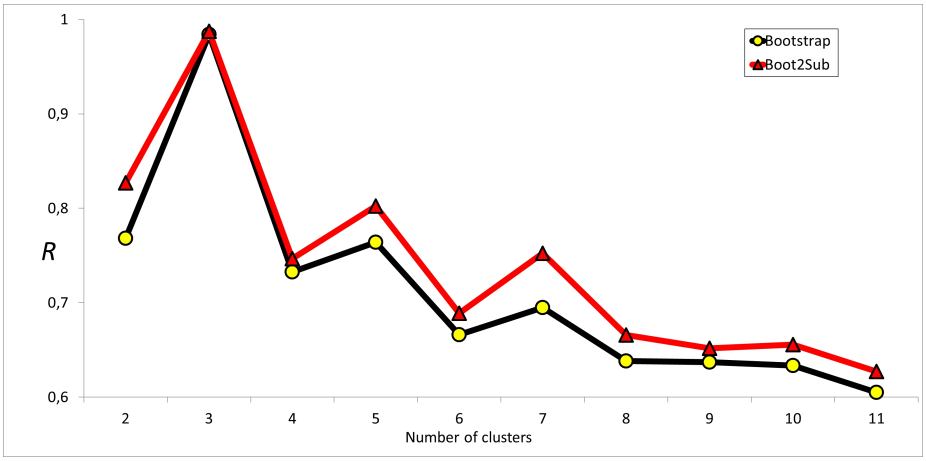

Fig. 6 Investigation of the stability of K-means clustering of three class data by bootstrapping and "Boot2Sub" based on the adjusted Rand index R.

generated data, here the average values of the stability values for the different datasets are shown. Both resampling techniques vote for the true number of clusters. However, "Boot2Sub" rates the stability of the partitions into wrong class numbers $K \neq 3$, higher than Bootstrap.

Figure 7 shows a typical bootstrap sample of such a three class data set in detail. All in all there are 300 observations in such a bootstrap sample. Concretely, in the bootstrap sample shown in Fig. 7, 186 observations out of $I = 300$ have a mass $m_i^b > 0$, and hence these 186 observations get the mass $m_i^s = 1$, i.e., here the cardinality of the drawn subsample is exactly 62%. However, there are 85 multiple observations among the 186 observations: 63 are drawn two times, 17 three times, 4 four times, and 1 six times.

Figure 8 and Fig. 9 show nonparametric density estimates of the original three class data set (see also Fig. 7, but here all four plots are rotated by 90°) and the different sets coming from bootstrapping: "Boot2Sub"-sample, bootstrap sample, and set of multiple points. Obviously, multiple points come mainly from dense regions, and therefore they are responsible for boosting dense regions in their importance for K-means clustering.

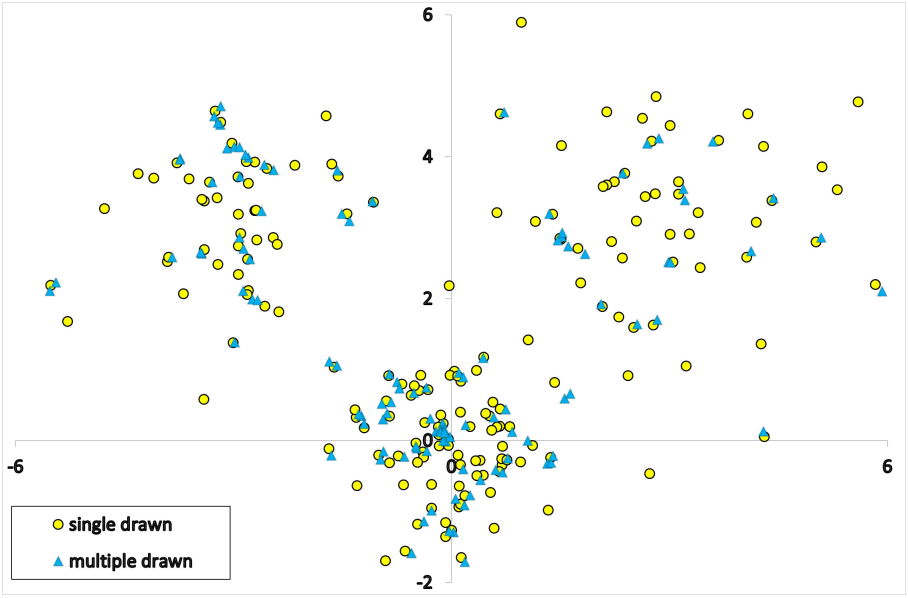

Fig. 7 XY-plot of a bootstrap sample of Gaussian three class data. Here multiple points were jittered and marked by triangles.

