Some Thoughts on Simulation Studies to Compare Clustering Methods

Christian Hennig

Abstract Simulation studies are often used to compare different clustering methods, be it with the aim of promoting a new method, or for investigating the quality of existing methods from a neutral point of view. I will go through a number of aspects of designing and running such studies, including the definition and measurement of clustering quality, the choice of models to generate data from, aggregation and visualisation of results, and also limits of what we can learn from such studies. The paper may be useful for researchers who run such simulation studies and for those interested in the results of them. Some aspects are relevant for more general simulation studies, also outside the domain of cluster analysis.

Christian Hennig
Department of Statistical Sciences “Paolo Fortunati”, University of Bologna
via delle Belle Arti, 41, 40126, Bologna, Italy

University of Statistical Science, University College London
🖂 c.hennig@ucl.ac.uk

ARCHIVES OF DATA SCIENCE, SERIES A
(ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/24
ISSN 2363-9881

1 Introduction

Simulation studies are often used to compare different clustering methods. This paper collects some hopefully useful thoughts on key issues regarding running and interpreting such simulation studies, inspired by my wide experience as designer and reviewer of such studies. Some of these thoughts will be specific to cluster analysis, but some apply to more general simulation studies for comparing statistical methods.

The term "clustering method" here is meant very generally. Usually one would probably think of comparing a new clustering method with k-means, spectral clustering and other competitors, but what follows is also relevant for comparing different versions of the same approach, such as comparing different algorithm initialisations for k-means, or different choices of tuning parameters, or for different methods for estimating the number of clusters used together with one or more approaches for clustering with a given number of clusters.

There are various ways to compare clustering methods:

- **Mathematical theory** can take various forms such as asymptotic theory (consistency, asymptotic normality) assuming probability models, for which clustering methods are usually interpreted as estimators of model parameters (such as the k means of clusters following a certain cluster model), or axiomatic theory that investigates certain desirable characteristics for clustering methods, c.f. Ackerman et al. (2010).

- **Comparison on real datasets with known grouping** by assessment of the amount of misclassification (comparison on real datasets without known grouping could also be done, but it is less clear then how to assess the quality of a clustering solution).

- **Simulation study with artificial datasets** generated with known "ground truth".

All these have advantages and disadvantages. Mathematical theory comes with a specified domain of validity; mathematical statements are general under the conditions under which they are proven, and the proofs are not affected by statistical uncertainty.

However, mathematical theory of cluster analysis is often difficult, and existing results are often either rather unspecific (e.g., just dividing methods into those that have a certain property and those that do not), come with restrictive assumptions, or give asymptotic characteristics that can be quite different from the finite sample behaviour relevant in practice. Furthermore, my experience is that cluster analysis is often rather sensitive to model assumptions, i.e., the behaviour of methods may be quite different if model assumptions are only slightly violated. An example for this is the issue of estimating the number of mixture components (interpreted as clusters) in a Gaussian mixture model; clusters that are approximately but not exactly Gaussian will be fitted by more than one Gaussian component if there are enough observations that the deviation from Gaussianity can be picked up.

The comparison based on real datasets is attractive because real datasets may represent better what is relevant in reality than simplifying models. A major problem is that it is not clear how results from a handful of real datasets generalise to any situation other than just these datasets. Normally the used real datasets are not a representative random sample of a well defined population of datasets; if this were the case, results could be generalised to that population, but in most situations such a sample is not available. Furthermore, even if a real dataset comes with a "true" grouping, it cannot be ruled out that there are other meaningful groupings in the same dataset, and therefore it cannot necessarily be held against a clustering method if it finds a very different clustering; ultimately science is about finding out something new, and reproducing what is already known is not always better.

Simulation studies can explore situations for which theory is not available, and compared to real datasets they can be used to explore more systematically a space of potential situations of interest, such as different numbers of clusters, different degrees of separation between clusters etc. Furthermore they can give an idea about the variation of results for datasets generated from the same model. They can also highlight ideal and problematic scenarios for specific methods. By "scenario" I mean a complete specification of a simulation setup, i.e., probability model and all its parameters, numbers of observations etc., but allowing for replication with new realisations of the involved random variables. In principle, artificial datasets can also be deterministic, but it is then inappropriate to speak of a "simulation study".

However, it is also important to have limitations in mind. Artificial datasets may deviate systematically from what happens in reality, and the issues of generalisation and uniqueness of the “true” clusters is not fully solved by them, see Key Issues 2, 3.

