Selection of the number of clusters via the bootstrap

Yixin Fang*
Division of Biostatistics, Department of Environmental Medicine
New York University
Email: Yixin.Fang@nyumc.org

Junhui Wang
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Email: junhui@uic.edu

Abstract

Here the problem of selecting the number of clusters in cluster analysis is considered. Recently, the concept of clustering stability, which measures the robustness of any given clustering algorithm, has been utilized in Wang (2010) for selecting the number of clusters through cross validation. In this manuscript, an estimation scheme for clustering instability is developed based on the bootstrap, and then the number of clusters is selected so that the corresponding estimated clustering instability is minimized. The proposed selection criterion's effectiveness is demonstrated on simulations and real examples.

Keywords: Cluster analysis; K-means; Spectral clustering; Stability.

*Correspondence to: Yixin Fang, 650 First Avenue, Room 551, New York, NY 10016, U.S.A.

1 Introduction

The goal of cluster analysis is to assign observations into clusters so that observations in the same cluster are similar in some sense. Popular methods include k-means clustering (MacQueen, 1967), hierarchical clustering (Johnson, 1967), k-medoids clustering (Kaufman and Rousseeuw, 1990), and spectral clustering (Shi and Malik, 2000). Many selection criteria have been proposed for selecting the number of clusters in a clustering algorithm. Most of them are based on between-cluster and/or within-cluster sum of squared distances; to name just a few, Calinski and Harabasz (1974), Hartigan (1975), and Krzanowski and Lai (1985). Milligan and Cooper (1985) conducted comprehensive simulation studies to evaluate the performances of 30 different such procedures. Additionally, other procedures such as the silhouette statistic proposed by Kaufman and Rousseeuw (1990), the gap statistic proposed by Tibshirani, Walther and Hastie (2001), and the jump statistic proposed by Sugar and James (2003) can also be applied to select the number of clusters.

Recently, the concept of clustering stability has drawn great attention from statisticians. Steinley (2008) developed a procedure, called stability analysis in his paper, for selecting the number of clusters by repeatedly performing k-means many times with different random initializations. Steinley (2008) also mentioned three other types of stability: Bryant (2002) developed a method to determine the general stability of the data with respect to the choice of hierarchical clustering methods; Cheng and Milligan (1996) and some other methods were concerned with character stability when characters are added to or removed from the system; and McIntyre and Blashfield (1980) defined stability as the ability of a cluster solution to be continually recognized in different random samples of the general population.

Here we are concerned with the fourth type of stability, the one measuring the clustering robustness against the randomness in the sample. Also see Fowlkes and Mallows (1983), Gnanadesikan (1997), Ben-Hur, Elisseeff, and Guyon (2002), Lange *et al.* (2004), Ben-David,

von Luxburg, and Pal (2006), and the references therein. As discussed in Wang (2010), the intuition is that if we repeatedly draw samples from the population and apply the given clustering algorithm, a good one should produce clusterings that do not vary much from one sample to another. The stability measure is assumption free and applicable to both distance based and non-distance based clustering algorithms.

It has been proposed to select the number of clusters as the one maximizing the clustering stability. Since maximizing the clustering stability is equivalent to minimizing the clustering instability, an estimation scheme for the clustering instability based on modified cross-validation was developed in Wang (2010). The key idea in Wang (2010) is to split the data into two training datasets and one validation dataset, where the two training datasets are used to construct two clusterings and the validation dataset is used to measure the clustering instability. However, the data splitting reduces the sizes of training datasets and therefore the cross-validation method is inefficient.

In this manuscript, we develop an estimation scheme for clustering instability based on the bootstrap. For an introduction to the bootstrap, please refer to Efron and Tibshirani (1993). The implementation of the bootstrap method is straightforward, and it has a number of advantages. First, the bootstrap samples are of the same size as that of the original data, and therefore the bootstrap method is more efficient. Second, the bootstrap estimate of the clustering instability is the nonparametric maximum likelihood estimate (MLE). Third, the bootstrap method can provide the instability path of a clustering algorithm for any given number of clusters. The last advantage is discussed in detail in Subsection 2.3.

The rest of the manuscript is organized as follows. In Section 2, the concept of clustering stability is introduced and the bootstrap method is developed. In Section 3, the effectiveness of the proposed method is demonstrated on simulations and two real data examples. Some discussion is presented in Section 4.

2 The bootstrap method

2.1 The clustering instability

This subsection is almost reproducing Section 2 in Wang (2010), but with different notation and expressions. Assume that $X^n = \{x_1, \ldots, x_n\}$ is a random sample of size n from some unknown distribution $F(x)$ with $x \in \mathbb{R}^p$, where p is the number of features. A clustering $\psi(x)$ is defined as a mapping $\psi : \mathbb{R}^p \to \{1, \ldots, k\}$, where k is the given number of clusters. A clustering algorithm $\Psi(\cdot; k)$ with a given number of clusters $k \geq 2$ yields a clustering mapping $\Psi_{X^n,k}(x)$ when applied to sample X^n. Here the case of $k = 1$ is excluded as $\Psi(\cdot; 1)$ leads to the same degenerate clustering $\Psi_{X^n,1}(x) = 1$ regardless of X^n.

Definition 1 (Clustering distance). *The distance between any two clustering $\psi_1(x)$ and $\psi_2(x)$ is defined as*

$$d_F(\psi_1, \psi_2) = E_{x^0 \sim F, y^0 \sim F}\{|I\{\psi_1(x^0) = \psi_1(y^0)\} - I\{\psi_2(x^0) = \psi_2(y^0)\}|\}, \tag{1}$$

where $I\{\cdot\}$ is the indicator function, and the expectation is taken over x^0 and y^0, two independent observations sampled from F.

The above expectation equals $Prob\{\psi_1(x^0) = \psi_1(y^0), \psi_2(x^0) \neq \psi_2(y^0)\} + Prob\{\psi_1(x^0) \neq \psi_1(y^0), \psi_2(x^0) = \psi_2(y^0)\}$, so this distance measures the disagreement between two clusterings.

Definition 2 (Clustering instability). *The clustering instability of $\Psi(\cdot; k)$ is defined as*

$$s(\Psi, k, n) = E_{X^n \sim F^n, \tilde{X}^n \sim F^n}\{d_F(\Psi_{X^n,k}, \Psi_{\tilde{X}^n,k})\}, \tag{2}$$

where the expectation is taken over X^n and $\tilde{X}^n$, two independent random samples of size n from F, and $\Psi_{X^n,k}$ and $\Psi_{\tilde{X}^n,k}$ are two clusterings trained from X^n and $\tilde{X}^n$ respectively.

