Weierstraß-Institut für
Angewandte Analysis und Stochastik

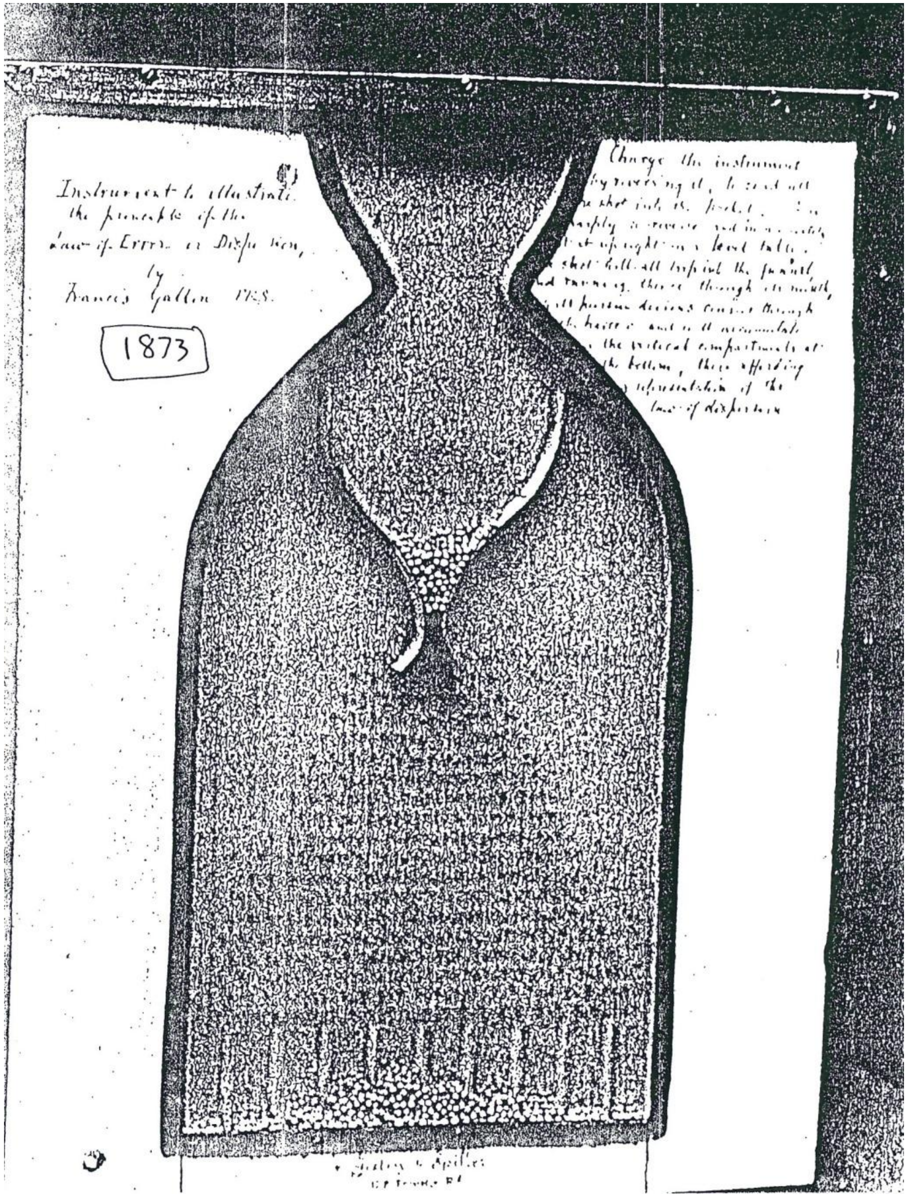
Variable Selection in Cluster Analysis Using Resampling Techniques: A Proposal

Hans-Joachim Mucha

AG DANK/BCS Meeting 2013 at University College London, 8/9 November

Prelude: The historical machine(s) of Galton

140 years ago: Sir Francis Galton (1822-1911) invented the Quincunx, a mechanical device to illustrate the central limit theorem and binomial probabilities. It consists of a vertical board with interleaved rows of pins. Balls are dropped from the top, and bounce left and right as they hit the pins.


Instrument to illustrate the principle of the Law of Error or Dispersion, by Francis Galton F.R.S.
1873
Charge the instrument by reversing it, to send all the shot into its funnel. Then simply reverse and in a moment set it upright on a level table. The shot will all drop into the funnel and running thence through its mouth will pursue devious courses through the [illegible] and will accumulate in the vertical compartments at the bottom, thus affording a representation of the law of dispersion
Tisley & Spiller

Prelude: The historical machine(s) of Galton

Later on Galton improved/modified his *Bean Machine*. With the help of his machines, Galton discovered basic properties and relationships of distributions.

In 1888, Galton realizes that if $(\mathbf{X},\mathbf{Y})$ are bivariate normal, standardized so both have expectation = 0 and standard deviation =1, then $E(\mathbf{X}|\mathbf{Y}=\mathbf{y}) = \mathbf{r}\,\mathbf{y}$ and $E(\mathbf{Y}|\mathbf{X}=\mathbf{x}) = \mathbf{r}\,\mathbf{x}$, and so $\mathbf{r}$ measures the association: The ***„co-relation“***.

Nowadays, computers replace mechanical devises in doing simulation experiments, see animations at: http://vis.supstat.com/2013/04/bean-machine/ . It's a pitty that Galton can't play with the electronic bean machine …

Outline

- Introduction
- Non-parametric resampling techniques
- Variable selection in clustering: a proposal
- Examples
- Summary

Introduction

- Variable selection is a well-known problem in many areas of multivariate statistics such as classification and regression. The hope is that ***the structure of interest may be contained in only a small subset of variables.***
- In contradiction to supervised classification such as discriminant analysis, variable selection in cluster analysis is a much more difficult problem because usually ***nothing is known about the true class structure.***
- In addition, in clustering, variable selection is highly related to the main problem of the determination of the number of clusters K to be inherent in the data.

Introduction

- There are many papers on variable selection in clustering, mainly based on special cluster separation measures such as the Davies and Bouldin (1979) criterion: ratio of within-cluster dispersions and between-cluster separation.
- For instance, Steinley and Brusco (2008) compared eight different variable selection procedures.
- Here we present a ***general approach to variable selection using non-parametric resampling techniques*** based on criteria of stability such as the adjusted Rand's index (ARI). General means, it makes use only of measures of stability of partitions, and so it can be applied to almost any cluster analysis method.

Non-parametric resampling techniques

- Usually, the starting point of cluster analysis is a data matrix $\mathbf{X} = (x_{ij})$ with I observations and J variables.
- Cluster analysis means finding a partition of the set of I observations into K non-empty clusters C_k, $k = 1, 2, ..., K$.
- The C_k should be stable, i.e. they should be reproduced to a high degree if the data set is changed in a non-essential way. Thus, ***clustering of a random drawn sample of the data should lead to similar results.***
- Obviously, the stability of clusters should decrease if we add some noisy (no-structure) variables J+1, J+2,… So, it appears plausible to start variable selection in clustering with as few variables as possible and to proceed by adding new ones until the stability increases.

Non-parametric resampling techniques

- Non-parametric ***bootstrapping*** is a statistical method for estimating the sampling distribution of an estimator by sampling ***with replacement*** from the original sample.

- Well-known alternative resampling methods are ***sub-sampling*** (draw a subsample to a smaler size ***without replacement***) and jittering (add noise to every single observation), and combinations of simulation schemes.

- These resampling techniques and simulation schemes allow the estimation of the sampling distribution of almost any statistic such as ARI or Jaccard measure. The latter assesses the ***stability of every original cluster*** C_k ***by the mean value (or median)*** γ_k over all Jaccard values of the bootstrap samples.

Non-parametric resampling techniques

- For a decision about the number of clusters K (as it is usual using ARI), ***an averaged Jaccard value*** γ_K ***regarding all*** γ_k ***of individual clusters*** C_k, k= 1,2,…, K, of a partition is recommended:

$$\gamma_K = \frac{1}{I}\sum_{k=1}^{K} n_k \gamma_k \quad , \text{where } \; I = \sum_{k=1}^{K} n_k$$

- Both the ARI R (Hubert and Arabie 1985) and the averaged Jaccard index γ_K are most appropriate to decide about the number of clusters K, i.e. to assess the stability of a partition consisting of all clusters C_1, C_2 , …, C_K.
- Alternative well-known measures of stability in cluster analysis are Dice, and Fowlkes and Mallow.

Selection of variables in clustering: a proposal

Here we propose a basic bottom-up variable selection:

1. The starting point is an assessment of the evidence of univariate clustering results. Concretely, ***we are looking for the most stable univariate clustering*** (i.e., the best variable) with respect to indexes such as ARI or Jaccard.

2. Subsequently, ***we are looking for the best partner of the variable found in step 1.*** The hope is to find the most stable bivariate clustering in that way.

3. We are going to find a third partner (variable) of the two variables found in step 2. Furthermore, ***we proceed the search for next variables as long as an "essential" improvement of the stability of the clustering is realized.***

Selection of variables in clustering: a proposal

- The computational complexity decreases with the number of steps: J univariate (original) clustering results have to be assessed, J-1bivariate ones, J-2 trivariate ones, and so on.
- ***This basic variable selection procedure can be modified in several ways:***
  - starting with J*(J-1)/2 bivariate clustering results,
  - or, starting with J*(J-1)*(J-2)/6 trivariate ones,
  - evaluate statistically the J ARI of univariate clustering results (=J*(J-1)/2 pairs) (Carmone et al. 1999), or the J-1 ARI of bivariate clustering …,
  - in between, switching to a top-down step.