Theory, real datasets, and simulation studies with artificially generated data all have their place in the assessment of quality of clustering methods, because all of them deliver something that cannot be replaced by the other two approaches. Part of the considerations when setting up a simulation study is to what extent theory is available to compare the methods, and it should be done consciously in order to give information that can neither come from theory nor from analysing real datasets.

Another key issue to understand when setting up a simulation study, and maybe even more importantly when appreciating the results of such studies, is the following.

Key Issue 1. ***“Method-centred” vs. neutral simulation studies.*** *Probably the majority of simulation studies in clustering are found in papers in which a new method is introduced, and the study then is used to demonstrate the benefits and (far less often) drawbacks of the new method. I call such studies “method-centred”. Obviously the authors of such studies cannot be neutral when comparing their own method to others; by a “neutral” study I mean a study the authors of which do not have any personal interest in the success of any of the compared methods. This does not mean that any study can be perfectly neutral and impartial, see below.*

It is important to acknowledge that there is a place for method-centred studies; not all of them should or can be neutral. Authors of a new method are for good reasons expected to make a case for their method. They need to demonstrate that the new method has something to offer that existing methods cannot offer already. A simulation study is often appropriate for this. It is illusory to demand that such studies should be fully impartial. The authors have to find scenarios in which their new method performs well, so it is most likely that they arrived at the scenarios that they finally present by looking around, trying out more than they present, and potentially also by adapting their own method so that it can deal better with the simulated scenarios. This often even happens before a systematic simulation study is in fact run, because “developing” the method may imply pre-testing it and improving it in case the pre-test fails; the model used for pre-testing, or a very similar one, may later appear again in the simulation study.

I think that this is legitimate as long as the authors do not try to convey that their study were in fact neutral. It is important that readers of the study do not interpret the study as neutral and as giving a general account of the performance of the new method and its competitors. Certainly it is not appropriate to cite a method-centred study to state in general terms that "as shown in the study S, the new method M is superior to k-means, spectral clustering and other clustering methods involved in the study", as can sometimes be seen in the literature!

Method-centred studies should still be fair, see Key Issue 4. Another desirable feature of a method-centred study is that it should also highlight limitations, i.e., situations in which the new method could be expected to perform well but does not. Authors may be reluctant to show weaknesses of their own method, but this contributes to the better understanding of the new method and can guide a potential user regarding when to use and when better not to use the new method. Obviously, reviewers should not use such aspects of a simulation study to reject the paper!

There is a strong need for neutral studies for comparing clustering methods. The authors of such a study should not only not be the authors of one or more of the involved methods, they should neither have an interest in any of the methods in particular for other reasons, be it because they are connected in some way to the authors of the methods, or be it that they lean toward one of the involved methods for other reasons, for example because they have used this method in their work and/or defended its use in a discussion. Ultimately it is hard or even impossible to avoid such bias as author, and we cannot hope for anything better than an honest "as-neutral-as-possible" attitude; as experts in cluster analysis, study authors may know method authors personally, and may have applied many methods in their own work. A certain unconscious bias can never be ruled out.

An alternative approach is that simulation studies could be put together involving all method authors, and giving all these authors the chance to provide a simulation scenario that they believe is favourable for their own method.

The paper will go through the following major steps in designing and running simulation studies:

- Problem definition (Section 2)
- What scenarios to simulate? (Section 3)

- What methods to include? (Section 4)
- How to evaluate performance? (Section 5)

The thoughts in this paper are a side product of my work in the IFCS (International Federation of Classification Societies) Cluster Benchmarking Task Force. This Task Force has produced a "White Paper" on cluster benchmarking (Van Mechelen et al. (2018)), in which the comparison of clustering methods is discussed systematically. There is some limited overlap with that paper (and some influence of the other Task Force members on this paper, which I acknowledge, as well as helpful comments by Tim Morris and two reviewers). The present paper is meant as a collection of thoughts rather than a systematic guideline for running simulation studies. See Morris et al. (2019) for a general tutorial on comparative simulation studies with some further interesting references.

2 Problem Definition

Normally, in cluster analysis, a clustering is seen as "good" if it matches the "true" clustering closely, assuming that such a true clustering exists (this point was made in the pioneering clustering simulation study of Milligan (1980) and has been taken up ever since). But there is no unique definition of a cluster; several definitions exist that in some models may define different "true" clusters. A simulation study should therefore come with a specific definition.