By definition, small values of $s(\Psi, k, n)$ indicate a stable clustering algorithm $\Psi(\cdot; k)$. Then the optimal number of clusters is defined as

$$k_0 = k_0(n) = \text{argmin}_{2 \leq k \leq K} s(\Psi, k, n), \tag{3}$$

where K is preset as the maximum number of clusters to be considered in practice.

2.2 The bootstrap method for selecting the number of clusters

Because the random sample X^n is generated from F, denote $s(\Psi, k, n)$ as $\theta(F)$, a function of F. We wish to estimate $\theta(F)$ on the basis of X^n. Let $\hat{F}$ be the empirical distribution, putting probability $1/n$ on each of the observed values x_i, $i = 1, \ldots, n$. We are interested in using the plug-in estimator, $\theta(\hat{F})$, to estimate $\theta(F)$. Note that $\hat{F}$ is the nonparametric MLE of F, and consequently $\theta(\hat{F})$ is the nonparametric MLE of $\theta(F)$.

Noting that $\theta(F)$ can be written as

$$E_{X^n \sim F^n, \tilde{X}^n \sim F^n} E_{x^0 \sim F, y^0 \sim F} \{ |I\{\Psi_{X^n,k}(x^0) = \Psi_{X^n,k}(y^0)\} - I\{\Psi_{\tilde{X}^n,k}(x^0) = \Psi_{\tilde{X}^n,k}(y^0)\}| \},$$

its plug-in estimator $\theta(\hat{F})$ is equal to

$$E_{X^{n*} \sim \hat{F}^n, \tilde{X}^{n*} \sim \hat{F}^n} E_{x^{0*} \sim \hat{F}, y^{0*} \sim \hat{F}} \{ |I\{\Psi_{X^{n*},k}(x^{0*}) = \Psi_{X^{n*},k}(y^{0*})\} - I\{\Psi_{\tilde{X}^{n*},k}(x^{0*}) = \Psi_{\tilde{X}^{n*},k}(y^{0*})\}| \}.$$

Since the inside expectation can be simplified, the plug-in estimator becomes

$$E_{X^{n*} \sim \hat{F}^n, \tilde{X}^{n*} \sim \hat{F}^n} \{ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |I\{\Psi_{X^{n*},k}(x_i) = \Psi_{X^{n*},k}(x_j)\} - I\{\Psi_{\tilde{X}^{n*},k}(x_i) = \Psi_{\tilde{X}^{n*},k}(x_j)\}| \}.$$

Now it is ready to present the bootstrap method for estimating the clustering instability $s(\Psi, k, n)$, as well as the optimal number of clusters k_0.

Algorithm 1: Bootstrap method for selecting the number of clusters

Step 1. Generate B independent bootstrap sample-pairs $(X_b^{n*}, \tilde{X}_b^{n*})$, $b = 1, \ldots, B$. Each sample consists of n observations generated from empirical distribution $\hat{F}$ with replacement.

Step 2. Construct $\Psi_{X_b^{n*},k}$ and $\Psi_{\tilde{X}_b^{n*},k}$ based on $(X_b^{n*}, \tilde{X}_b^{n*})$, $b = 1, \ldots, B$.

Step 3. For each pair, $\Psi_{X_b^{n*},k}$ and $\Psi_{\tilde{X}_b^{n*},k}$, calculate their empirical clustering distance

$$d_{\hat{F}}(\Psi_{X_b^{n*},k}, \Psi_{\tilde{X}_b^{n*},k}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |I\{\Psi_{X^{n*},k}(x_i) = \Psi_{X^{n*},k}(x_j)\} - I\{\Psi_{\tilde{X}^{n*},k}(x_i) = \Psi_{\tilde{X}^{n*},k}(x_j)\}|. \quad (4)$$

Then the clustering instability $s(\Psi, k, n)$ can be estimated by

$$\hat{s}_B(\Psi, k, n) = \frac{1}{B} \sum_{b=1}^{B} d_{\hat{F}}(\Psi_{X_b^{n*},k}, \Psi_{\tilde{X}_b^{n*},k}). \quad (5)$$

Step 4. Finally, the optimal number of clusters can be estimated by

$$\hat{k} = \hat{k}(n) = \text{argmin}_{2 \le k \le K} \hat{s}_B(\Psi, k, n). \quad \square \tag{6}$$

Hereafter the maximum number of clusters K is set as 10. However, the selection of K depends on one's interest and should be adjusted according to the data. If $\hat{k} = K$, that is the instability achieves its minimum at the boundary, we should increase the value of K.

Algorithm 1 is demonstrated in Figure 1, the schematic diagram of the bootstrap method, where "$\rightarrow$" means data generating and "$\Rightarrow$" means estimating. This diagram is similar to Figure 8.1 in Efron and Tibshirani (1993).

Insert Figure 1 about here

Furthermore, using the bootstrap, we can estimate the standard error of the estimated clustering instability. The following algorithm describes a straightforward bootstrap procedure, and some other more sophisticated procedures are discussed in Section 4.

Algorithm 2: Bootstrap method for estimating the standard error

Step 1. Generate C independent bootstrap samples X_c^{n*}, $c = 1, \ldots, C$. Each sample consists of n observations generated from empirical distribution $\hat{F}$ with replacement.

Step 2. For sample X_c^{n*} , calculate the estimated clustering instability $\hat{s}_B^{c*}(\Psi, k, n)$ based on B bootstrap samples generating from X_c^{n*}.

Step 3. Calculate the sample standard deviation of $\hat{s}_B^{c*}(\Psi, k, n)$, $c = 1, \cdots, C$, which estimates the standard error of the estimated clustering instability, $\hat{s}_B(\Psi, k, n)$. $\square$

Ideally, $\hat{s}_\infty(\Psi, k, n) = \theta(\hat{F})$. In practice, based on our limited experience, moderate large B, say $B = 20$ or 50, can result in very stable estimate of $s(\Psi, k, n)$ and then very stable estimate of k_0. Efron and Tibshirani (1993, Section 6.4) gave some rules of thumb for selecting the number of replications B, and demonstrated that the selection of B can be guided by the

coefficient of variation. Note that we can apply Algorithm 2 to estimate the standard error used in calculation of the coefficient of variation.

