Validating Synthetic Health Datasets for Longitudinal Clustering

SHIMA GHASSEM POUR[1], ANTHONY MAEDER[1] and LOUISA JORM[2]

[1] School of Computing, Engineering and Mathematics
University of Western Sydney
Campbelltown, Australia
Email: A.maeder@uws.edu.au

[2]School of Medicine
University of Western Sydney
Campbelltown, Australia
Email: L.jorm@uws.edu.au

Abstract

Clustering methods partition datasets into subgroups with some homogeneous properties, with information about the number and particular characteristics of each subgroup unknown a priori. The problem of predicting the number of clusters and quality of each cluster might be overcome by using cluster validation methods. This paper presents such an approach incorporating quantitative methods for comparison between original and synthetic versions of longitudinal health datasets. The use of the methods is demonstrated by using two different clustering algorithms, K-means and Latent Class Analysis, to perform clustering on synthetic data derived from the 45 and Up Study baseline data, from NSW in Australia.

Keywords : Cluster analysis; longitudinal synthetic data; Cluster validation

1 Introduction

Unsupervised learning methods (such as clustering) are based on discovering statistically reliable, unknown previously, and actionable insights from datasets, with information about structure of the datasets (such as cluster number and size) unknown a priori. On the other hand, some clustering algorithms seek to determine the number of clusters in advance. In data that are not clearly separated into groups, identifying the number of clusters becomes difficult. Various validity indexes are available to measure the quality of each cluster, such as Silhouette index (Rousseeuw 1987), Dunn's index (Dunn 1973) and Davies-Bouldin index (Davies & Bouldin 1979). In addition, the BIC index (Schwarz 1978) has been used because it is closely associated with the Latent Class Analysis method.

The organization of the paper is as follows: in Section2 we describe two different clustering methods; in Section 3 we introduce relevant validation methods and Section 4 presents construction of a synthetic dataset. Comparison of the validation methods is presented in section5 and a conclusion in section 6.

2 Two different clustering methods

Cluster analysis refers to partitioning the data into meaningful subgroups, when the information about their composition and the number of subgroups are unknown (Jain et al. 1999). In this paper we use two different kinds of clustering algorithms to cluster our datasets, in order to also investigate the effect of the algorithm choice.

The K-means algorithm is a point-based clustering method which places cluster centers in an arbitrary position and relocates them at each step to optimize the clustering error. Despite being widely used in many clustering applications, this method suffers from sensitivity to initial position of the cluster centers (Likas et al. 2003).

Latent Class Analysis (LCA) is a statistical clustering approach that attempts data reduction by classifying objects into one of K homogeneous clusters, where within-group-objects similarity is minimized and the between-group-objects dissimilarity is maximized, and where K is fixed and known. LCA applies a probabilistic clustering approach: this means that although each object is assigned to belong to one cluster, it is taken into account that there is uncertainty about an object's class membership (Magidson & Vermunt 2002, Lanza et al. 2003).

3 Validation methods

There are several validation methods available to validate the quality of clusters resulting from a given clustering method. One approach consists of running a clustering algorithm several times for different numbers of clusters and computing validity indexes to assess the quality of each cluster. Validation indexes can be divided into two categories: external index and internal indexes. External index techniques use a dataset with known cluster configurations and measure how well clustering methods perform with respect to these known clusters. Internal indexes techniques are used to evaluate the goodness of a cluster configuration without any prior knowledge of the nature of the clusters (Rendón et al. 2011). In practice, external information such as class labels is often not available in many application scenarios. Therefore, in the situation where there is no external information available, internal validation indexes are the only option for cluster validation. This section presents four widely used and well-known internal validation indexes: Silhouette index (Rousseeuw 1987), Davies-Bouldin index, Dunn's index (Dunn 1974) and BIC

index (Schwarz 1978); used to assess the ideal number of clusters and the quality of clusters (Liu et al. 2010). Useful reviews of available validation techniques have been presented elsewhere (Halkidi et al. 2002, Datta & Datta 2003, Kryszczuk & Hurley 2010).