Key Issue 2. ***Definition of "true clusters"*** *Much literature suggests that given a dataset or mode, it is clear what the "true" clusters are that a clustering method can be expected to find. Usually, a formal definition of the clustering problem is not given, and often the "true clusters" are defined by fiat, appealing to intuition, often using two-dimensional images (which cannot easily be done for higher dimensional data). This is not good practice. There should be a clear definition of the clustering problem that the simulation study is meant to deal with. In Hennig (2015) I have discussed different possible definitions of clusters. Here are some that are explicitly or implicitly used in some literature:*

- *Clusters can be defined as components of a mixture probability model. Without further constraints such models are not identifiable, and therefore* *this approach needs an identifiable specification of the parametric form or shape of the components (e.g., Gaussian).*

- *Clusters can be defined as density level sets or as associated to density modes (in which case there needs to be a specification how observations are assigned to modes).*
- *Clusters can be defined as represented by centroids minimising a certain criterion that formalises how observations are assigned to centroids.*
- *Clusters can be defined as sets between which there is sufficiently strong separation (there is more than one possibility what exactly that means).*
- *Clusters can correspond to distributions formalising certain geometrical shapes such as spherical or linear.*
- *Clusters can be defined as optimal for serving other aims, e.g., clustering may be used for data compression, dimension reduction through clustering of variables, or prediction of external variables, and suitable optimality criteria can assess such uses.*

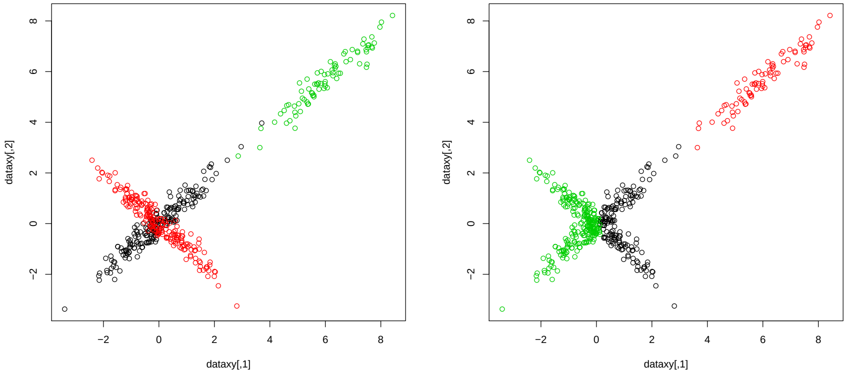

Figure 1: Artificial dataset generated from a three-component Gaussian mixture, illustrating different cluster definitions. Left side: if clusters are defined as Gaussian mixture components, there are three clusters here, as indicated by the three colours. If clusters are associated to density modes, there are two clusters here (the green one, and the red and black components taken together). If clusters are defined as linear patterns, there are two clusters here (the red one, and the green and black components taken together). Right side: optimal representation of the data by three centroids according to 3-means.

These definitions do not necessarily agree, as is illustrated in Figure 1. The assessment of cluster quality (Section 5) and also the decision of what to simulate (Section 3) needs to depend on the kind of clusters that are of interest, and therefore this should be explicitly defined.

A formal definition of true clusters may be seen as too restrictive, for example, the experimenter (study author) may be interested in approximately but not necessarily precisely Gaussian clusters that are also well separated, excluding outliers. Giving such informal descriptions of the kind of clusters of interest is certainly better than not giving a definition at all and appealing to the reader's intuition. The study design and performance measurement can then be chosen accordingly.

Some definitions may seem inappropriate for certain datasets. For example, one may think that the clustering on the right side of Figure 1 looks counter-intuitive and k-means is just inappropriate for these data. However, there are applications such as data compression, for which such a definition may still be useful (Jain (2010)). The experimenter may be interested in the chosen cluster definition in general, or only in certain situations, for example situations in which it is in agreement with their intuition or with other definitions (i.e., centroid-representation only if there also is separation). The simulated scenarios (see Section 3) can be chosen accordingly. Actually, by choosing particular scenarios, the study design always implies a certain restricted problem definition.

Here is an example for considerations connected to the problem definition. In Coretto and Hennig (2016) we were interested in robust clustering in the presence of outliers. The clusters were modelled by Gaussian distributions, but we did not want to restrict considerations to precisely Gaussian distributions; we also wanted to know how methods performed with distributions that generate clusters for which in practice a Gaussian distribution would not be seen as totally inappropriate. In particular, we generated data from mixtures of (multivariate) t-distributions. The issue with t-distributions with a low number of degrees of freedom is that they also generate data that looks outlying, and we were interested as well in the correct detection of outliers. Therefore we needed a definition that allowed us to declare the *centre part* of the data generated by a t-distribution as "true cluster" and the outlying part as "true outlier". We achieved this by defining a robust covariance matrix functional for distributions, and by declaring data "outliers" that are "too far" away from the (robust) cluster mean, where a threshold for "too far" was required that was chosen so that the probability of being a “true outlier” was very low given that the data indeed was defined by a Gaussian distribution. The resulting definition reproduced truly Gaussian clusters almost perfectly, but was applicable also to other distributions, defining certain low probability regions formally as “outlying”, and also defining what is meant by a “true cluster” generated by a non-Gaussian distribution, but with a Gaussian shape defining the kind of clusters that we were interested in.