2.3 The bootstrap estimates of the instability paths

Although the clustering instability has attracted increasing attention, its theoretical justification remains unclear. One important issue, as pointed out in Krieger and Green (1999) and Ben-David *et al.* (2006), is that any clustering algorithm is asymptotically stable as long as it can be formulated as an optimization problem with a certain objective function that has a unique global minimizer. On the other hand, it has been noted that although the instability measures may converge to zero, the rates of convergence can behave differently when different numbers of clusters are specified (e.g., Shamir and Tishby, 2007).

Denote the instability path for clustering algorithm $\Psi(\cdot; k)$ as $\{s(\Psi, k, N) : N \geq n\}$. To estimate $s(\Psi, k, N)$ for any given N and then visualize the convergence path of the clustering instability for any given number of clusters, we modify Algorithm 1 in what follows. The same idea has been used in Efron and Tibshirani (1993, Section 25.4) for power calculation.

Algorithm 3: Bootstrap method for estimating the instability paths

Step 1. Generate B independent bootstrap sample-pairs $(X_b^{N*}, \tilde{X}_b^{N*})$, $b = 1, \ldots, B$. Each sample consists of N observations generated from empirical distribution $\hat{F}$ with replacement.

Step 2. Construct $\Psi_{X_b^{N*},k}$ and $\Psi_{\tilde{X}_b^{N*},k}$ based on $(X_b^{N*}, \tilde{X}_b^{N*})$, $b = 1, \ldots, B$.

Step 3. For each pair, $\Psi_{X_b^{N*},k}$ and $\Psi_{\tilde{X}_b^{N*},k}$, calculate empirical distance $d_{\hat{F}}(\Psi_{X_b^{N*},k}, \Psi_{\tilde{X}_b^{N*},k})$. Then estimate $s(\Psi, k, N)$ by $\hat{s}_B(\Psi, k, N) = \frac{1}{B}\sum_{b=1}^{B} d_{\hat{F}}(\Psi_{X_b^{N*},k}, \Psi_{\tilde{X}_b^{N*},k})$. □

3 Numerical results

3.1 An illustrative example

We first examine an illustrative example in great detail. The simulated sample contains three clusters in a two-dimensional space, each with size 50 and sampled from bivariate

normal distributions with a common identity covariance matrix and distinct centers $(2, 0)$, $(-1, 2)$, and $(-1, -2)$, respectively. Therefore, the sample size is $n = 150$ and the optimal number of clusters is $k_0 = 3$. An observed sample is displayed in the left panel of Figure 2.

Insert Figure 2 about here

For this observed sample, we apply both the newly proposed bootstrap method and the cross-validation method to select the number of clusters. (Here the cross-validation method is the one based on averaging, that is CV_a in Wang (2010), where the cross-validation method based on voting, CV_v, was also proposed.) In the bootstrap method, 50 bootstrap sample-pairs are generated, while in the cross-validation method, 50 data splittings are performed and the splitting ratio is set as 1/3 (one third as the validation dataset, the other two thirds as two training datasets). Here the k-means algorithm is applied. Both methods produce estimates of the clustering instability for $k = 2, \ldots, 10$. The results are displayed in the right panel of Figure 2. Both methods select the optimal number of clusters as $\hat{k} = 3$, which correctly estimates the true optimal number of clusters.

From the right panel of Figure 2, we also find that the cross-validation estimated instabilities are consistently larger than those estimated using the bootstrap method, because the sizes of two training datasets in the cross-validation method are only $n/3$. Actually, the cross-validation only provides estimates for $s(\Psi, k, n/3)$ instead of $s(\Psi, k, n)$, whereas the bootstrap method provides nonparametric MLE for $s(\Psi, k, n)$.

Figure 2 also displays the estimated standard errors, shown as the bars. Here the standard errors are estimated using Algorithm 2 with $C = 100$. It is found that the standard errors in the bootstrap method are smaller than those in the cross-validation method, due to the same reason that the effect simple size in the bootstrap method is larger.

We go further to see if the bootstrap distribution of $d_{\hat{F}}(\Psi_{X^{n*},k}, \Psi_{\tilde{X}^{n*},k})$ is similar to the true distribution of $d_F(\Psi_{X^n,k}, \Psi_{\tilde{X}^n,k})$. We only consider the case where k is preset as 2.

For this aim, we generate 1000 bootstrap sample-pairs $(X_b^{n*}, \tilde{X}_b^{n*})$ from the given empirical distribution $\hat{F}$, $b = 1, \ldots, 1000$. Then the clustering distances $d_{\hat{F}}(\Psi_{X_b^{n*},2}, \Psi_{\tilde{X}_b^{n*},2})$ are calculated. On the other hand, we generate 1000 random sample-pairs $(X_b^n, \tilde{X}_b^n)$ from the true simulation distribution F, $b = 1, \ldots, 1000$. Then the clustering distances $d_F(\Psi_{X_b^n,2}, \Psi_{\tilde{X}_b^n,2})$ are calculated using Mente Carlo method based on 50 newly generated copies of X^n. The histograms of $\{d_{\hat{F}}(\Psi_{X_b^{n*},2}, \Psi_{\tilde{X}_b^{n*},2}), b = 1, \ldots 1000\}$ and $\{d_F(\Psi_{X_b^n,2}, \Psi_{\tilde{X}_b^n,2}), b = 1, \ldots, 1000\}$ are displayed respectively in Figure 3.

Insert Figure 3 about here

From Figure 3, we find that the two histograms are similar to each other; both of them are in U-shape, with large frequencies at the two ends and small frequencies in the between. Also, the means and the standard deviations of these two histograms are very close. Noting that the two means are respectively the bootstrap estimate and the Monte Carlo estimate of the true clustering instability $s(\Psi, 2, n)$, we are further assured that the bootstrap method gives almost unbiased estimates of the clustering instabilities.