3.1 Silhouette index

For a given cluster, $X_j(j = 1..c)$, a quality measure assigned to the ith sample of X_j which known as silhouette width. This value is a confidence indicator on the membership of ith sample in the cluster X_j and defined as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

where $a(i)$ is the average distance between the ithsample and all of samples belong to X_j, $b(i)$ is the minimum distance between the ithsample and all of samples clustered in $X_k(k = 1...c; k \neq j)$. Thus for a given cluster $X_j(j = 1..c)$, it is possible to calculate a cluster silhouette S_j, which characterizes the heterogeneity and isolation properties of such a cluster:

$$S_j = \frac{1}{m}\sum_{i=1}^{m} s(i) \tag{2}$$

where m is number of samples in X_j. It has been shown that for any partition $U \longleftrightarrow X : X_1 \bigcup ...X_i \bigcup ...X_c$, Global Silhouette value can be used as an effective validity index for U.

$$GS_u = \frac{1}{c}\sum_{j=1}^{c} S_j \tag{3}$$

if c is the number of clusters for partition U: a maximum value of GS_u indicates the better cluster configuration for a given dataset.

3.2 Dunn's index

The idea of this index (Dunn 1974) is based on clustering compactness and good separation:

$$D(U) = \min_{1\leq i\leq c}\left\{\min_{1\leq j\leq c_{j\neq i}}\left\{\frac{\delta(X_i, X_j)}{\max\Delta(X_k)}\right\}\right\} \tag{4}$$

if $U = X_i \bigcup ...X_j \bigcup ...X_c$, the $\delta(X_i, X_j)$ is the inter-cluster distance between clusters i and j and $\Delta(X_k)$ is the intra-cluster distance for cluster K. The main goal of this measure is to minimize intra-cluster distance and maximize the inter-cluster distance.

3.3 Davies-Bouldin index

To express how far clusters are located from each other and how compact they are, the Davies-Bouldin index can be used. The Davies-Bouldin index can be defined as:

$$DB(U) = \frac{1}{c}\sum_{i=1}^{c}\max_{i\neq j}\left\{\frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)}\right\} \tag{5}$$

if $U = X_i \bigcup ...X_j \bigcup ...X_c$, $\Delta(X_i)$ and $\Delta(X_j)$ represent the intra-cluster distance and $\delta(X_i, X_j)$ define as the inter-cluster distance. Therefore, the number of clusters that minimizes DB is chosen as the optimal number of clusters.

3.4 BIC index

This index is presented to avoid over-fitting in a dataset and is defined as:

$$BIC = -ln(L) + vln(n) \tag{6}$$

where n is the number of objects, L is the likelihood of the parameters to generate the data in the model and v is the number of free parameters in the Gaussian model. The BIC index takes into account both the fit of the model to the data and the complexity of the model: a model that has a smaller BIC is better.

There are different methods available to calculate the intra-cluster and inter-cluster distances based on functions defined on the set of all sample pairs (Azuaje 2002). Here we present two examples for each of these distances to show their diversity. Various well known metrics are used to calculate the distance between two samples, $d(x, y)$, such as Euclidean and Manhattan metrics (Salzberg 1991). In this paper we use the Euclidean metric as it is computationally simple.

3.4.1 Inter-cluster Distance

Single linkage is defined as the closest distance between two samples which belong to different clusters.

$$\delta(S, T) = \min d(x, y)_{x\in S, y\in T} \tag{7}$$

Complete linkage is represented by the distance between two remote samples in different clusters.

$$\delta(S, T) = \max d(x, y)_{x\in S, y\in T} \tag{8}$$

If S and T represent clusters from partition U and $d(x, y)$ defines the pair-wise distance between samples in S, T.