Selection of variables in clustering: a proposal

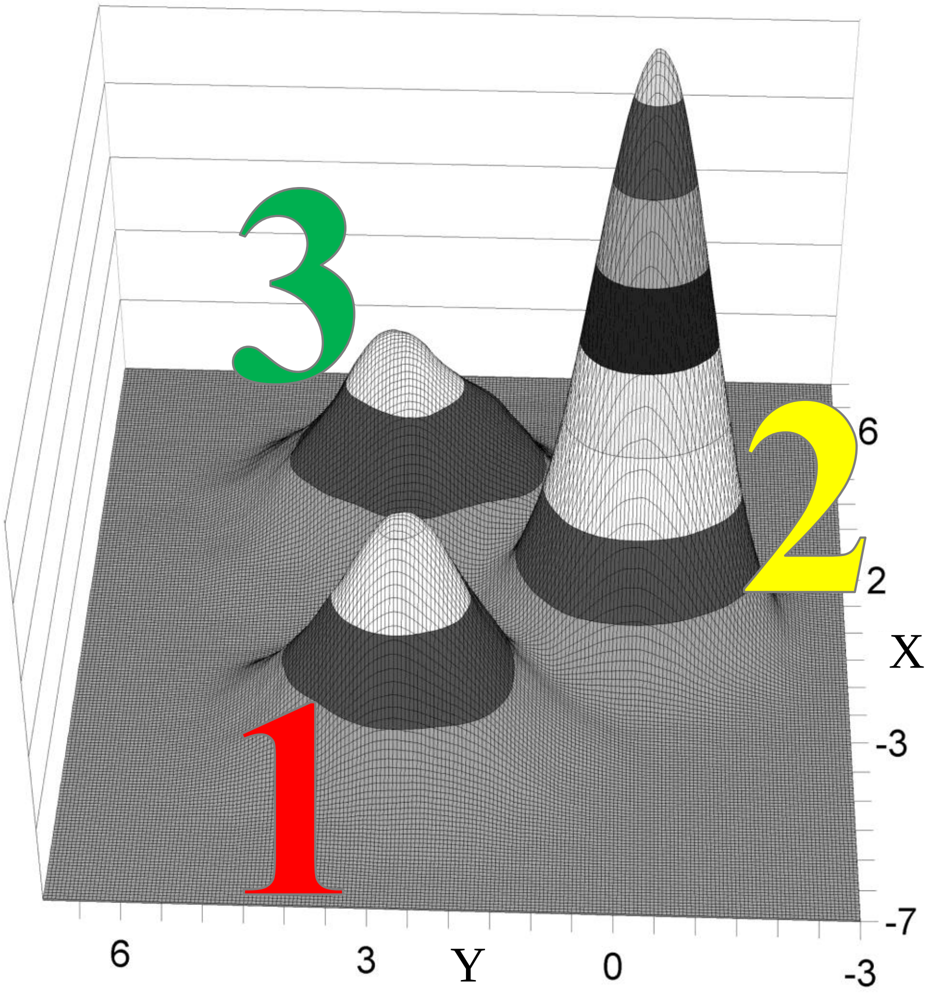
- However, the computational complexity should be taken into consideration. So, starting with the assessment of trivariate clustering results seems to be unrealistic …
- There are two main families of clustering techniques, hierarchical and partitional clustering.
- ***Hierarchical clustering*** looks fit and proper for our resampling proposal because of the ***(usual) unique and parallel clustering of the*** I ***observations into partitions of*** K**=2,** K**=3,…** ***clusters.*** In addition, pairwise distances, the starting point of hierarchical cluster analysis, are not affected by bootstrapping/subsampling.

Selection of variables in clustering: a proposal

- In contradiction, the results of ***partitional (iterative) clustering*** methods are dependent on the initial partition into a fixed number of clusters K. Usually, 50 different initial partitions are used to get up to 50 different locally optimal solutions. The best solution is taken for the investigation of stability.
- Moreover, you have to do this for different K (K=2, K=3, K=4,…).
- Finally, you have to do all things outlined above also for each bootstrap sample (or subsample). So, our proposal starts with altogether $50{*}K_{\max}{*}(B+1){*}J$ univariate partitional clusterings… (B is the number of bootstrap samples).

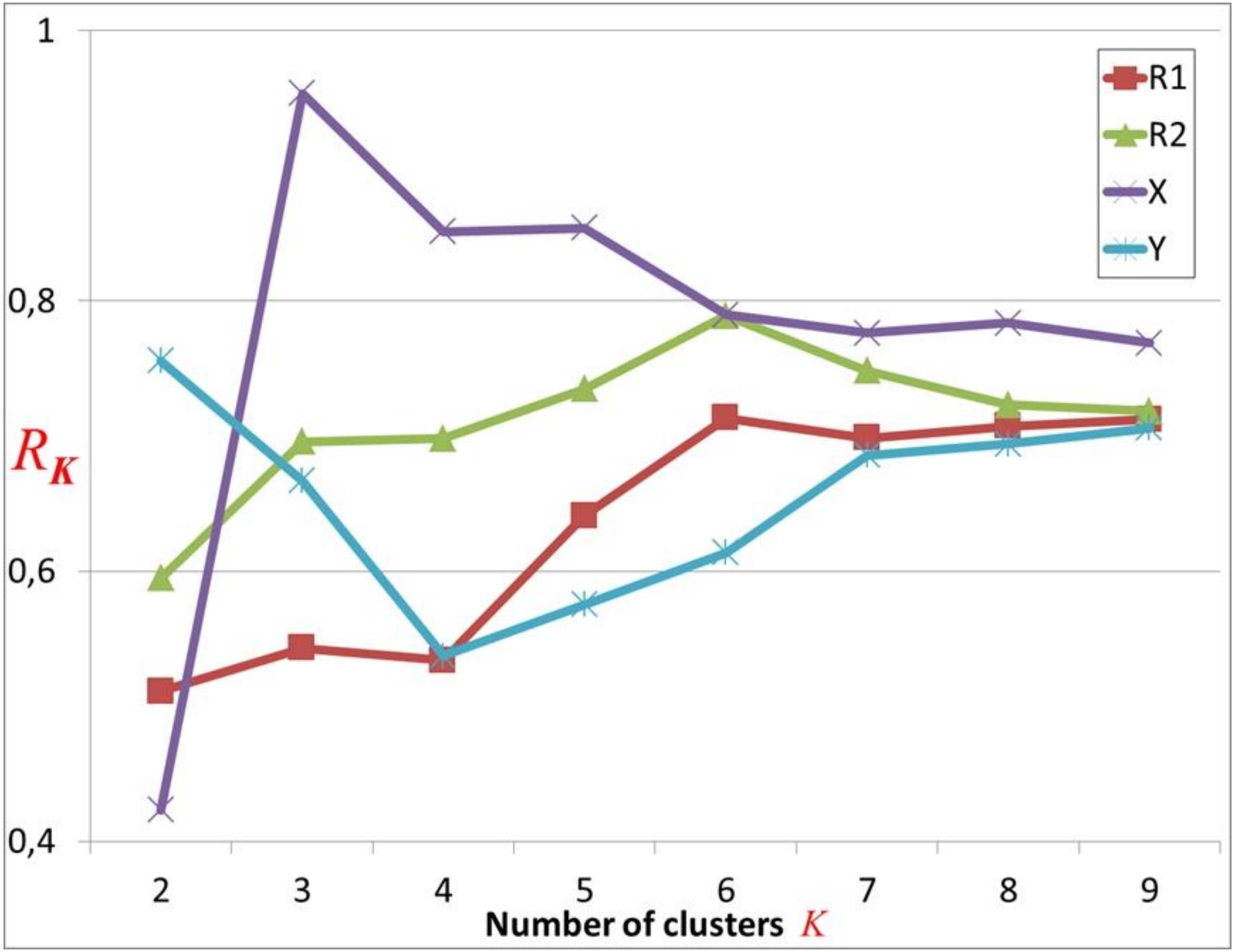
Example: Hierarchical clustering of synthetic data

Bivariate density surface of the variables X and Y of four-dimensional three class data. The other two variables R1 and R2 are masking variables without any class structure. Concretely, they are uniformly distributed in (-5; 5). The Gaussian sub-populations were generated with the following different parameters: cardinalities 80, 130, and 90, mean values (-3 , 3), (0 , 0), and (3 , 3), and standard deviations (1, 1), (0.7, 0.7), and (1.2,1.2).