3.2 Four simulation examples in Wang (2010)

We examine the newly proposed bootstrap method using the same four simulation examples as in Wang (2010). The first example consists of two elongated clusters in a three-dimensional space, each of size 100. This is also the example (e) in Tibshirani *et al.* (2001). Each cluster is generated as follows. Set $x_1 = x_2 = x_3 = t$ with t taking 100 equally spaced values from -0.5 to 0.5 and then Gaussian noise with standard deviation 0.1 is added to each feature. Cluster 2 is generated in the same way, except that the value 10 is added to each feature at the end. The second example contains four non-Gaussian clusters in a ten-dimensional space, each of size 100. The clusters reside in a two-dimensional subspace, and are sampled from bivariate exponential distributions with location parameters $(4, 4), (4, -4), (-4, 4), (-4, -4)$

and a common scale parameter 1, and the remaining eight dimensions are noises sampled from standard exponential distribution. The other two are non-distance based examples, the two-moon example and the bull's eye example, whose two realizations are displayed in Figure 4 (a) and (b) respectively. Wang (2010) compared his two cross-validation methods with six other methods and concluded that the cross-validation based on averaging, CV_a, performs the best in these four simulation examples.

We perform the k-means algorithm for the first two examples (distance-based) and the spectral clustering algorithm (Ng *et al.*, 2001) for the other two examples (non-distance-based). The k-means algorithm assigns each observation to the cluster whose center is nearest, and the center is the average of all the observations in the cluster. The spectral clustering algorithm defines a similarity matrix first and then makes use of the spectrum of the similarity matrix to perform principle component analysis for clustering in fewer dimensions.

We compare the bootstrap method with CV_a, along with three recent methods, the silhouette statistic in Kaufman and Rousseeuw (1990), the gap statistic in Tibshirani *et al.* (2001), and the jump statistic in Sugar and James (2003). Kaufman and Rousseeuw (1990) defined the silhouette statistic as $s(k) = n^{-1}\sum_{i=1}^{n}(b(x_i) - a(x_i))/\max\{a(x_i), b(x_i)\}$, where $a(x_i)$ is the averaged distance to other observations in its cluster, and $b(x_i)$ is the averaged distance to observations in its nearest cluster. Tibshirani *et al.* (2001) defined the gap statistic as $\mathrm{Gap}(k) = B^{-1}\sum_{b=1}^{B}\log(W_k^{*b}) - \log(W_k)$, where B reference datasets are generated and W_k and W_k^{*b} are the within-cluster sums of squares from the observed dataset and the bth reference dataset respectively. Sugar and James (2003) defined the jump statistic as $\mathrm{Jump}(k) = \hat{d}_k^{-p/2} - \hat{d}_{k-1}^{-p/2}$, where $\hat{d}_k = p^{-1}\sum_{i=1}^{n}\min_{c_r}(x_i - c_r)'(x_i - c_r)$. The number of clusters is chosen such that $s(k)$, $\mathrm{Gap}(k)$ or $\mathrm{Jump}(k)$ is maximized.

Each simulation example is repeated 50 times and the results are summarized in Table 1. Here the number of bootstrap sample-pairs in the bootstrap method, the number of random

splits in the cross-validation method, and the number of reference datasets are all set as 50.

Insert Table 1 about here

From Table 1, we find that the bootstrap method performs as well as the cross-validation method for selecting the number of clusters, and clearly outperforms the others. The gap statistic also performs very well in Example 1, which is of Gaussian distribution, and in the two-moon example. However, the gap statistic does not perform well in Example 2, where the clusters are non-spherical, and is not working in the bull's eye example, one of the non-distance based examples. Neither the silhouette statistic nor the jump statistic performs well here. Because the sample sizes in these examples are in hundreds, the advantage of the bootstrap method over the cross-validation method is not clear. However, as byproducts, the bootstrap method can provide almost unbiased estimates for the clustering stabilities and the stability paths. The latter will be further demonstrated by two real examples.

3.3 More simulation examples

To further compare the bootstrap method with the cross-validation method and the gap statistic, we revisit the illustrative example but under various conditions, such as the true optimal number of clusters, dimension of features, correlation among features, and clusters overlapping level.

Two groups of simulation settings are considered. In the first group, there are 3 clusters, that is $k_0 = 3$, while in the second group, $k_0 = 5$. Each cluster contains 50 observations, and the number of features p is either 2 or 5. We describe the first two dimensions as follows. In the first group, the three cluster centers are on the vertices of an equilateral triangle with the edge length d. In the second group, four cluster centers are on the vertices of a square and the fifth cluster center is at the middle with the diagonal length $2d$. For the settings where $p = 5$, the remaining three dimensions contain no clustering information.

The observations are generated from multivariate normal distributions with means being their corresponding centers, variances 1, and correlation coefficients ρ, which is equal to either 0 or 0.6. The length d, which controls the clusters overlapping level, is set as 3, 4, or 6. When $d = 6$, clusters are moderately overlapped, while $d = 4$ indicates serious overlapping and $d = 3$ indicates extremely serious overlapping.

Each simulation setting is repeated 50 times and the results are presented in Table 2. Again, the k-means is performed, the number of bootstrap sample-pairs in the bootstrap method, the number of random splits in the cross-validation method, and the number of reference datasets are all set as 50. The maximum number of clusters K is set as 10.

Insert Table 2 about here

From Table 2, we find that the bootstrap method performs slightly better than the cross-validation method. In the settings where $p = 2$, all the three methods performs very well even when the clusters are seriously overlapped, and the correlation between variables does not cause any problem.

Both the stability selection methods are outperformed by the gap statistic. This is not surprising because here the data are generated from Gaussian distribution and the gap statistic is based on the Euclidean distance. In the previous subsection, it has shown that the gap statistics works well with Gaussian clusters or other spherical distributed clusters, but it fails when the clusters are non-spherical.

In the settings where $p = 5$ and three features contain no clustering information, no method is working except in the settings where $k_0 = 3$ and $\rho = 0$. This indicates that it is necessary to consider some feature selection procedure to get rid of redundant features before conducting cluster analysis; for example, Witten and Tibshirani (2010) proposed a general sparse clustering framework and Sun, Wang, and Fang (2011) proposed a regularized k-means algorithm for high-dimensional data.

In the settings where $p = 5$, $k_0 = 3$, and $\rho = 0$, the gap statistic is working with both $d = 4$ and $d = 6$, while the bootstrap method is working only when $d = 6$ and the cross-validation method is not working. This indicates that the stability selection methods are not working well for the cases where clusters are seriously overlapped.

3.4 Two real data examples

We examine the effectiveness of the bootstrap method in two real data examples, comparing it with the cross-validation method in Wang (2010). The first one is iris data (Fisher, 1936), which contains 150 observations from three different species of iris, each with four measures on the length and width of the sepal and petal. The second one is wine data (Aeberhard, Coomans and de Vel, 1992), which consists of 12 measurements for each of 178 alcohols from three different kinds. The iris dataset is available in R, and the wine dataset is downloaded from University of California Irvine machine learning repository. Figure 5 displays the scatterplots of these two datasets.