3 What Scenarios to Simulate?

The performance of clustering methods can depend strongly on many factors, e.g.,

- the number of clusters,
- distributions within clusters, including dependence/covariance structure, geometrical “shape” etc.
- whether and how these differ between clusters,
- the relative position of the clusters (separation etc.),
- the absolute and relative size of the clusters,
- existence of outliers or “noise”,
- the type of distance (for distance-based methods),
- dimensionality of the dataset.

It is hardly possible to cover all these aspects exhaustively in a single simulation study, particularly because there are very many possibilities for aspects such as the distributional shape of clusters and outliers and their relative position. Therefore a simulation study in cluster analysis will always be restricted in certain ways. A systematical study may choose to vary some of these factors, holding others constant, optimally using a factorial experimental design combining a number of levels per factor. Fractional factorial designs allow for a larger number of factors, but sometimes interactions between factors are important and a fractional design loses too much information. The decision what *not* to simulate is important and unavoidable. Ultimately the study needs to differentiate between the compared methods, so aspects are of most interest where different levels may imply different relative performances of the different methods. In my experience, almost all simulation studies in clustering vary the dataset size, and this is very often the least interesting factor.

In some literature the impossibility to run an "exhaustive" experiment is used as an excuse to only run a tiny one, with only, say, one or two different scenarios. This is hardly informative, and authors of method-centred studies give far more credibility to their new method if they make an effort to test it over a wide range of scenarios.In particular, this makes it far more difficult to cherry-pick scenarios in which the new method works well if these are difficult to find.

In method-centred simulations, authors could look for

- a "proof of concept", i.e., a scenario in which the new method is optimal, in order to show that at least there exist situations for which it is more valuable than anything else in the literature,
- scenarios that are not perfect for the new method (for example because model assumptions are slightly violated, outliers added etc.). but where one could in reality reasonably expect that the new method should work,
- scenarios in which one could reasonably expect the new method to work, but it does not (or rather, another method works better); as discussed earlier, such scenarios contribute strongly to the understanding of the new method.

Key Issue 3. ***Generalisation*** *The hope is that a simulation study can give us general information about the performance of the involved methods, not only for the specific datasets that are generated. In order to achieve this at least within the simulated scenarios, replication is required in order to explore the variation of results. To what extent we can generalise outside the simulated scenarios is more difficult to assess.*

Generalisability is made possible by equivariance (or invariance) results, i.e., theoretical results that state that a clustering method behaves in a certain appropriate way if the dataset is transformed in specific ways, e.g., all distances are multiplied by a constant. Equivariance results are often easy to obtain, but unfortunately they only allow generalisation to a rather restricted set of further situations.

Apart from such results generalisability cannot be taken for granted. If there are simulation results for scenarios with 2 and 5 clusters, say, can it be expected that 3 and 4 clusters would produce results in between? This will probably happen more often than not, but exceptions exist. Particularly, the relative position of the clusters often matters a lot, and relative positions of clusters for certain numbers of clusters do not imply relative positions for other numbers that automatically allow generalisation. Generalisation outside the simulated range is even more problematic.

Distributional shapes within clusters also often matter a lot, and good simulation results for Gaussian clusters will not necessarily imply good results for other cluster-wise distributions, not even for elliptical ones. The best that can probably be done is to think hard about heuristic reasons why or why not some not simulated scenarios can be expected to behave in line or differently from the simulated ones. At least simulation studies offer a better framework for such thoughts than experiments based on real datasets.