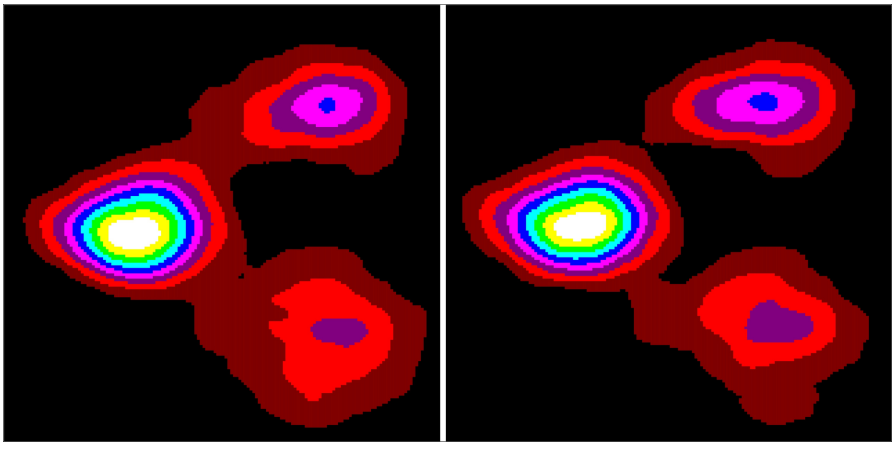

Fig. 8 Cuts at several levels of the bivariate nonparametric density estimate of (a) the original 300 observations of the Gaussian three class data of Fig. 7 (on the right hand side) and (b) the 186 observations of "Boot2Sub" (left).

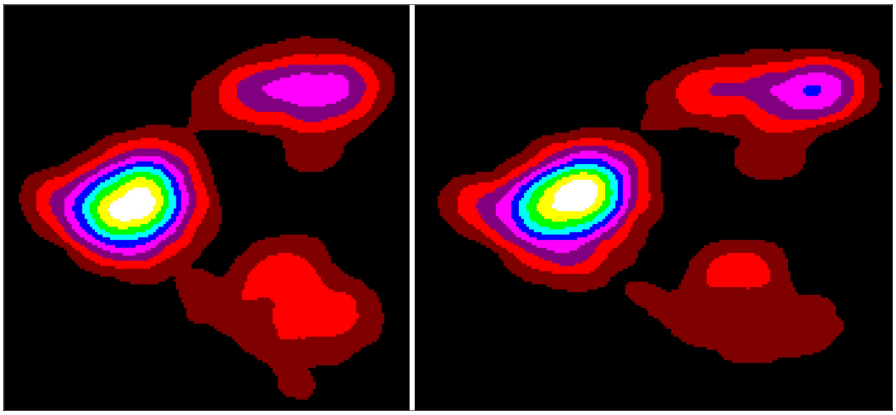

Fig. 9 Cuts at several levels of the bivariate nonparametric density estimate of (a) the 300 observations of bootstrapping of the Gaussian three class data (on the right hand side) and (b) the 114 (= 300 – 186) multiple observations (left).

5 K-means clustering of Swiss banknotes

Let us consider the well-known Swiss banknotes data (Flury and Riedwyl 1988). All together $I = 200$ Swiss banknotes are characterized by six measurements. The 100 genuine bank notes are more homogeneous than the 100 forged ones (see Mucha 2016). Here the true classes are known beforehand, and so, in addition, we are able to look alternatively also at the error rates.

Figure 10 shows the result of the investigation of the stability of K-means clustering of the Swiss bank notes based on the ARI R. Both, bootstrapping and "Boot2Sub" vote for the true number of clusters. However, "Boot2Sub" again overestimates the stability of all other remaining partitions of clustering into the wrong number of clusters.

By the way, the assessment of stability by the ARI seems to be a good choice because the error rates shown in Fig. 11 have a quite similar behaviour for different number of clusters.

6 Summary

First, in case of no-structure data, we found out that the reference line of ARI R of "Boot2Sub"-subsampling shows a much higher degree of stability of K-