In both real examples, the true numbers of clusters are unknown, but their available class memberships provide a helpful reference from the true clustering. The results are displayed in Figure 6 and Figure 7 respectively. In both examples, k-means is performed, 50 bootstrap sample-pairs are generated for the bootstrap method, and 50 data-splittings with ratio 1/3 are generated for the cross-validation method.

In the iris example, from the left panel of Figure 6, both the bootstrap and cross-validation methods select the optimal number of clusters as $\hat{k} = 2$, while the dataset contains three species of iris. Similar observation was made in Sugar and James (2003). This example shows the weakness of the stability selection methods when some clusters are seriously overlapped, which has already been shown in Subsection 3.3. We also see that the bootstrap estimated instabilities are larger than those by the cross-validation method for the same reason given in Section 4. From the right panel of Figure 6, we see that the instability path of $k = 2$

converges to zero the fastest.

Insert Figure 6 about here

In the wine example, from the left panel of Figure 7, both the bootstrap and cross-validation methods select the optimal number of clusters as $\hat{k} = 3$, which agrees with the number of classes in the dataset. Similarly, the bootstrap estimated instabilities are larger than those by the cross-validation method. And from the right panel of Figure 7, we see that the instability path of $k = 3$ converges to zero much faster than the other two.

Insert Figure 7 about here

4 Discussion

We propose the bootstrap method for estimating the clustering instability and then selecting the number of clusters. It can be applied to any distance-based or non-distance-based clustering algorithm. It is an alternative to the cross-validation method in Wang (2010), whose selection consistency has been established. Since the bootstrap method provides nonparametric MLE of the clustering instability, which are believed to enjoy many good properties, we don't attempt to work on the selection consistency of the bootstrap method.

The main goal in this manuscript is selecting the number of clusters, but the method can also be used to construct confidence interval estimate for the clustering instability. The jackknife-after-bootstrap method (Efron, 1992), the BC_a (Efron, 1987) and the ABC (DiCiccio and Efron, 1992) methods can provide desirable confidence interval estimates.

The bootstrap method has other applications rather than selection of the number of clusters. For example, the bootstrap method can also be applied to select tuning parameters in the regularized k-means (Sun *et al.,* 2011), the penalized cluster analysis (Fang and Wang, 2011) and the sparse clustering (Witten and Tibshirani, 2010).

Acknowledgement

We thank the Editor, an Associate Editor and two referees for their constructive comments that improve the manuscript significantly.

References

[1] Aeberhard, S., Coomans, D., and de Vel, O. (1992). Comparison of classifiers in high dimensional settings, *Technical Report, 92-02*, Department of Computer Science and Department of Mathematics and Statistics, James Cook University of North Queensland.

[2] Ben-David, S., von Luxburg, U., and Pal, D. (2006). A sober look at stability of clustering, *19th Annual Conference on Learning Theory (COLT 2006).*

[3] Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data, *Pacific Symposium on Biocomputing* **7**: 6-17.

[4] Bryant, P. (2002). More on the stability of hierarchical clustering, *Paper presented at the Classification Society of North America Meeting, Madison, WI.*

[5] Calinski, R. B. and Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics - Simulation and Computation* **3**: 1-27.

[6] Cheng, R. and Milligan, G. W. (1996). K-means clustering methods with influence detection, *Educational and Psychological Measurement* **56**: 833-838.

[7] Efron, B. (1987). Better bootstrap confidence intervals (with discussion), *Journal of American Statistical Association* **82**: 171-200.

[8] Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions, *Journal of Royal Statistical Society, Series B* **54**: 83-127.

[9] DiCiccio, T. J. and Efron, B. (1992). More accurate confidence intervals in exponential families, *Biometrika* **79**: 231-245.

[10] Efron B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall/CRC.

[11] Fang, Y. and Wang, J. (2011). Penalized cluster analysis with applications to family data, *Computational Statistics and Data Analysis* **55**: 2128-2136.

[12] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**: 179-188.

[13] Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierachical clusterings. *Journal of the American Statistical Association* **78**: 553-584.

[14] Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations, 2nd Edition*, John Wiley & Sons, Inc., New York.

[15] Hartigan, J. A. (1975). *Clustering Algorithms*, Wiley, New York.

[16] Johnson, S. C. (1967). Hierarchical Clustering Schemes, *Psychometrika* **2**: 241-254.

[17] Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An introduction to Cluster Analysis*, Wiley, New York.

[18] Krieger, A. W. and Green, P. E. (1999). A cautionary note on using internal crossvalidation to select the number of clusters. *Psychometrika* **64**: 341-353.

[19] Krzanowski, W. J. and Lai, Y. T. (1985). A criterion for determining the number of clusters in a data set, *Biometrics* **44**: 23-34.

[20] Lange, T., Roth, V., Braun, M., and Buhmann, J. (2004). Stability-based validation of clustering solutions, *Neural Computation* **16**: 1299-1323.

[21] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, **1**: 281-297.

[22] McIntyre, R. M. and Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum variance clustering procedure, *Multivariate Behavioral Research* **15**: 225-238.

[23] Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a dataset, *Psychometrika* **50**: 159-179.

[24] Ng, A., Jordan, M. and Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. In *Adv. Neural. Info. Processing Sys. (NIPS2001)*, Ed. T. Dietterich, S. Becker and Z. Ghahramani, pp. 84956. Cambridge: MIT Press

[25] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**: 888-905.

[26] Shamir, O. and Tishby, T. (2007). Cluster stability for finite samples. In *Adv. Neural Info. Processing Sys. (NIPS2007)*, Ed. J. Platt, D. Koller, Y. Singer and S. Roweis, pp. 1297-1304. Cambridge: MIT Press.

[27] Steinley, D. (2008). Stability analysis in K-means clustering, *British Journal of Mathematical and Statistical Psychology* **61**: 255-273.

[28] Sugar, C. and James, G. (2003). Finding the number of clusters in a data set: an imformation theoretic approach, *Journal of American Statistical Association* **98**: 750-763.

[29] Sun, W., Wang, J. and Fang, Y. (2011). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Manuscript.*

[30] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic, *Journal of Royal Statistical Society, Series B*, **63**: 511-528.