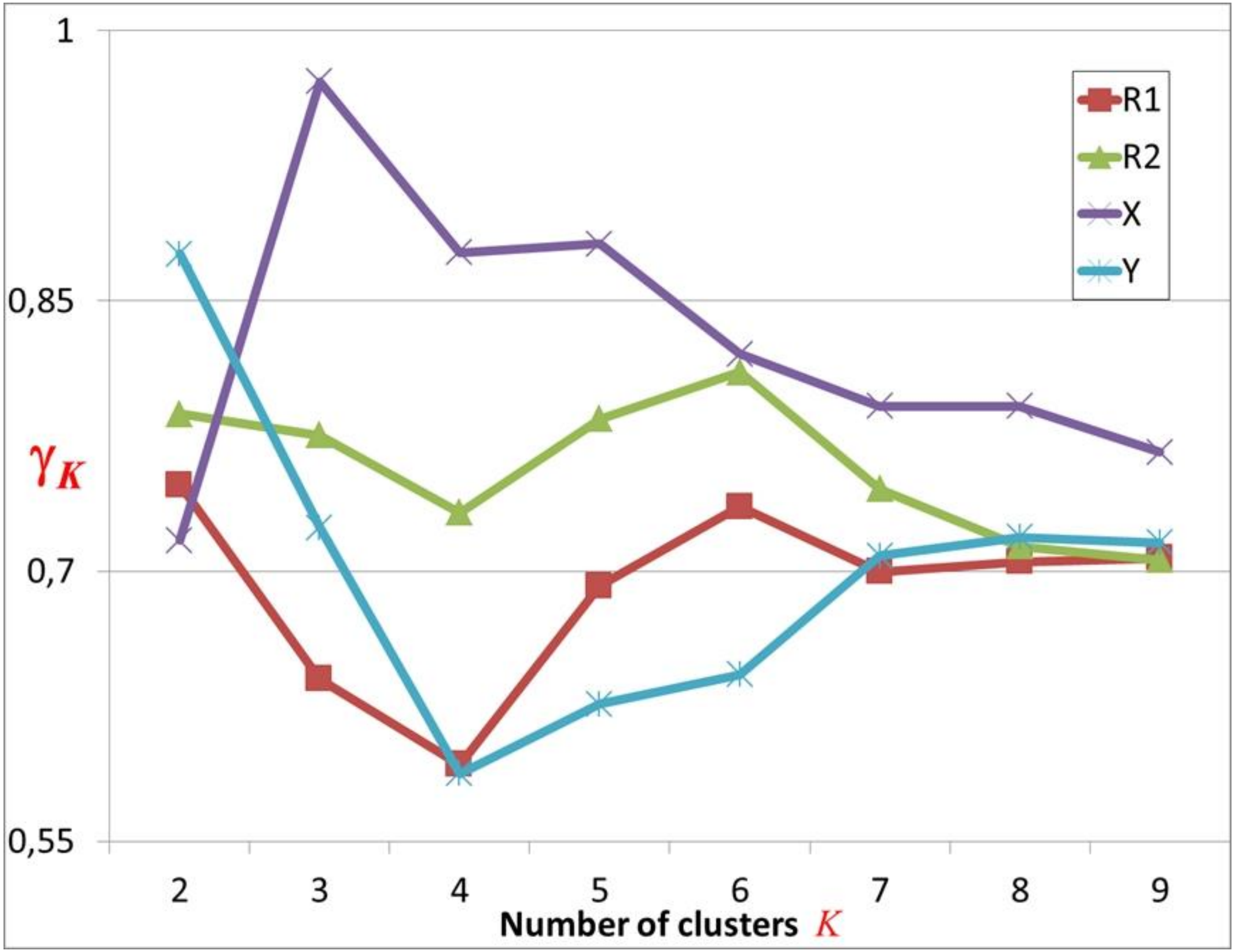
Example: Hierarchical clustering of synthetic data

Investigation of the stability of the **univariate clustering results** based on the adjusted Rand index R coming from comparisons with cluster analyses of 250 sub-samples of cardinality 180 (60%). Ward's method is used.



Example: Hierarchical clustering of synthetic data

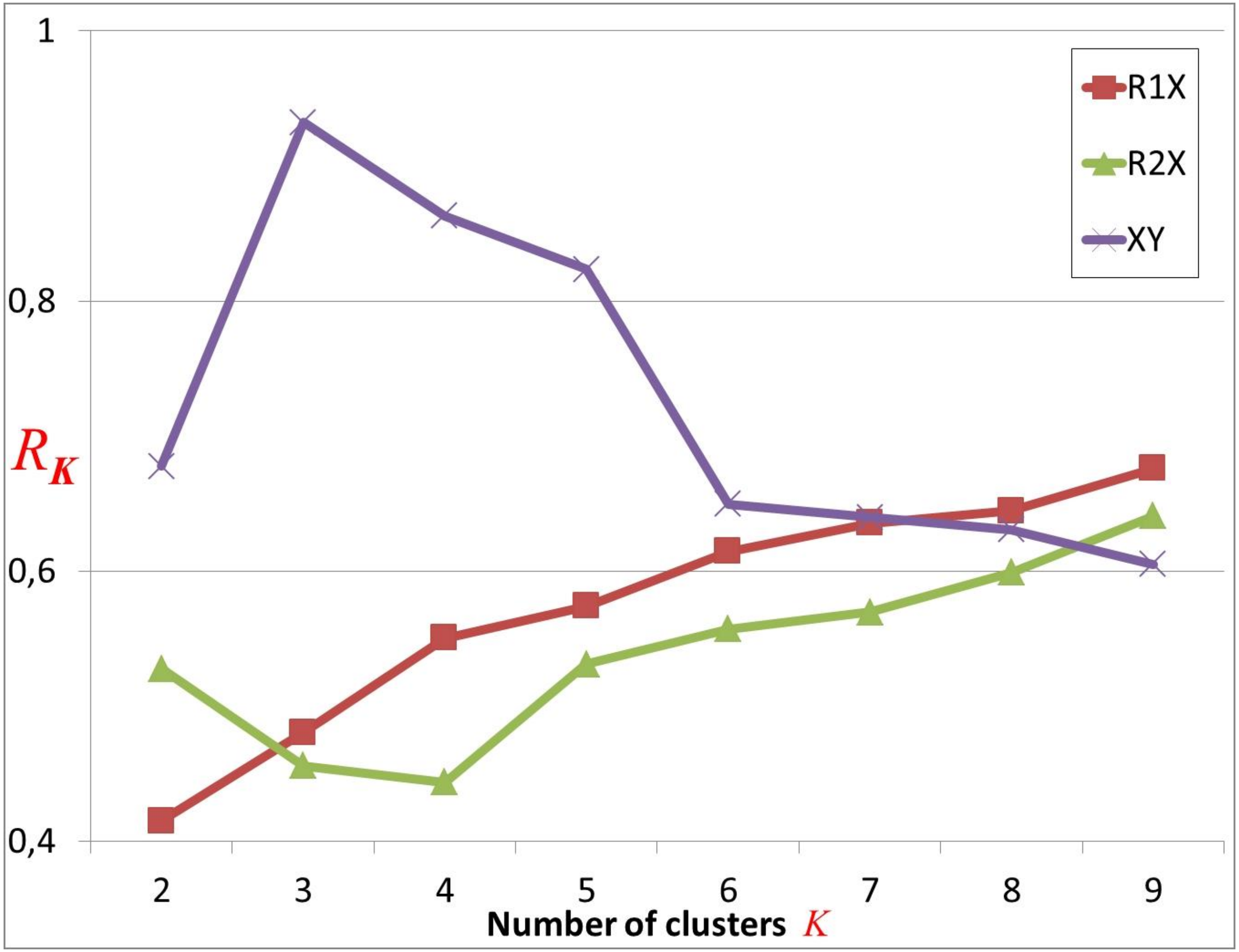
The similarity measure "averaged Jaccard index" γ_K behaves similar to the adjusted Rand index R: clustering based on variable X is most stable for K=3 clusters (with an additional most steep rise from K=2 to K=3 .



(X: 32 errors, Y: 78 errors)

Example: Hierarchical clustering of synthetic data

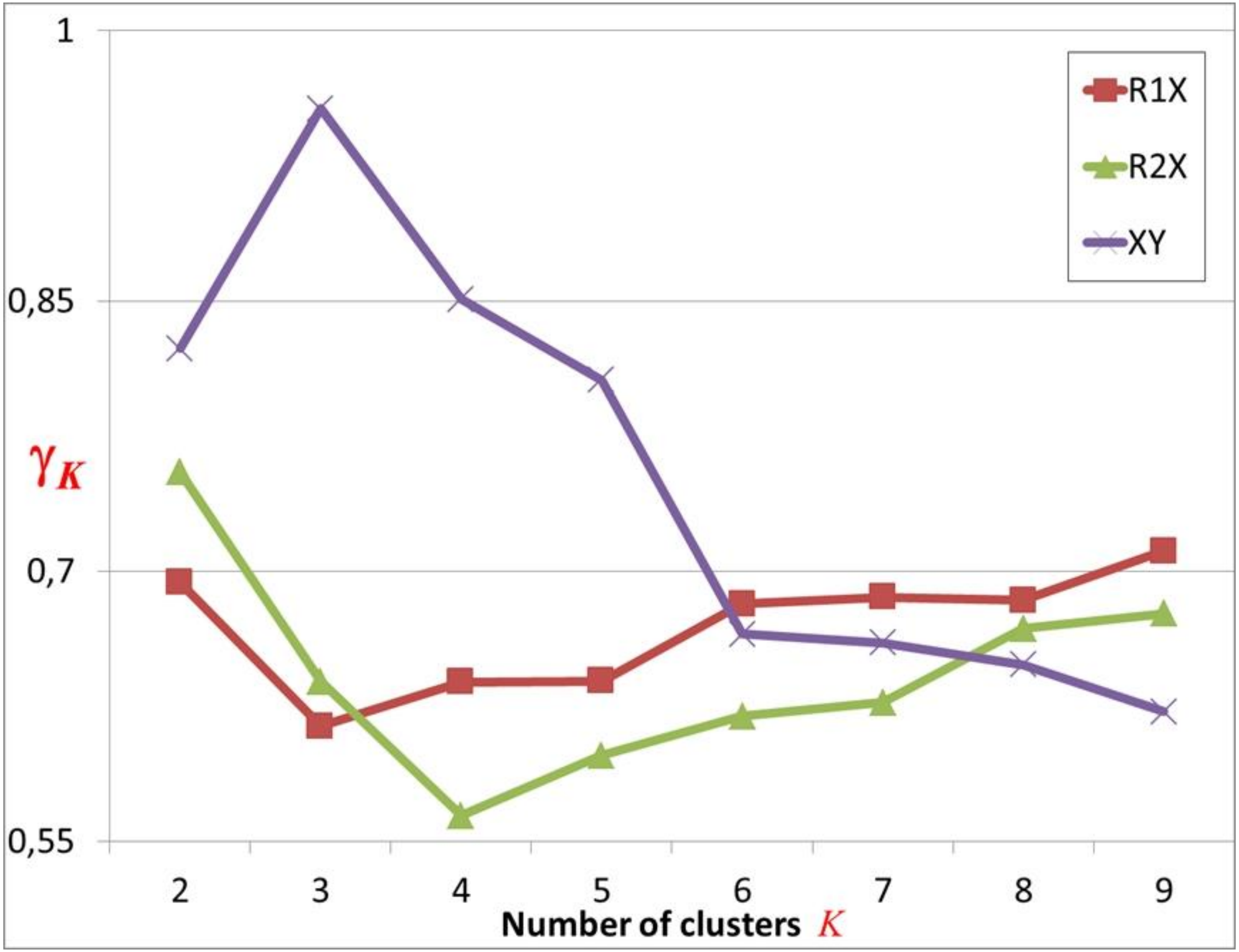
Step 2 of the procedure: Investigation of the stability of the three bivariate Ward's clustering results based on the adjusted Rand index R. The clustering based on variables X and Y is most stable for K=3 .