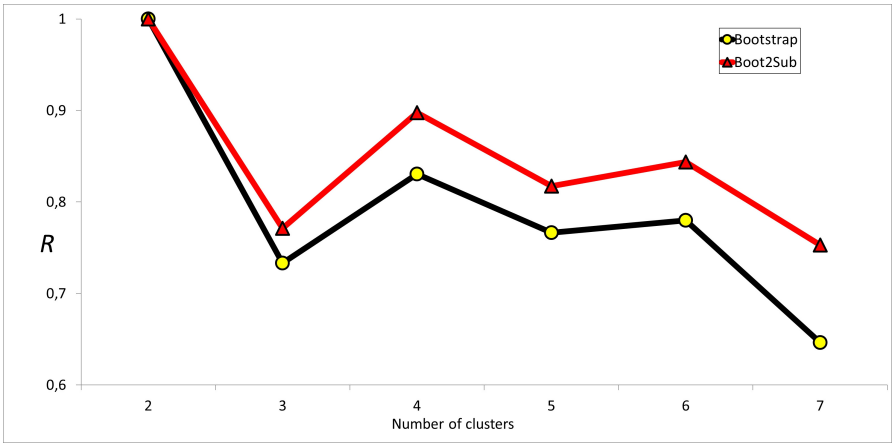

Fig. 10 Swiss banknotes: Investigation of the stability of K-means clustering by bootstrapping and "Boot2Sub" based on the adjusted Rand index R.

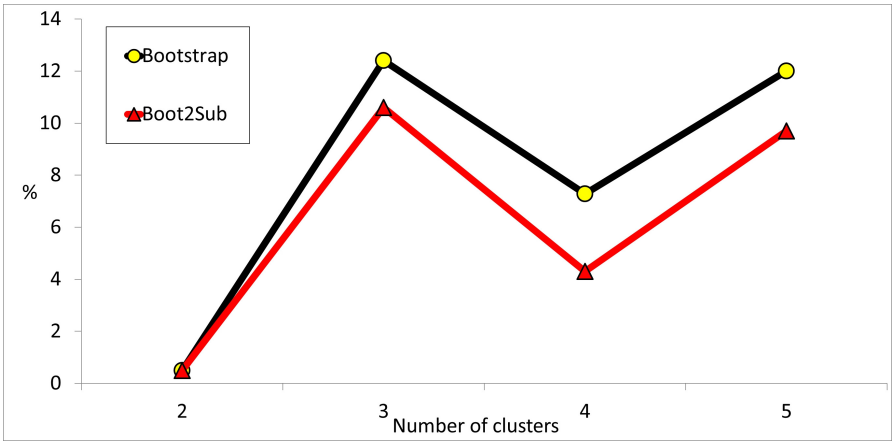

Fig. 11 Swiss banknotes: Error rates in percent of bootstrapping versus "Boot2Sub".

means partitions than the bootstrapping technique. That is, it overestimates the stability of partitions of no-structure data in comparison with bootstrapping. At the end, we were able to find out the most likely reason why bootstrapping outperforms subsampling. Obviously, it is because multiple points give essential additional information in bootstrapping for the investigation of stability of K-means clustering as well as for the determination of the (true) number

of clusters. At this time, it is a conjecture only which requires further investigations accompanied by theoretical considerations. Bootstrapping seems to be the first and best choice. And it is as simple as possible, and there is no additional serious problem with a parameter such as H as it is the case in subsampling.

Of course, for clustering techniques that make no use of weights of observations such as the hierarchical methods Single Linkage and Complete Linkage, the results of the investigation of stability by bootstrapping and "Boot2Sub" are identical. That is because such methods are not affected by multiple points.

References

Banfield JD, Raftery AE (1993) Model-based gaussian and non-gaussian clustering. Biometrics 49(3):803–821, DOI 10.2307/2532201, URL http://www.jstor.org/stable/2532201

Flury B, Riedwyl H (1998) Multivariate statistics: A practical approach. Biometrical Journal 32(5):640–640, DOI 10.1002/bimj.4710320519

Hennig C (2007) Cluster-wise assessment of cluster stability. Computational Statistics & Data Analysis 52(1):258 – 271, DOI 10.1016/j.csda.2006.11.025, URL http://www.sciencedirect.com/science/article/pii/S0167947306004622

Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2(1):193–218, DOI 10.1007/BF01908075

Mucha HJ (2007) On Validation of Hierarchical Clustering, Springer, Berlin, pp 115–122. DOI 10.1007/978-3-540-70981-7_14

Mucha HJ (2016) Assessment of stability in partitional clustering using resampling techniques. Archives of Data Science, Series A 1(1):21–39, DOI 10.5445/KSP/1000058747/02

Mucha HJ, Bartel HG (2014) Soft Bootstrapping in Cluster Analysis and Its Comparison with Other Resampling Methods, Springer International Publishing, Cham, pp 97–104. DOI 10.1007/978-3-319-01595-8_11

Mucha HJ, Bartel HG (2015) Resampling Techniques in Cluster Analysis: Is Subsampling Better Than Bootstrapping?, Springer, Berlin, pp 113–122. DOI 10.1007/978-3-662-44983-7_10

Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66(336):846–850, DOI 10.

1080/01621459.1971.10482356, URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356

Späth H (1986) Cluster dissection and analysis: theory, fortran programs, examples. Biometrical Journal 28(2):182–182, DOI 10.1002/bimj.4710280207