[31] Wang, J. (2010). Consistent selection of the number of clusters via cross-validation, *Biometrika* **97**: 893-904.

[32] Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering, *Journal of the American Statistical Association* **105**: 713-726.

Table 1: Four simulation examples in Wang (2010)

Method	Estimated number of clusters								
	2	3	4	5	6	7	8	9	10
Example 1: Two elongated clusters in 3 dimensions									
Silhouette	50	0	0	0	0	0	0	0	0
Gap	50	0	0	0	0	0	0	0	0
Jump	0	0	0	0	4	0	36	7	3
CV_a	50	0	0	0	0	0	0	0	0
Bootsrap	50	0	0	0	0	0	0	0	0
Example 2: Four Exponential clusters in 10 dimensions									
Silhouette	0	1	36	8	4	1	0	0	0
Gap	46	0	1	3	0	0	0	0	0
Jump	0	0	50	0	0	0	0	0	0
CV_a	0	0	50	0	0	0	0	0	0
Bootstrap	0	0	50	0	0	0	0	0	0
The two-moon example									
Silhouette	0	0	1	2	9	5	13	13	7
Gap	50	0	0	0	0	0	0	0	0
Jump	39	0	0	3	0	3	1	3	1
CV_a	50	0	0	0	0	0	0	0	0
Bootstrap	50	0	0	0	0	0	0	0	0
The bull's eye example									
Silhouette	0	0	6	9	2	5	10	9	9
Gap	10	27	10	2	1	0	0	0	0
Jump	0	4	12	11	4	9	3	4	3
CV_a	50	0	0	0	0	0	0	0	0
Bootstrap	50	0	0	0	0	0	0	0	0

Table 2: More simulation examples

			$k_0 = 3$				$k_0 = 5$				
Estimated number of clusters			2	3	4	≥ 5	≤ 3	4	5	6	≥ 7
$p = 2$	$d = 3$	$\rho = 0$ Gap	0	50	0	0	0	0	50	0	0
		$\rho = 0$ CV_a	0	41	4	5	0	0	39	2	1
		$\rho = 0$ Boot	0	48	0	2	0	9	38	2	1
		$\rho = 0.6$ Gap	0	47	3	0	0	0	46	2	2
		$\rho = 0.6$ CV_a	0	40	5	5	0	3	40	3	4
		$\rho = 0.6$ Boot	0	41	4	5	0	4	41	3	2
	$d = 4$	$\rho = 0$ Gap	0	50	0	0	0	0	50	0	0
		$\rho = 0$ CV_a	0	50	0	0	0	0	50	0	0
		$\rho = 0$ Boot	0	50	0	0	0	0	50	0	0
		$\rho = 0.6$ Gap	0	50	0	0	0	0	50	0	0
		$\rho = 0.6$ CV_a	0	50	0	0	0	2	48	0	0
		$\rho = 0.6$ Boot	0	50	0	0	0	6	44	0	0
$p = 5$	$d = 4$	$\rho = 0$ Gap	0	35	9	6	4	1	9	19	17
		$\rho = 0$ CV_a	0	0	1	49	0	0	0	0	50
		$\rho = 0$ Boot	0	5	3	42	0	0	0	0	50
		$\rho = 0.6$ Gap	0	1	3	46	0	4	2	17	30
		$\rho = 0.6$ CV_a	1	0	0	49	0	0	0	0	50
		$\rho = 0.6$ Boot	5	0	0	45	7	0	0	1	42
	$d = 6$	$\rho = 0$ Gap	0	50	0	0	1	0	13	19	17
		$\rho = 0$ CV_a	0	26	17	7	0	0	0	0	50
		$\rho = 0$ Boot	0	50	0	0	0	0	0	0	50
		$\rho = 0.6$ Gap	0	0	5	45	0	2	2	25	21
		$\rho = 0.6$ CV_a	0	0	1	49	0	0	0	0	50
		$\rho = 0.6$ Boot	1	0	5	44	3	0	0	2	45