An approach used by some experimenters to improve generalisability is to define scenarios involving randomly generated parameters of the data generating distributions, i.e., to use newly generated parameter values for every simulation run (see DeSarbo and Cron (1988) for an example of combining random parameters with exploring several deterministic factors). For example, if data is generated from Gaussian mixtures, mean vectors or even covariance matrix entries could be randomly generated, covering different levels of cluster separation, shape, and within-cluster dependence. The hope in doing this is that results can then be generalised over the whole range of the space of simulated parameters. To what extent this is justified depends on the variation in results depending on the generating parameters. Certainly it will require a large number of replicates. Different parameters may produce systematically different results, and aggregating results over a distribution of parameters may hide the fact that and how strongly results depend on the specific parameters. Another disadvantage is that two possible sources of variation, namely different parameters, and random variation given fixed parameters, can no longer be told apart. To some extent these issues can be treated by relating the results to the underlying parameter values using techniques such as regression or visualisation.

Here is a last remark on the scenarios to be simulated. There are a number of existing software packages that generate data to be clustered for simulation studies fulfilling certain specifications. Using these is certainly tempting and has some advantages, for example the possibility to directly compare results with other studies that used the same generator. Also, some choices that are required for the simulation study are not trivial, such as the precise definition of separation, and some of these may be handled intelligently in some of these packages. An important issue with using such packages is reproducibility. Packages may change over time, and occasionally package documentations are deficient.

Simulations should always be documented in such a way that a reader can reproduce them (if necessary using appendixes and online supplements to journals), and a simple reference to a data generation package is not enough, even if parameters are given. The experimenter should always precisely know, and write down in terms of probability models with full specification, how data is generated. If this is not clearly and fully explained in the package documentation, the package should not be used.

4 What Methods to Include?

The definition of what the study is meant to achieve also involves the choice of methods to be compared; often an experimenter aims at comparing the best available methods for the chosen cluster definition and population of datasets, but occasionally the scope is more restricted. For example, the experimenter may be only interested in comparing different algorithms for optimising the same objective function, or two specific methods that have been discussed in the literature for a specific aim. Actually, there are thousands of clustering methods in the literature, and simulation studies can only ever cover a very restricted set of methods.

In any case, the experimenter needs to consider the literature and should include methods that can be expected, for example from past performance, to be good for the range of scenarios under investigation, or that are in widespread use for such data. This also applies to studies with a restricted scope, because the relevance of comparing some algorithms optimising the same objective function, say, is questionable if other clustering methods can produce better clusterings in the simulated scenarios than the objective function of interest.

Key Issue 4. ***Fair comparison*** *Most clustering methods require some kind of tuning, may exist with different implementations, or require initialisation for which there are different options. This can be a major factor causing the* *simulation results. As long as the aim of the study is not the comparison of such different versions of the same clustering approach, it is important that the methods to be compared are run in ways that put them on equal footing. In method-centred studies there is always a danger that authors of new methods work hard on tuning their own method to optimal performance, whereas competing methods are used in a much less sophisticated out-of-the-box fashion.*

Particularly, methods should not be tuned for optimising the performance in the specific simulated scenarios. Firstly, this produces selection bias, and secondly, it cannot represent the methods in the way they are used in practice, because in practice the "true" clustering or true parameter values (as used for measuring performance) are not known, so in practice a default tuning or implementation will be used and performance optimisation is not possible. Methods need to be run in the simulation study in the same way in which the authors would recommend to run them in practice without knowing the truth.

Generally it is desirable that experimenters try to make the same amount of effort for tuning all competing methods, and this should not involve comparing performances between different tunings on the simulated setups. Possible considerations are for example whether the same initialisation can be used for all methods, or whether it is possible to tune all methods to the same speed (in case that existing algorithms allow performance improvement at the expense of computing time, for example, in k-means, by initialising the algorithm randomly many times and choosing the best solution according to the objective function, which of course is observable also in practice).

In method-based studies, sometimes authors of competing methods may be available to suggest specific tunings of "their" method. In any case the used tunings need to be seen as essential part of the methods compared in the simulation study, and results cannot be expected to generalise to different ways of tuning.

5 How to Evaluate Performance?

Key Issue 5. ***Performance measurement on a single dataset.*** *The most popular way to measure the performance of clustering methods in a simulation study is probably the comparison between the "true" clustering and the one produced by a clustering method, where a clustering is understood as a set of subsets (clusters) of the data, based on counts of observations that are in a particular* *cluster in the "true" clustering and in a particular cluster in the one generated by the method. Meila (2016) lists and discusses a number of criteria to measure distances between partitions. The most popular is probably the adjusted Rand index (ARI; Hubert and Arabie (1985)), which is based on classifying and counting pairs of observations depending on which they are in the same or different clusters in the two compared partitions. Note that this approach does neither require a matching of clusters in the different clusterings, nor that the numbers of clusters are the same. The definition of "misclassification rates", as often used in supervised classification, requires that every found cluster is assigned to a true cluster; this is probably done best by finding the permutation of cluster numbers that minimises the misclassification rate; if the found clustering structure is essentially different from the true one, this may be seen as inappropriate.*