3.4.2 Intra-cluster distance

Complete diameter is defined as maximum distance between two samples belonging to the same cluster:

$$\Delta(S) = \max d(x, y)_{x,y\in S} \tag{9}$$

Average diameter is defined as the average distance between all of the samples in same cluster.

$$\Delta(S) = \frac{1}{|S|.(|S| - 1)}\sum_{x,y\in S, x\neq y} d(x, y) \tag{10}$$

Where $|S|$ represents the number of samples in the cluster S and $d(x, y)$ is the distance between two samples $x, y \in S$.

In this paper we use complete linkage for the intra-cluster distance and complete diameter for inter-cluster distance to calculate the Dunn's index, Davies-Bouldin index and Silhouette index.

We used the Cluster Validity Analysis Platform (CVAP) (Wang et al. 2009) to run K means and compute Davies-Bouldin, Dunn's index and Silhouette index. Latent Gold software(*Welcome to Statistical Innovations Inc.* 2011) was used to run Latent Class Analysis and compute BIC. Also, we used Matlab to compute Dunn's index and Silhouette index for LCA .

4 Constructing a synthetic dataset

In our data mining research, we are using the 45 and Up Study baseline dataset (*Study Overview* 2011). The 45 and Up Study is a large-scale cohort involving 266,848 men and women aged 45 years and over from New South Wales (NSW), Australia. Participants in the 45 and Up Study were randomly sampled from the database of Australia′s universal health insurance provider, Medicare Australia, which provides virtually complete coverage of the general population. Participants joined the Study by completing a baseline questionnaire (between February 2006 and April 2009) and giving signed consent for follow-up and linkage of their information to a range of health databases. The baseline questionnaire (available at http://www.45andup.org.au) collected measures of general health, health related behaviors and demographic and social characteristics.The overall response rate was 18%. The Study is described in detail elsewhere (Banks et al. 2008). In addition, it is planned to follow up the cohort every five years (Banks et al. 2009). Thus this study is of interest for many researchers to evaluate and develop longitudinal data mining methods. However, this study has finished only its first stage of collecting data and is currently entering the second phase to provide the first time step after baseline. It is necessary to have longitudinal datasets to test and evaluate longitudinal clustering methods; therefore as part of our project we are interested to create a synthetic longitudinal dataset based on this study. This section aims to explain our procedure for creating of a synthetic dataset. For the sake of simplicity we chose two variables, Body Mass Index (BMI) and the amount of Physical Activity (PA) for each case. Body Mass Index (BMI) was calculated from weight and height as self-reported on the baseline survey. After excluding people with a reported BMI of <15 or >50 kg/m2 or unknown BMI, BMI was categorized using the following cut-points: 15 (underweight), 18.5, 20 and 22.5 (normal weight), 25 and 27.5 (overweight), 30 (obese). Participants overall level of physical activity was classified according to their responses to elements of the Active Australia Questionnaire (of Health & Welfare AIHW), comprising information on number of weekly sessions (of any duration) of moderate and vigorous physical activity and episodes of walking for longer than 10 min. A weighted weekly average for number of sessions was calculated for each participant by adding the total number of sessions, with vigorous activity sessions receiving twice the weighting of moderate activity or walking sessions, and was categorized as $0-3, 4-9, 10-17 and 18$or more sessions per week. From about 160,000 cases we randomly chose about 1,000 cases from the first stage of the 45 and Up Study dataset, to represent the first time step of data. The LCA method was used to cluster our baseline data and based on the established BIC (Schwarz 1978) we determined the number of clusters which minimized the BIC index. The 45 and Up Study has primary ethical approval from the University of New South Wales Human Research Ethics Committee (HREC 05035). The main goal in creating a longitudinal synthetic dataset is to explain cluster behaviour in the next time step, as exhibited by either merging (clusters vanishing) or splitting (creating new clusters). Therefore we seek a parameter point in our sequence of different synthetic datasets at which the number of clusters change in situation, as a means to explain where the characteristics of the dataset changes. What proportion of data and by how much our data should change, is a fundamental question in this step. To address this question we investigated two different scenarios.