Figure 1: Schematic diagram of the bootstrap method

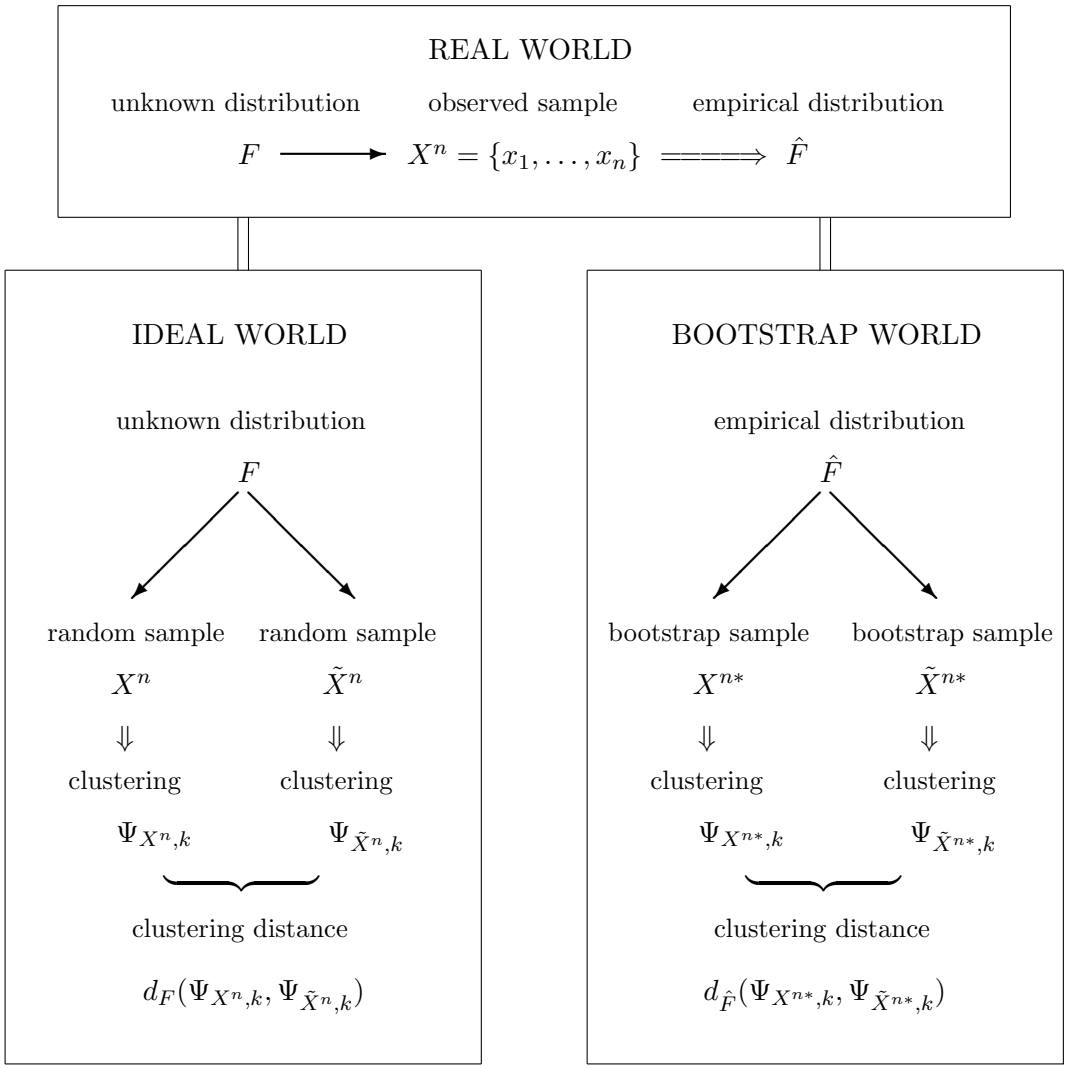

Figure 2: Illustrative example

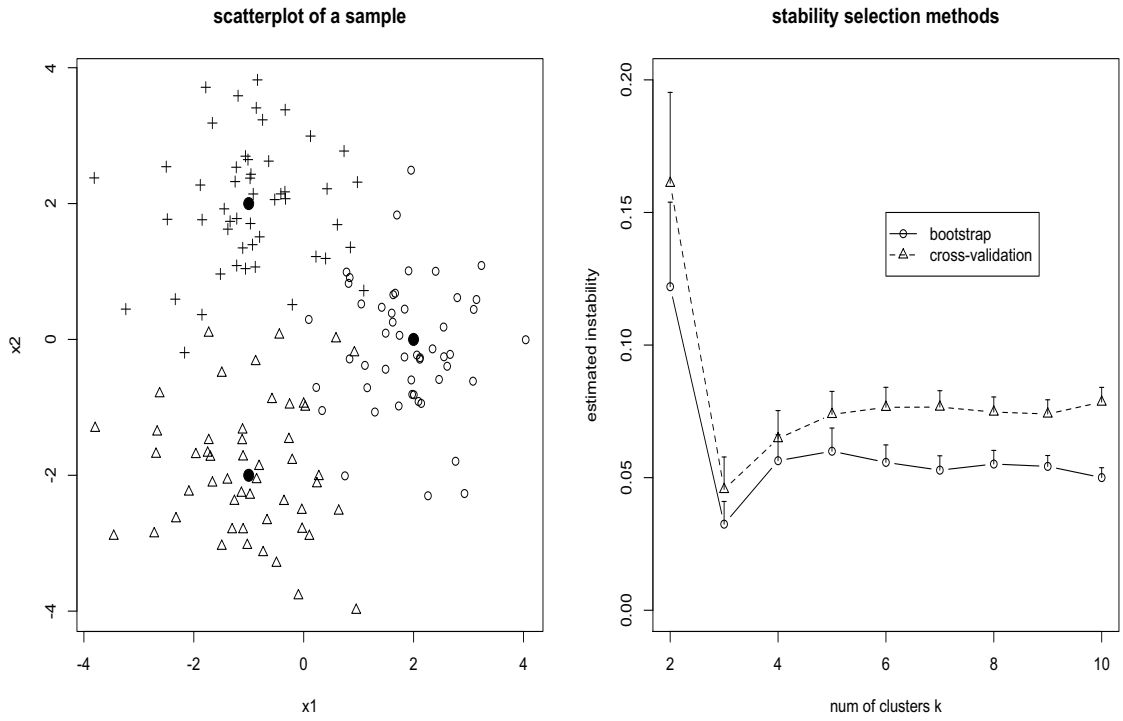

Figure 3: True distribution and bootstrap distribution of clustering distance