(XY: 4 errors only !)

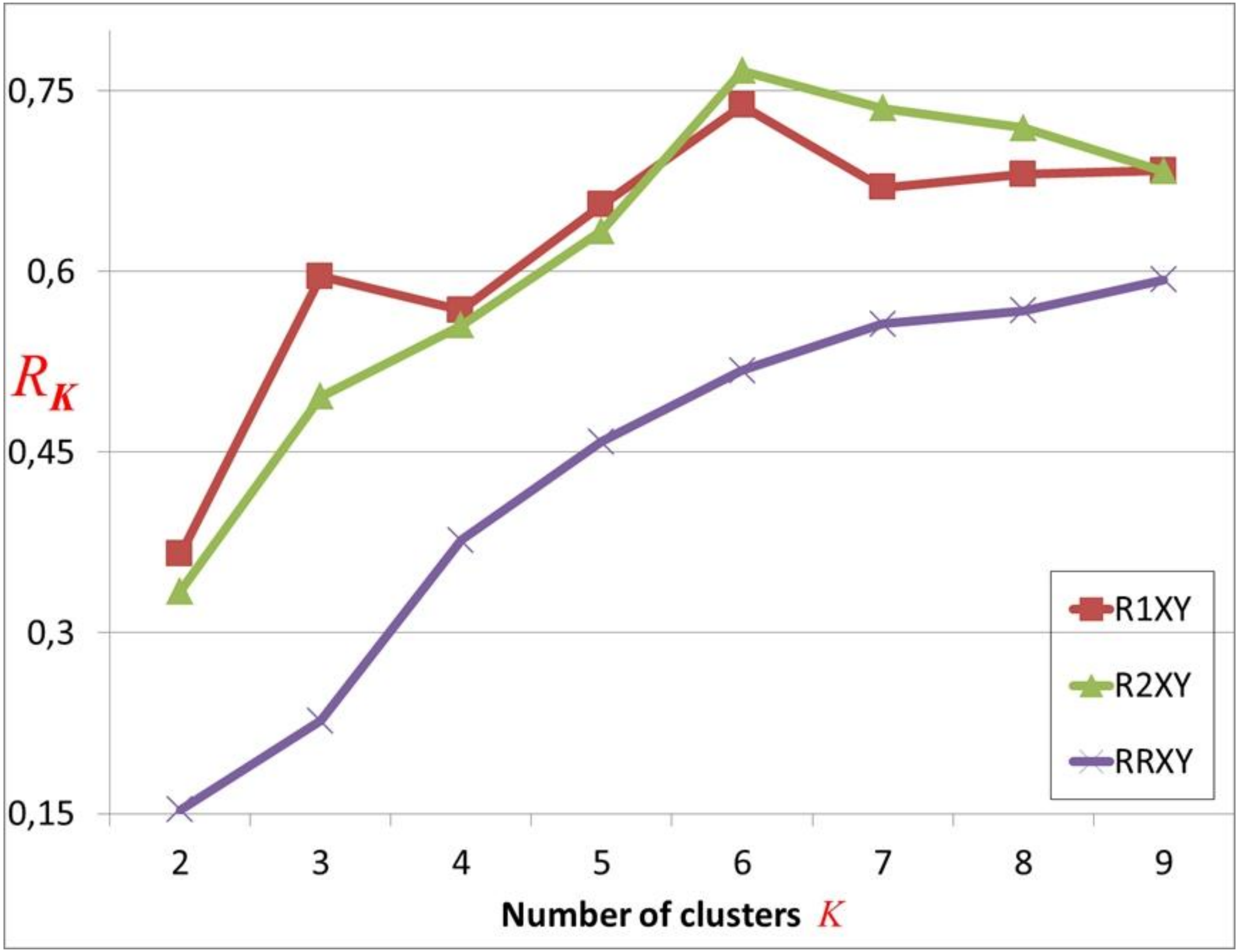
Example: Hierarchical clustering of synthetic data

Step 2 of the procedure: Investigation of the stability of the three bivariate Ward's clustering results based on the averaged Jaccard index" γ_K. Here, for K=2, the situation looks not so clear.



Example: Hierarchical clustering of synthetic data

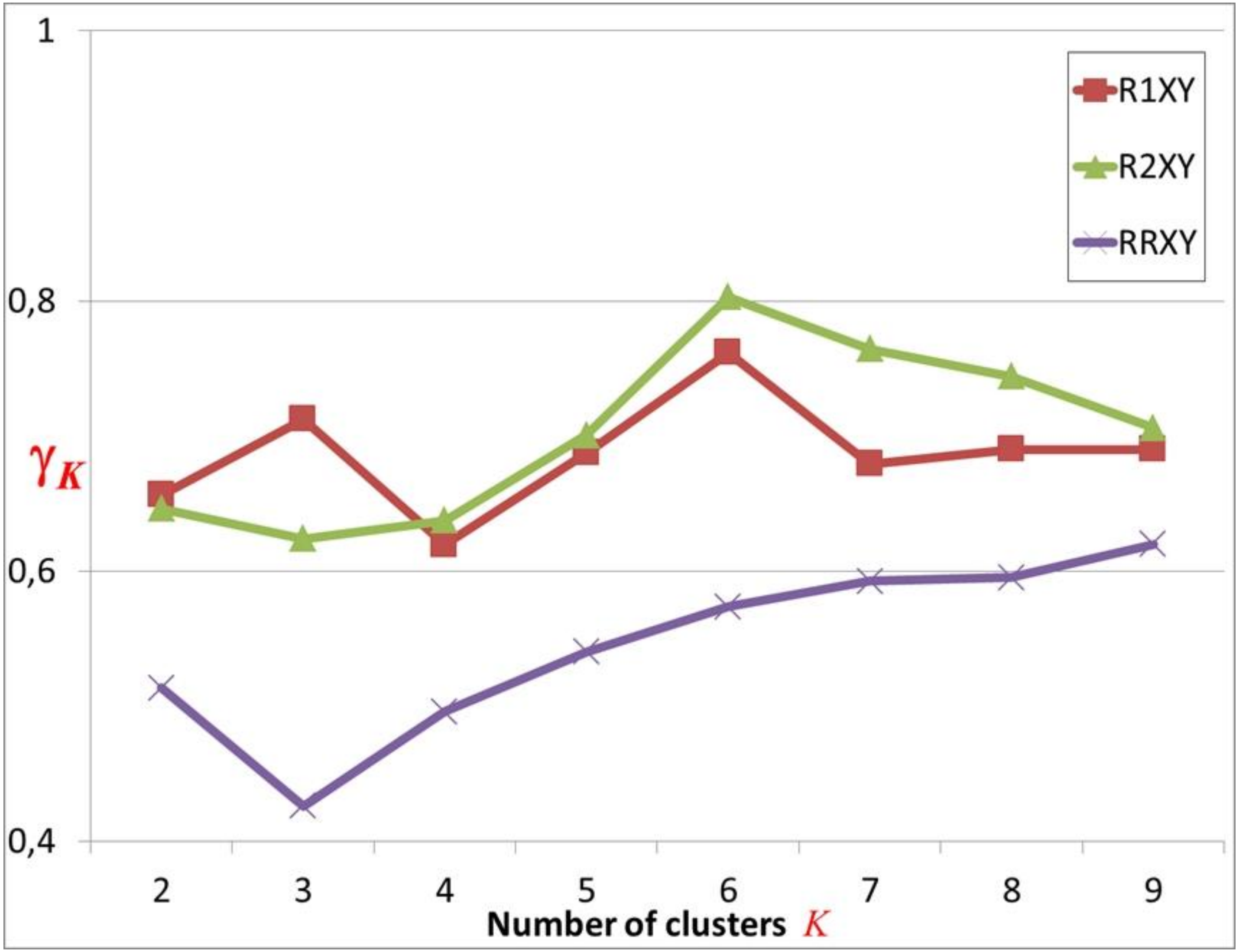
Step 3 of the procedure: Investigation of the stability of the two tri-variate Ward's clustering results based on the adjusted Rand index R. Additionally, the clustering was investigated that is based on all four variables.