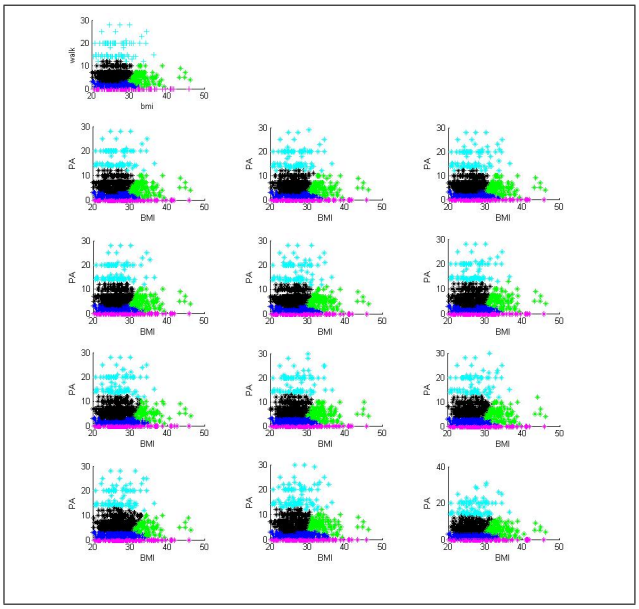

Figure 1: sequence of changing elements of larger cluster (BMI and PA increased by random number up to variance)

Scenario 1: we decided to follow a systematic change pattern for successively every five percent of the larger cluster (in terms of number of elements), for each element, adding a normally distributed random number in the range(-variance, +variance) of the targeted cluster. At each step we ran LCA to cluster the new time step data and we determined how many elements moved from one cluster to another cluster. The results in Figure 1 show that by applying this amount of change we observed cluster boundary positions changing somewhat, as might be expected.

Senario2 : we decided to change successively the element values of every five percent of the larger cluster (in terms of number of elements), for each element adding a normally distributed random number in the range(-2*variance, +2*variance) of the targeted cluster. At each step we ran LCA to cluster the new time step data and we determined how many elements actually moved from one cluster to another cluster. With this amount of change we might expect the targeted cluster to split as well as experiencing element movements. The results in Figure 2 show that after changing 40% of elements, our targeted cluster was splitting.

Based on this approach, three different datasets were chosen, one being the baseline dataset and the other two from the synthetic datasets, to compare LCA and K-means. The first synthetic dataset was created by increasing Physical Activity and Body Mass index for 40% of samples, in the range of a random number with normal distribution up to twice the variance. Finally, we chose the second synthetic dataset with only 20% of samples changed in the range of a random number with normal distribution up to the variance. With these three datasets we investigated the stability of each clustering method in the situations resulting from the changes in datasets.

5 Comparison of validation techniques

As discussed before in section 4, we chose three different datasets to compare LCA and K-means methods

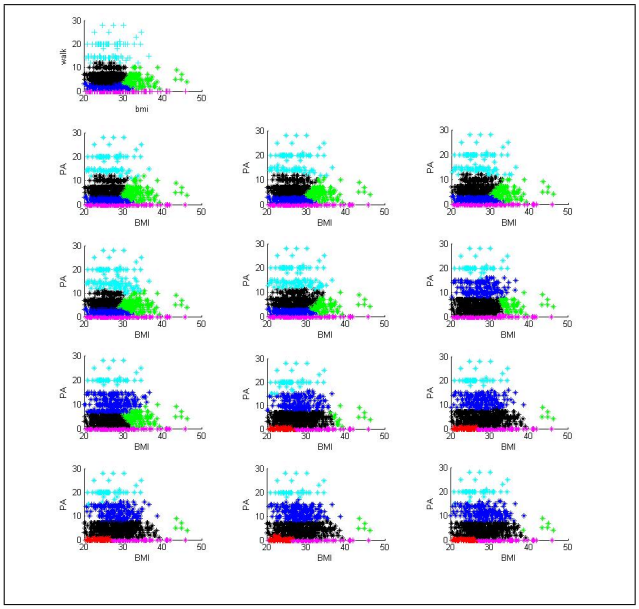

Figure 2: sequence of changing elements of larger cluster (BMI and PA increased by random number up to twice the variance)