The ARI has certain disadvantages (see Meila (2016) and references given there), but in my experience overall simulation results will rarely differ strongly from what other existing indexes deliver. However, in the study in Coretto and Hennig (2016), already discussed in Section 2, we decided to compute misclassification rates based on optimal matching, under the side condition that observations classified as outliers are mapped to "true outliers" (which cannot be done using the ARI). Other performance measures may occasionally be of interest, depending on the problem definition:

- *parameter estimation where clustering is done based on estimating a parametric model (mean squared error etc.),*
- *quality of the approximation of the true density, where the clustering method involves (parametric or nonparametric) density estimation,*
- *a target criterion formalising the cluster concept (e.g. for representation of objects by centroids),*
- *measuring aspects of specific interest of the clustering such as how often the true number of clusters is estimated,*
- *indirect criteria if clustering is done for serving other aims (e.g., quality of prediction of an external variable using the clusters as predictors).*

It is often worthwhile to consider the implications of such measures. For example, very different clusterings may go with similar density estimates (and the quality *of density estimation may therefore not be suitable if the clustering of the observations is the primary interest). One issue with estimating the number of clusters and counting how often the true number of clusters is found is that arguably a good clustering with the wrong number of clusters is better than a bad clustering with the correct number of clusters, see Figure 2. Therefore even in studies comparing methods to estimate the number of clusters, it may be worthwhile to use the ARI together with some measurement based on the estimated numbers of clusters to assess the performance.*

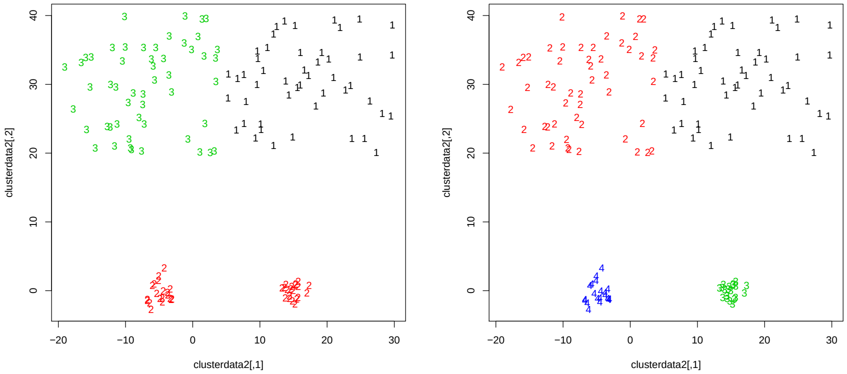

Figure 2: Artificial dataset generated from a mixture with two Gaussian and one uniform component with 3-means (left) and 4-means (right) clustering. The true number of clusters can be reasonably defined to be three (e.g., based on separation), corresponding to the mixture components, and then the 4-means solution looks better than the 3-means solution.

Most simulation studies will be so big that it is not practical to present the results of every method on every single dataset. Furthermore, often the experimenter may want to produce an overall ranking of the compared methods, or (often more appropriately) a differentiated set of recommendations or rankings, recommending for example one method for low dimensional and another one for high dimensional data. This requires the aggregation of results from the potentially many simulated datasets.

Key Issue 6. ***Aggregation (single scenario).*** *Consider aggregating results from a single scenario, and assume that on a single dataset the performance is measured by a single real number. The most popular choice for aggregation is certainly to take the mean. One particular issue with the mean is that it is sensitive to outliers. This can mean that the "mean performance" of a certain method looks bad because out of, say, 100 replicates one was really bad whereas the other 99 were fine. If the performance measure is bounded (such as the ARI between -1 and 1), the effect of outliers on the mean is bounded as well and may sometimes not be a problem. However, if the variation of the non-outlying results is low, outliers may still have a strong impact. One could instead use robust aggregation such as taking the median, but that is not really satisfactory either, because although one would not want one bad result to dominate the overall performance, one would still want to have a measure in which it is reflected that occasionally the method does badly. There are more sensitive robust estimators that are good compromises between means and medians such as M-estimators or* α*-trimmed means with small* α*, meaning that if a very small proportion of results is really bad, it does not influence the result how bad exactly they are; however, the aggregate is still sensitive to changes in the vast majority of results. If one can specify a threshold for "a performance so bad that it is fundamentally wrong and it is no longer of interest how bad it is exactly", one could assign a constant to these results (e.g., the threshold value itself), and take a plain mean afterwards.*