(RRXY: 121 errors are counted)

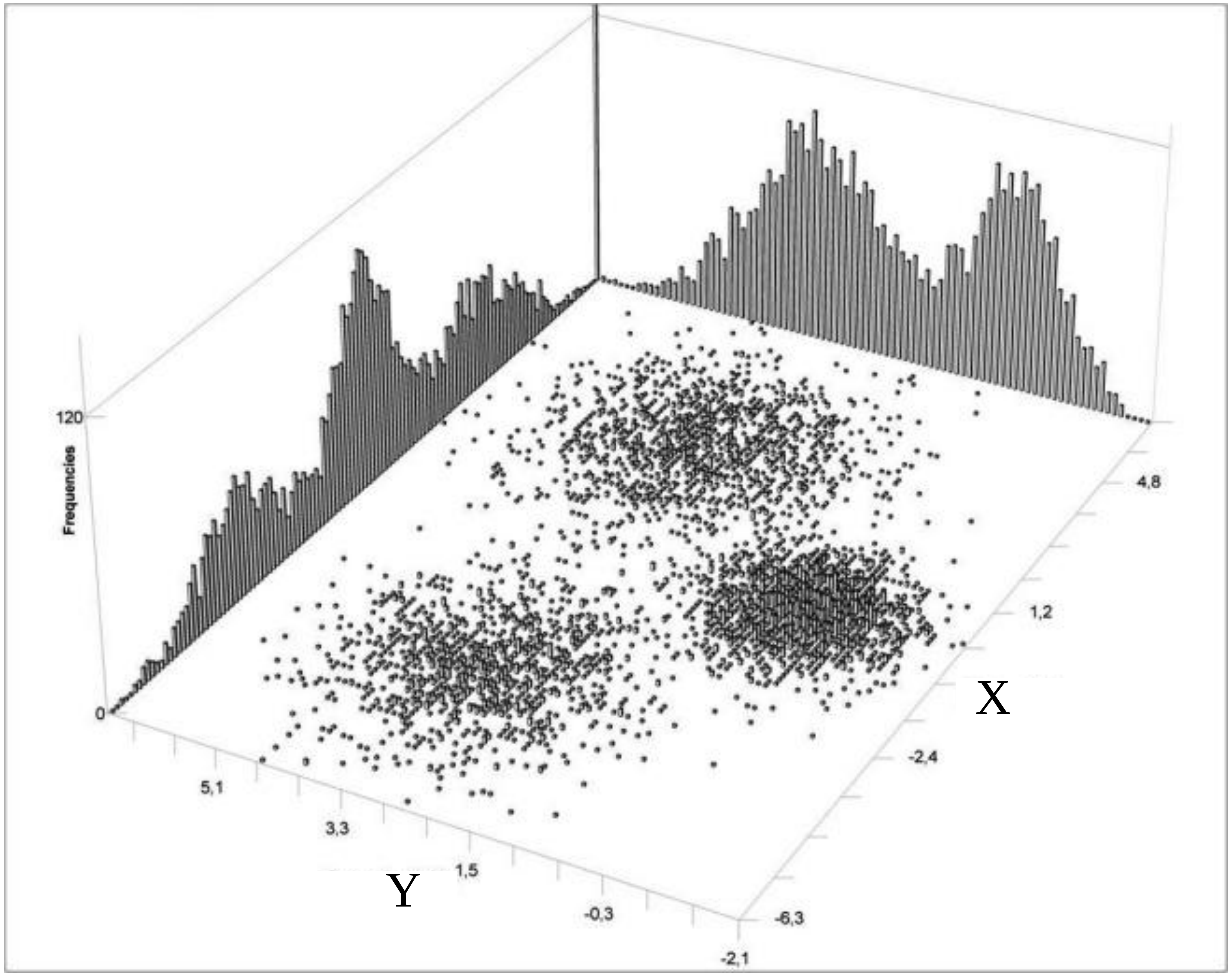
Example: Hierarchical clustering of synthetic data

Step 3 of the procedure: Investigation of the stability of the two tri-variate Ward's clustering results based on the averaged Jaccard index" γ_K. Additionally, the clustering was investigated that is based on all four variables.



Example: Hierarchical clustering of synthetic data

The procedure stops at step 3 because the stability of trivariate clustering decreases rapidly. ***Ward's method based on the selected variables X and Y is successful in finding the 3 classes: 4 errors are counted only.*** One gets quite similar partitional clustering results in the case of 4000 points (1100, 1600, 1300, see the histograms of X and/or Y).



Example: Clustering of the Swiss bank notes data

The data set contains six measurements made on 100 genuine and 100 counterfeit old-Swiss 1000-franc bank notes with the following variables:

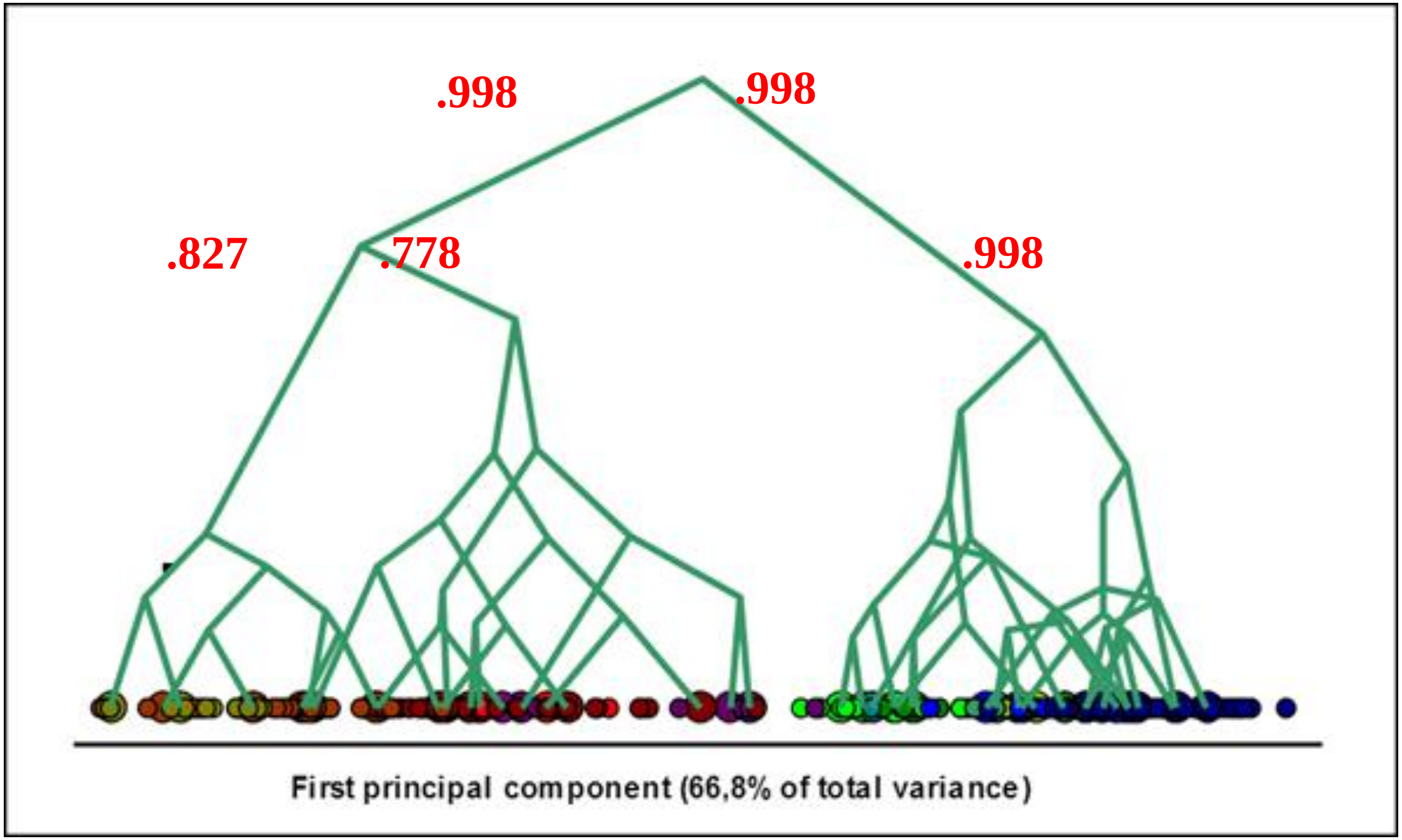
- Length: Length of bill (mm)
- Left: Width of left edge (mm)
- Right: Width of right edge (mm)
- Bottom: Bottom margin width (mm)
- Top: Top margin width (mm)
- Diagonal: Length of diagonal (mm)