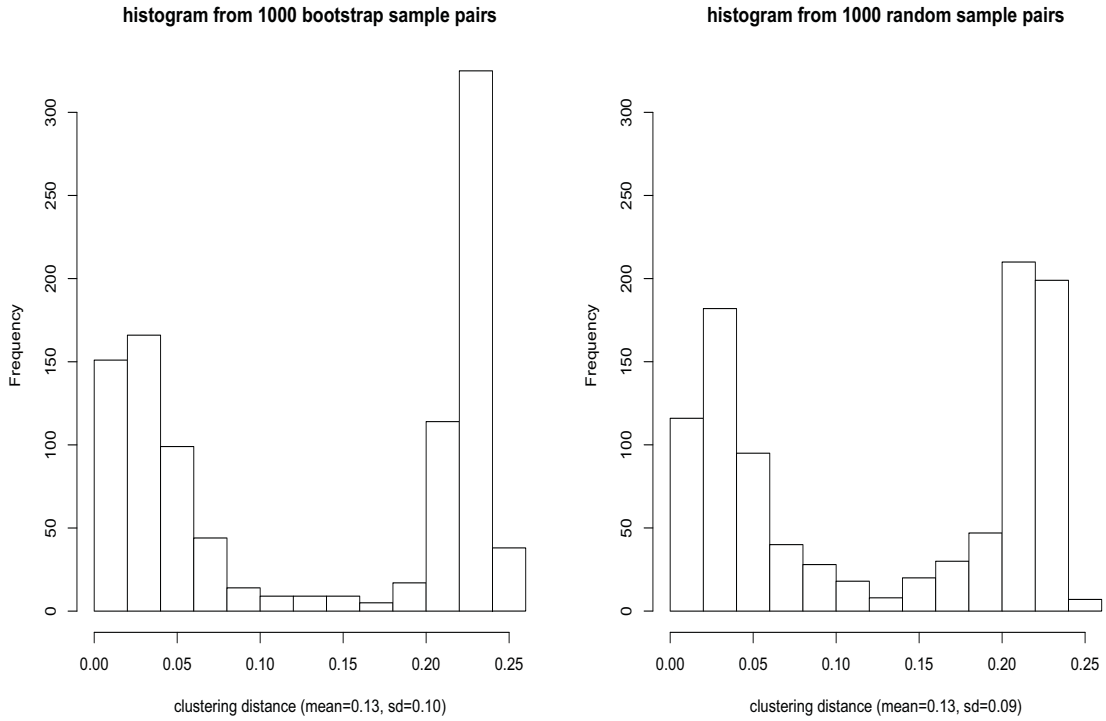

Figure 4: Some scatterplots

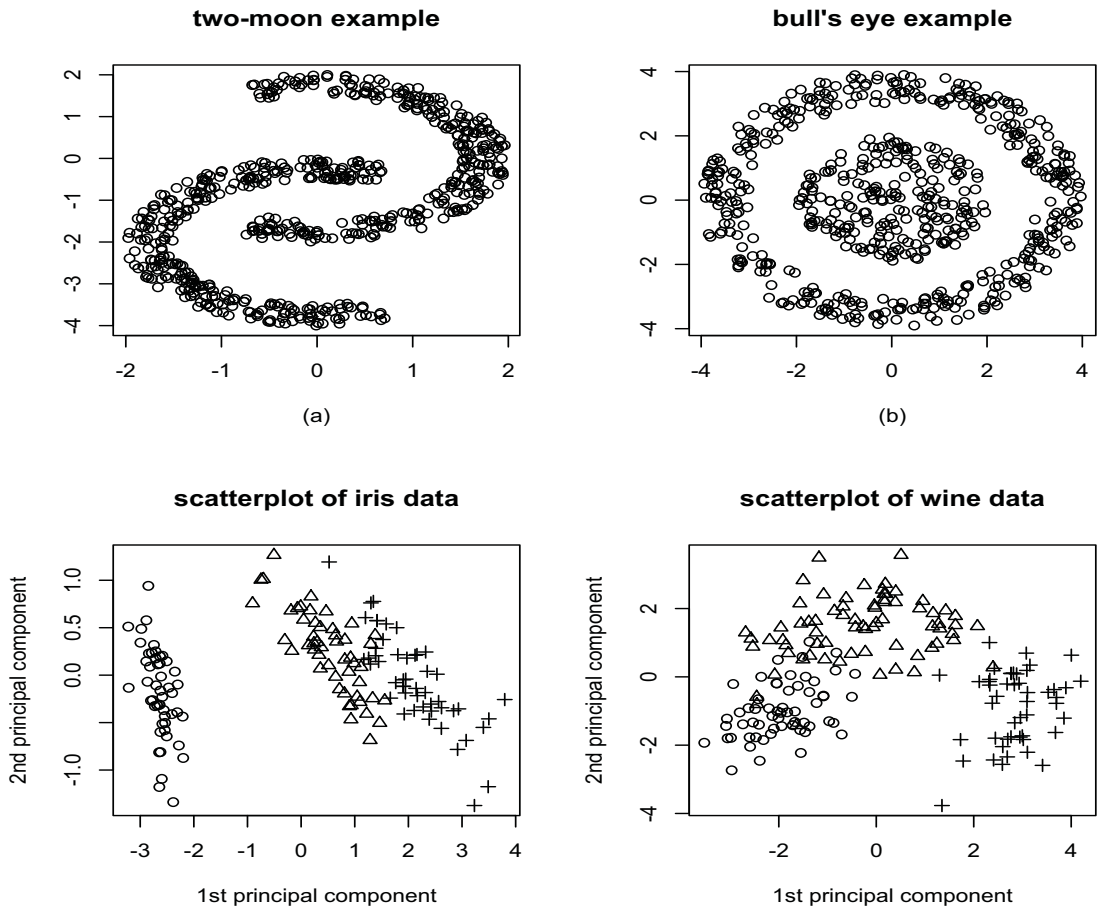

Figure 5: Iris data example ($\hat{k} = 2$)

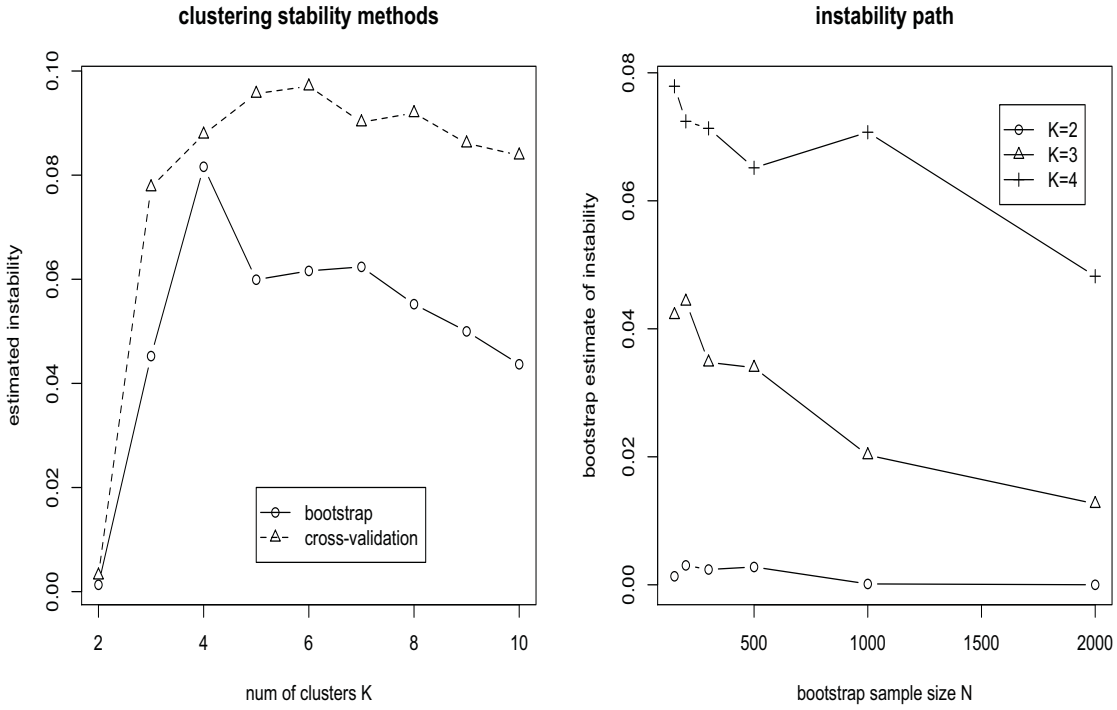

Figure 6: Wine data example ($\hat{k} = 3$)