for clustering our datasets (K-means method is chosen due to wide use). Table 1 shows the different range of validation methods for LCA and K-means clustering methods to cluster our baseline dataset. There are four different validation indexes computed for both LCA and K-means methods, while the number of clusters in each method varied from $K = 2...8$. The bold entries correspond to the optimal value predicted by each validation index.

Table 1: validation index- LCA and K means clustering for baseline dataset

Latent Class Analysis			
k	BIC	Silhouette	Dunn
2-Cluster	12099.3657	**0.5999**	1.3078
3-Cluster	11673.6888	0.4348	1.3251
4-Cluster	11606.441	0.4458	1.3723
5-Cluster	**11590.8389**	0.2919	1.3724
6-Cluster	11599.435	0.3435	1.9116
7-Cluster	11608.1122	0.4257	1.9115
8-Cluster	11634.5377	0.4538	**1.9979**
K means			
K	Silhouette	Davies-Bouldin	Dunn
2-Cluster	**0.46296**	0.88306	**2.0997**
3-Cluster	0.39821	0.85408	1.4749
4-Cluster	0.36353	0.82917	1.1644
5-Cluster	0.34664	0.8032	1.1575
6-Cluster	0.35132	0.73965	1.0911
7-Cluster	0.37077	**0.68475**	1.2532
8-Cluster	0.36513	0.69344	1.0437

In Table 1, clustering the baseline dataset using LCA shows the Silhouette index suggests that $K = 2$ has the best cluster configuration and may also suggest $K = 8$ be considered as a second option because it has second highest value for this index. Dunn's index is maximized at $K = 8$ and it might be of interest to consider $K = 6, 7$ as other options for choosing the number of clusters as they have the highest values. The BIC is minimized at $K = 5$ and based on Silhouette index and, Dunn's index at least $K = 5$ would be a reasonable choice for the number of clusters by using LCA clustering. Using K-means method, the Silhouette and Dunn's indexes are maximized at $k = 2$ ($K = 3$ has the second highest value for both of these indexes). However, the Davies-Bouldin index indicates that $K = 7$ has the best cluster configuration.

Table 2 shows the result for a synthetic dataset

Table 2: validation index- LCA and K means clustering for synthetic dataset with 20 percent change

Latent Class Analysis			
K	BIC	Silhouette	Dunn
2-Cluster	11251.4222	**0.5915**	1.3158
3-Cluster	10848.3937	0.4319	1.3336
4-Cluster	10805.5082	0.4472	1.3723
5-Cluster	**10790.0733**	0.3058	1.3723
6-Cluster	10800.7002	0.3137	1.3723
7-Cluster	10808.5815	0.4747	1.3723
8-Cluster	10826.1286	0.3669	**1.9167**
K means			
K	Silhouette	Davies-Bouldin	Dunn
2-Cluster	**0.4490**	0.9025	**2.0737**
3-Cluster	0.3968	0.8388	1.4866
4-Cluster	0.3632	0.8230	1.2061
5-Cluster	0.3417	0.7306	1.1751
6-Cluster	0.3527	0.8009	1.1495
7-Cluster	0.3749	0.7407	1.2887
8-Cluster	0.3669	**0.6597**	1.0308

with 20% change of elements; Silhouette index for LCA clustering is maximized in the 2 cluster solution, however, the second highest value is the 7 cluster solution that one may infer as a second option for number of clusters. BIC suggests the 5 cluster solution and Dunn's index indicates 8 cluster solution for this synthetic dataset. With K-means clustering algorithm, Silhouette and Dunn's indexes suggest the 2 cluster solution while Davies-Bouldin index indicates that the 8 cluster solution is the optimal number of clusters.