Source: Flury, B. and Riedwyl, H. (1988). Multivariate Statistics

Example: Clustering of the Swiss bank notes data

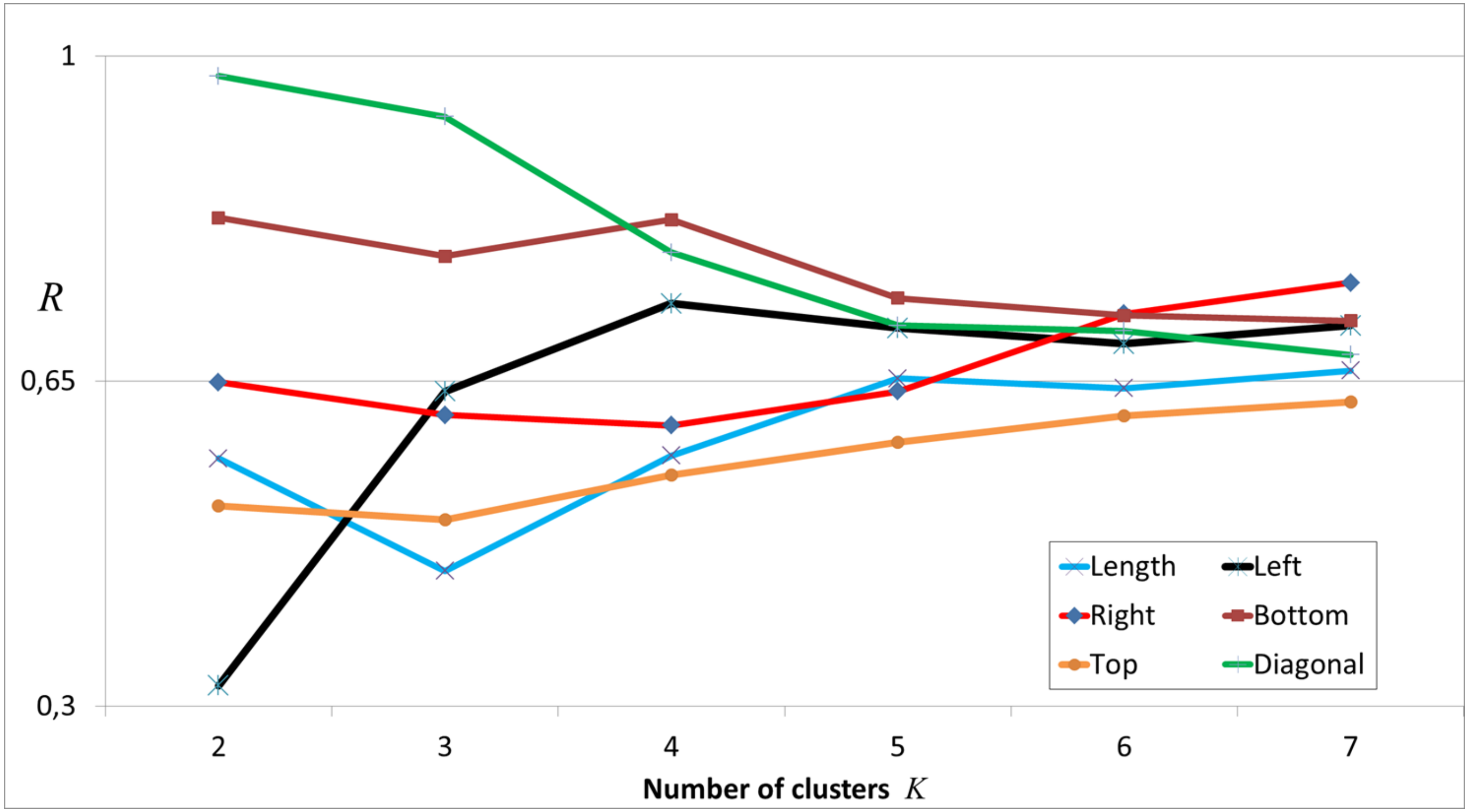
Dendrogram of the hierarchical ***Ward***'s clustering of 200 Swiss bank notes based on all 6 variables (one error only!). The genuine bank notes on the right hand side look more homogeneous than the forged ones.

The γ_k value of an original cluster come from a comparison with 250 Ward's clustering results of bootstrap samples.



Example: Clustering of the Swiss bank notes data

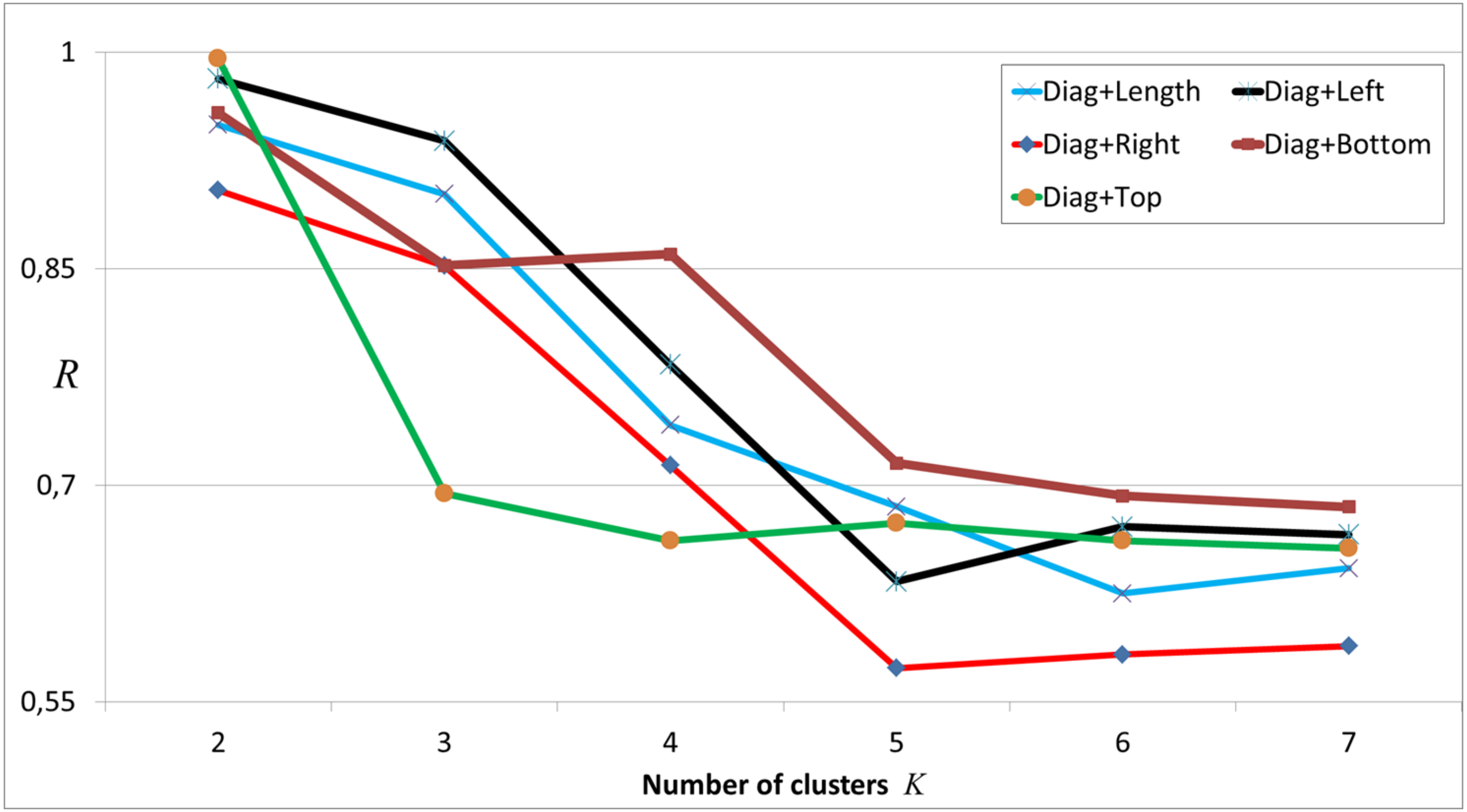
Investigation of the stability of the six univariate Ward's clustering results based on the adjusted Rand index R coming from comparisons with Ward's cluster analyses of 250 bootstrap samples.



(Diagonal: 2 errors are counted only)