Another issue is with missing values. Many clustering methods are implemented in such a way that occasionally the implementation does not deliver a result (this can happen because of lack of convergence, singularity issues etc.). It would be a mistake in such cases to just aggregate the non-missing results, because a missing result indicates that something undesirable has happened. If missing values are discarded, methods that produce several missing values but do well where they give results have an unfair advantage in comparison to methods that produce results more often; particularly it may be the most "difficult" datasets that attract missing values. A valid approach would be to compute an aggregate without the missing values, but to report the percentages of missing values separately, with the understanding that an overall good performance does not only require a good aggregate result but also a low or zero percentage of missings.

Key Issue 7. ***Aggregation of several scenarios.*** *When aggregating results from several scenarios, an additional problem is that it may not be straightforward to compare the results in the different scenarios. For example, certain scenarios may be more difficult or less stable than others, which may lead to substantially different value ranges and variances, and aggregating the different results in a straightforward manner may implicitly mean that the results of some scenarios dominate the overall aggregate and others have hardly any impact. One could standardise results for each scenario separately before aggregation, but this may be undesirable for the opposite reason: It may be that if the variation of results in a certain scenario is low, one can say that more or less all the methods are of the same quality there, and one would not want to weight this scenario up by standardising to unit variance, say. This depends strongly on the specific performance measure in use, and whether the meaning of absolute differences is the same or different in different regions of its range. Furthermore, the variance in a given scenario consists of both the variance between the different methods and the variance within the same method between replicates. If the within-method variance is high, results in the scenario are unstable and should potentially have low weight, whereas the between-methods variance points at strong differences between the methods that can be of primary interest, in which case it is not desirable to weight them down.*

For example, a 1% difference in misclassification rates has the same meaning regardless of the actual values of the misclassification rates, so one would not want to standardise these to unit variance, because a low variance means in the first place that there is not much relevant difference between the different methods. However, there is little "space" for low misclassification rates to vary and within-method variances are often low but differences between methods are very meaningful, and very large misclassification rates may just indicate that a method gets the clustering structure totally wrong, and then a few percents difference between two large rates may not be of interest.

In such a situation, instead of standardising to unit variance, it may be sensible to standardise by dividing by the mean misclassification rate for the given setting, or for a set of "similar and comparable" settings (it should be avoided to divide by something that is very close to zero because very small differences should not dominate the aggregation too much even where it seems justified to weight them up a bit).

Key Issue 8. ***Visualisation.*** *Visualisation allows to show more differentiated information than computing a single aggregate, and is therefore often preferable. The considerations regarding aggregation apply to some extent as well to visualisation. Sometimes it pays off to use nonstandard displays.*

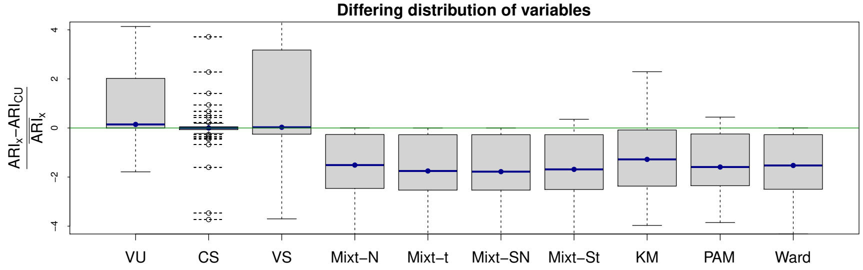

Figure 3: Relative performance in ARI of 10 clustering methods (along the x-axis) compared with method "CU" computed as ARI of the method minus ARI of CU on the same dataset, divided by the average ARI over a set of comparable scenarios, for details see Hennig et al. (2019)..

Figure 3 shows an example for the visualisation of results for lots of scenarios at the same time (it would have taken very many pages to show such graphs for every single scenario), taken from Hennig et al. (2019). A problem here was that there were many missing values (imputed by ARI= 0*, as bad as a totally random clustering) that had substantial influence on scenario means and medians. Therefore we decided to make the different scenarios comparable by comparing all methods to a reference method (called "CU" in Hennig et al. (2019) that never yielded missing values; this was actually a method-centred simulation study) and dividing by the mean ARI over a set of comparable scenarios to make sure that this mean was sufficiently far away from zero; the division was chosen because without it the different scenarios would still have been visually so different that differences between methods would have been dominated by differences between scenarios.*

This is not something that I generally recommend but it was chosen based on the specific results of that study, but not in order to make a particular method look better or worse, but rather in order to produce an image (for which we used standard boxplots), of which the most striking features corresponded to the most relevant results of the comparison.