Example: Clustering of the Swiss bank notes data

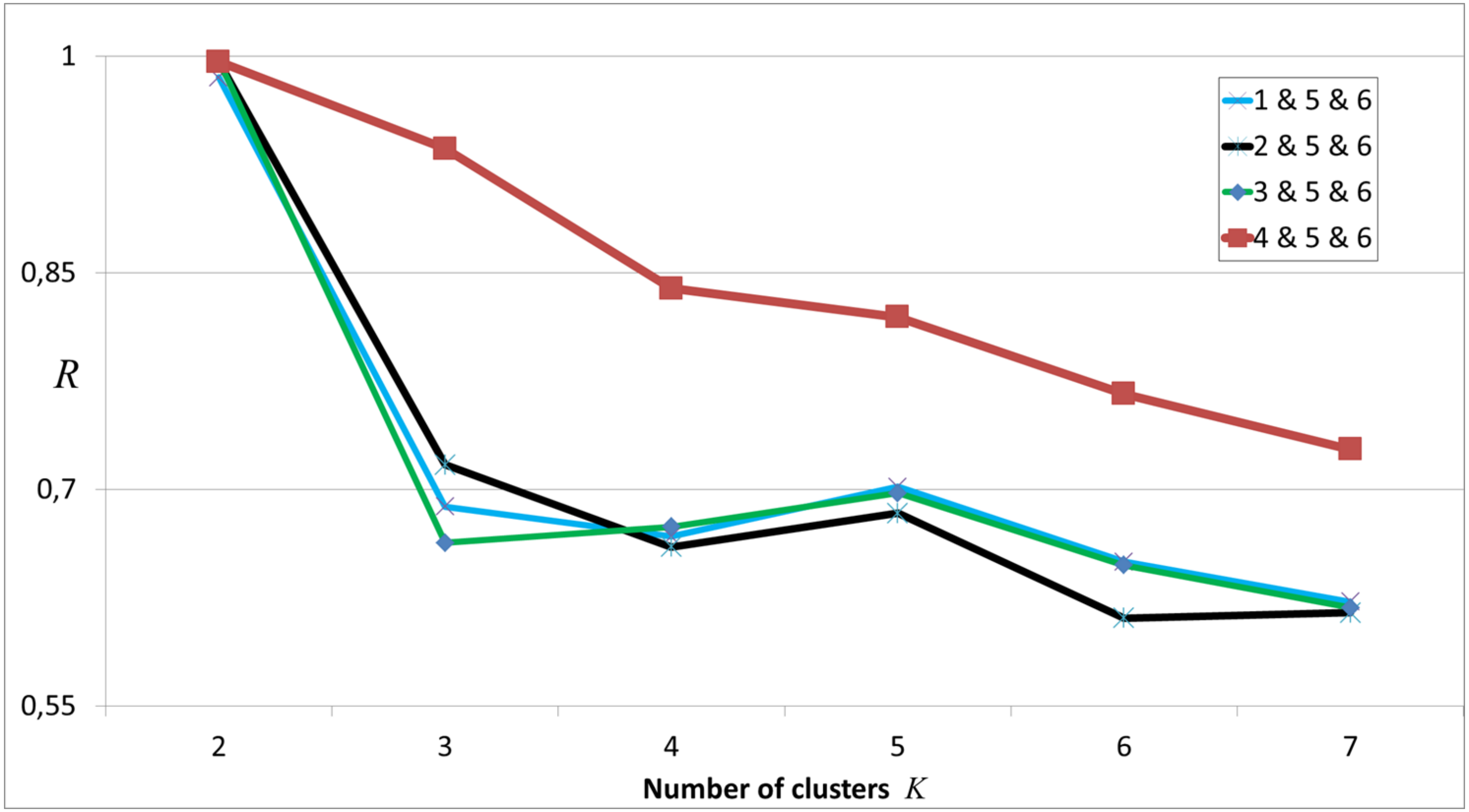
Step 2: Investigation of the stability of the five bivariate Ward's clustering results based on bootstrapping the adjusted Rand index *R*. See the high increase of *R* of "Diag+Top" when going from *K*=3 to *K*=2.



(Diagonal + Top: 1 error is counted.)

Example: Clustering of the Swiss bank notes data

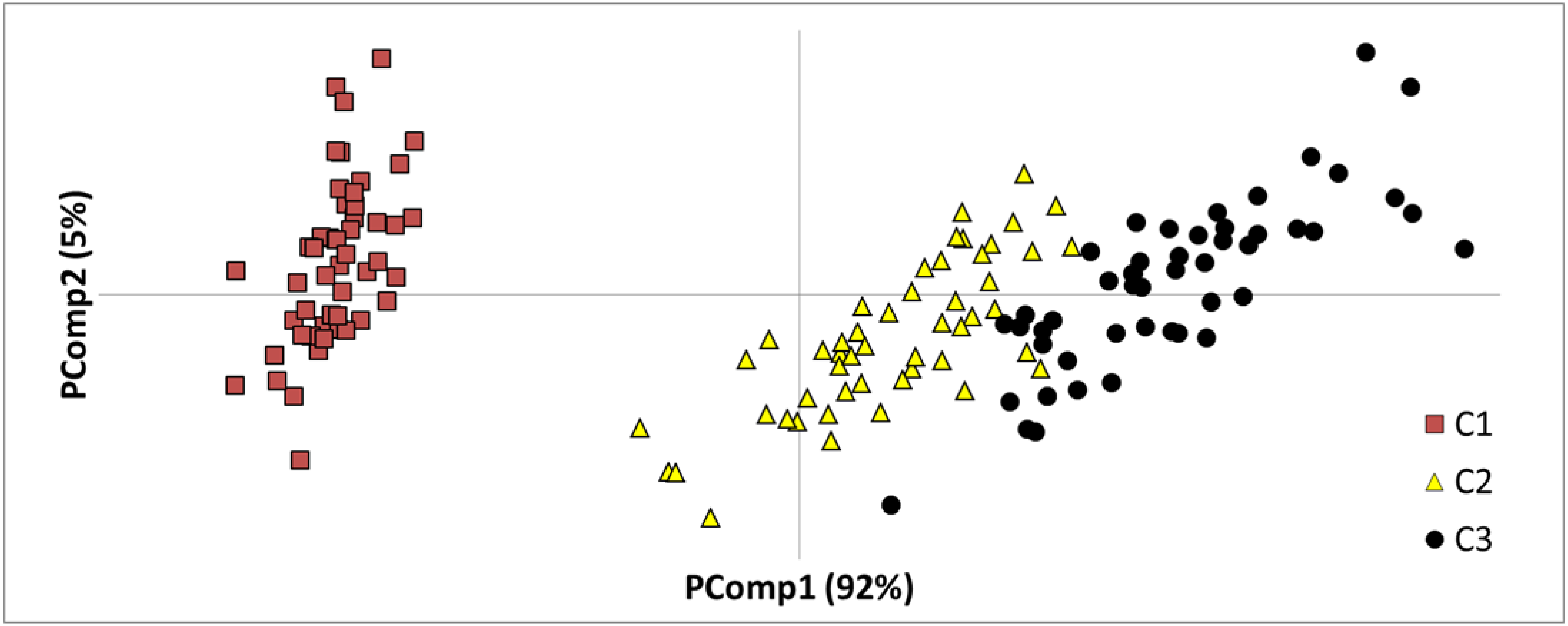
Step 3: Investigation of the stability of the four tri-variate Ward's clustering results based on bootstrapping the adjusted Rand index R. Is the fourth variable "Right" important in dividing the forged bank notes?



(Bottom + Top + Diagonal (4 & 5 & 6): 1 error is counted.)

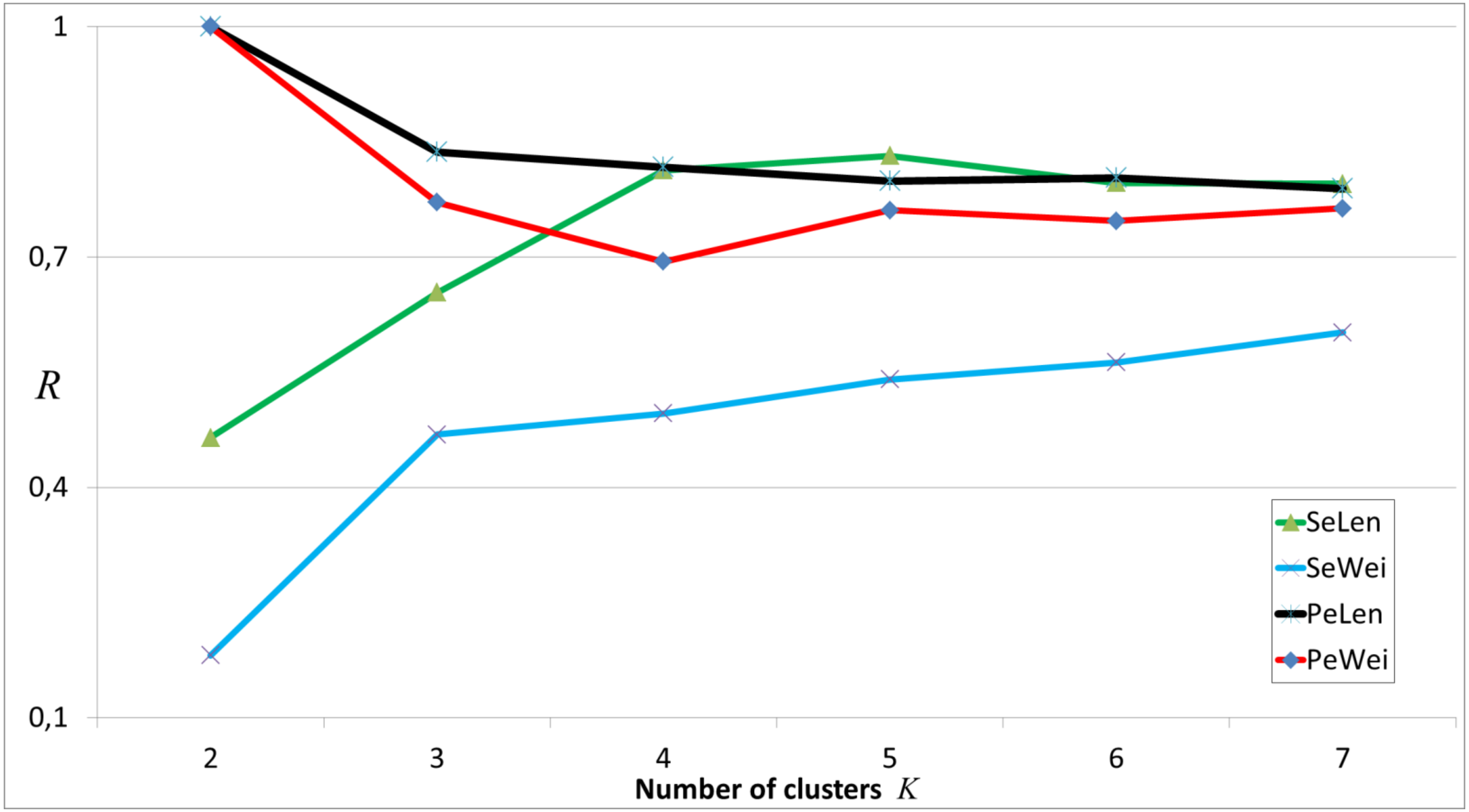
Example: Hierarchical clustering of the Iris data

Iris flower data (Fisher, 1936). The PCA-plot shows the 150 observations. The class (species) on the left hand side is easy to find. The other two species are not separated of each other. 16 errors are counted when using Ward’s method.



Example: Hierarchical clustering of the Iris data

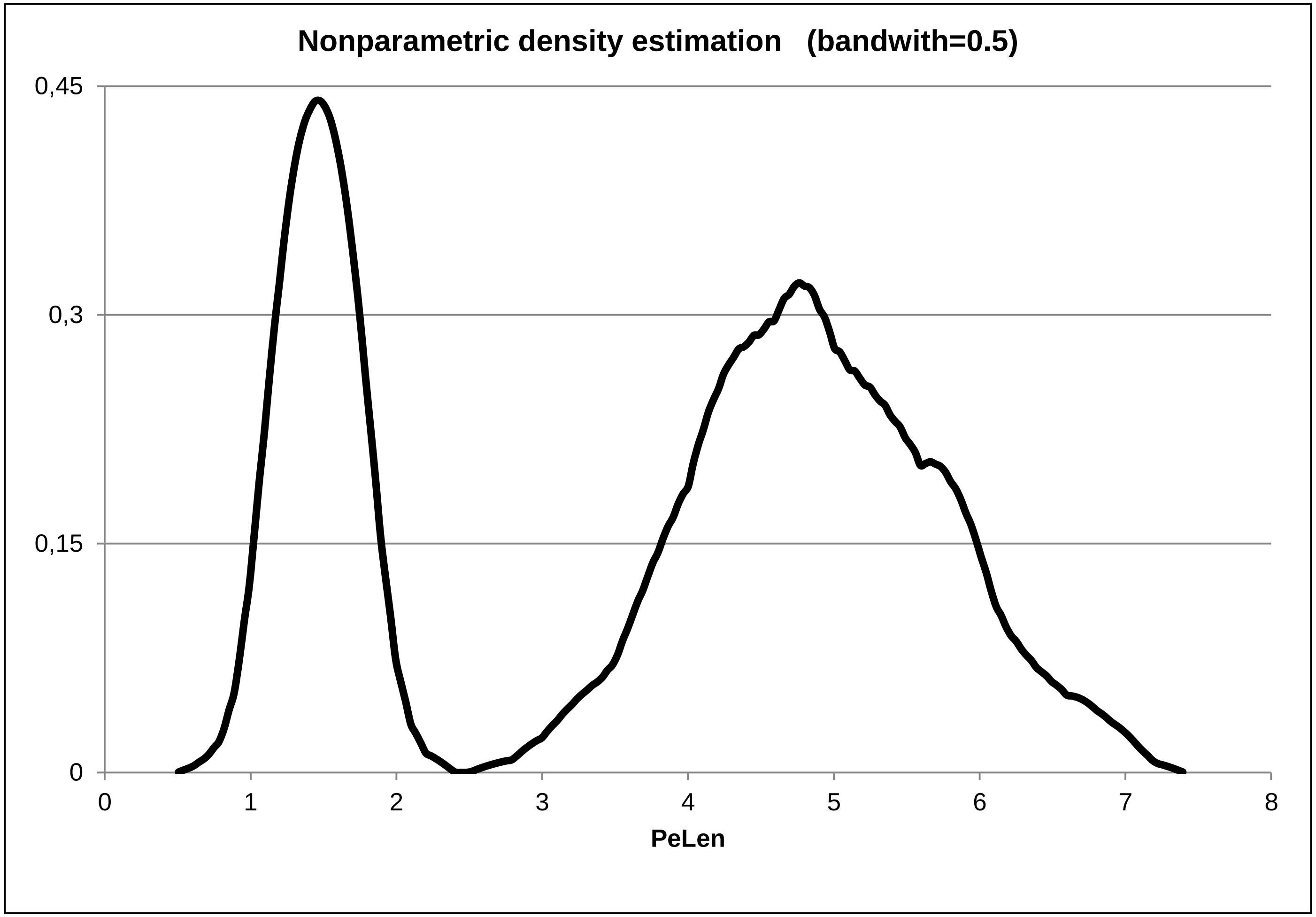
Investigation of the stability of the four univariate Ward's clustering results based on the adjusted Rand index R coming from comparisons with Ward's cluster analyses of 250 bootstrap samples.



(K=2: PeLen, PeWei: no errors, K=3: PeLen: 16, PeWei: 22)

Example: Hierarchical clustering of the Iris data

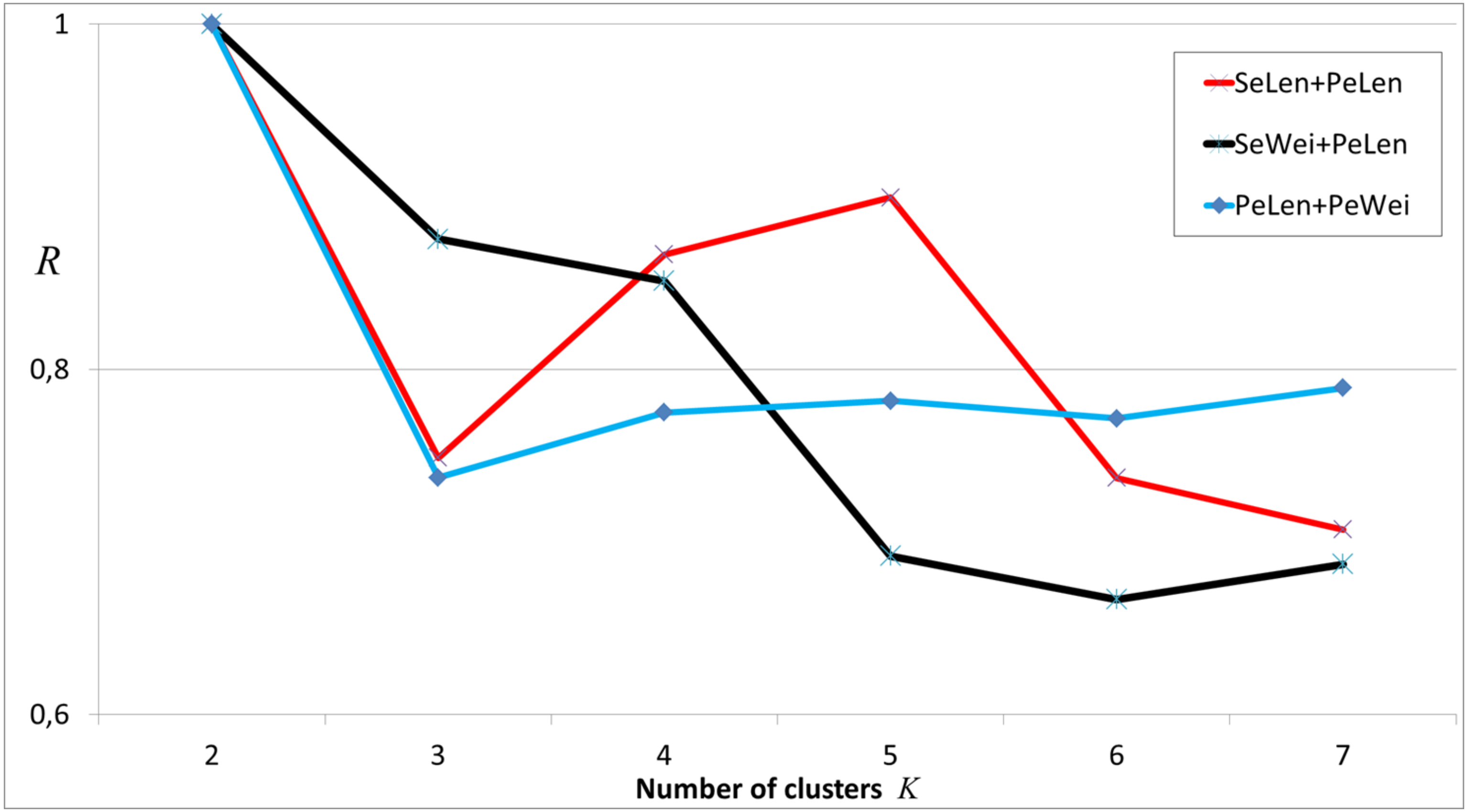
Indeed, the variable petal length can be used alone to find a stable two class solution.



Example: Hierarchical clustering of the Iris data

Step 2: Investigation of the stability of the tri-variate Ward's clustering results. All bivariate clustering results are most stable for $K = 2$.

Conclusion: Ward's method fails in finding the three true classes.



(K=2: no errors, K=3: PeLen + PeWei: 22)

Summary

- The general approach to variable selection proposed here works without using special clustering criteria such as within-cluster or between cluster variances.
- It is based only on non-parametric resampling techniques and criteria of stability such as the adjusted Rand's measure or the averaged Jaccard measure.
- This basic variable selection procedure can be modified in several ways. For instance, one can start with $J*(J-1)/2$ bivariate clustering results.

Thank you very much for your kind attention!