Key Issue 9. ***Variation and testing.*** *A major benefit of a simulation study based on artificial data compared to using a number of real datasets is that the simulation study can be used to assess the variability of results on data generated from the same scenario. Particularly, a method A can look better than another method B in a study based on some aggregate result, but it is of interest to assess whether this can be explained by random variation alone, in which case it would not be clear that method A is really better even in that specific scenario. The practice to show scenario-wise standard errors of simulation results is rather widespread and certainly better than not giving any indication of variation. But if the different methods are run on the same datasets in the simulation study (which is advisable because it reduces the variation of the comparison between methods) the methods' results are dependent, and therefore the standard errors do not allow to compute a straightforward test of the difference between two methods. This requires paired tests, which one could apply to all pairwise comparisons between methods of interest (potentially with appropriate correction for multiple testing), or random effects model based tests in case that more than two methods are compared at the same time. Tests may be run on nonstandard aggregates (for the reasons discussed in Key Issue 6) using bootstrap or other resampling approaches.*

6 Conclusion

I have presented a number of thoughts to take into account when designing and evaluating simulation studies comparing different methods, particularly in cluster analysis. Major issues are the distinction between method-centered and neutral studies and its implications; the necessity to give a proper problem definition of what kind of "true" clusters are of interest; threats to generalisability; fairness of comparisons, particularly regarding method tuning; issues with evaluation and aggregation of results, in particular the treatment of "performance outliers" and missing results, and making results from different scenarios comparable; the importance of results visualisation; and how to assess variation and whether differences between methods are meaningful.

I end with another recommendation from painful own experience. Sometimes simulations come up with very surprising results (and it would be good if the experimenter had in advance at least a rough idea of what kinds of results to expect). It is very important in such cases to find the reasons for these.

If a certain method performs very differently from what the experimenter thinks of as realistic, it means that the experimenter needs to improve his/her understanding of the methods. But if a convincing data analytic explanation cannot be given, a coding error is a far more likely explanation than the belief that something astonishing has been discovered. Coding errors happen all the time. Sometimes the method-centred study author's favourite method "wins" a study not because it is best, but rather because the author is very critical about good looking results of competing methods and much less critical of good looking results of their own.

References

Ackerman M, Ben-David S, Loker D (2010) Towards Property-Based Classification of Clustering Paradigms. In: Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A (eds.), Advances in Neural Information Processing Systems (NIPS), Vol. 23, pp. 10–18. URL: https://papers.nips.cc/book/advances-in-neural-information-processing-systems-23-2010.

Coretto P, Hennig C (2016) Robust Improper Maximum Likelihood: Tuning, Computation, and a Comparison With Other Methods for Robust Gaussian Clustering. Journal of the American Statistical Association 111:1648–1659. ISSN: 0162-1459, DOI: 10.1080/01621459.2015.1100996.

DeSarbo WS, Cron WL (1988) A Maximum Likelihood Methodology for Clusterwise Linear Regression. Journal of Classification 5(2):249–282, Springer. DOI: 10.1007/BF01897167.

Hennig C (2015) What are the true clusters? Pattern Recognition Letters, pp. 53–62. DOI: 1502.02555.

Hennig C, Viroli C, Anderlucci L (2019) Quantile-based clustering. Electronic Journal of Statistics 13(2):4849–4883. DOI: 10.1214/19-EJS1640.

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

Jain AK (2010) Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters 31(8):651–666. DOI: 10.1016/j.patrec.2009.09.011.

Meila M (2016) Criteria for Comparing Clusterings. In: Hennig C, Meila M, Murtagh F, Rocci R (eds.), Handbook of Cluster Analysis. Chapman & Hall/CRC, Boca Raton FL, chap. 27, pp. 619–636. DOI: 10.1145/1102351.1102424.

Milligan GW (1980) An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. Psychometrika 45(3):325–342. DOI: 10.1007/BF02293907.

Morris TP, White IR, Crowther MJ (2019) Using Simulation Studies to Evaluate Statistical Methods. Statistics in Medicine (published online). DOI: 10.1002/sim.8086.

Van Mechelen I, Boulesteix AL, Dangl R, Dean N, Guyon I, Hennig C, Leisch F, Steinley D (2018) Benchmarking in Cluster Analysis: A White Paper. URL: `https://arxiv.org/abs/1809.10496`. Manuscript submitted for publication.