Table 3: validation index- LCA and K means clustering for synthetic dataset with 40 percent change

Latent Class Analysis			
K	BIC	Silhouette	Dunn
2-Cluster	12215.7611	0.5023	1.3349
3-Cluster	11891.1496	0.3667	1.3185
4-Cluster	11873.3579	0.3752	1.3368
5-Cluster	11859.5514	0.3792	1.3723
6-Cluster	**11857.5536**	**0.5044**	1.86
7-Cluster	11870.1069	0.4038	1.9478
8-Cluster	11885.8412	0.3941	**1.9765**
K means			
K	Silhouette	Davies-Bouldin	Dunn
2-Cluster	**0.45064**	0.88824	**2.1344**
3-Cluster	0.38521	0.84444	1.4926
4-Cluster	0.35882	0.89291	1.3874
5-Cluster	0.34747	0.79671	1.2
6-Cluster	0.36214	0.81538	1.2627
7-Cluster	0.36771	**0.7499**	1.3059
8-Cluster	0.36395	0.80687	1.2128

Based on reported results in Table 3, using K-means for that synthetic dataset, the Silhouette index and Dunn's index suggest the 2 cluster solution and Davies-Bouldin index suggests the 7 cluster solution. Validation indexes for Latent class analysis show a more promising result, as presented in Table 3, with the optimal number of clusters based on BIC and Silhouette index a 6 cluster solution and the highest range of Dunn's index for $K = 6, 7, 8$.

6 Conclusions

The fundamental problem in unsupervised learning using clustering methods is to determine the number of clusters. Some methods like K-means try to assign each case to a cluster based on distance from each cluster center while others work based on the posterior probability of each case. Either way, using validation methods would give some insight for better understanding the quality of each cluster, and then based on specific domain knowledge decide on which clustering method is used and which model best ex-

plains the characteristics of our data. In this paper, we have used the K-means and LCA algorithms to cluster our data, and to explain a clustering solution for a synthetic dataset. Results of this work indicate that with a small change in data, LCA would still discover almost the same clusters as in the baseline dataset.

7 Acknowledgment

This research was completed using data collected through the 45 and Up Study (www.45andUp.org.au). The 45 and Up Study is managed by the Sax Institute in collaboration with major partner Cancer Council New South Wales; and partners the National Heart Foundation (NSW Division); NSW Health; Beyondblue: the national depression initiative; Ageing, Disability and Home Care, NSW Department of Human Services; and Uniting Care Ageing. We thank the many thousands of people participating in the 45 and Up Study.

References

Azuaje, F. (2002), 'A cluster validity framework for genome expression data', *Bioinformatics* **18**(2), 319–320.

Banks, E., Jorm, L., Lujic, S. & Rogers, K. (2009), 'Health, ageing and private health insurance: baseline results from the 45 and up study cohort', *Australia and New Zealand health policy* **6**(1), 16.

Banks, E., Redman, S., Jorm, L., Armstrong, B., Bauman, A., Beard, J., Beral, V., Byles, J., Corbett, S., Cumming, R. et al. (2008), 'Cohort profile: the 45 and up study', *Int J Epidemiol* **37**(5), 941–947.

Datta, S. & Datta, S. (2003), 'Comparisons and validation of statistical clustering techniques for microarray gene expression data', *Bioinformatics* **19**(4), 459.

Davies, D. & Bouldin, D. (1979), 'A cluster separation measure', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2), 224–227.

Dunn, J. (1973), 'A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters'.

Dunn, J. (1974), 'Well-separated clusters and optimal fuzzy partitions', *Journal of cybernetics* **4**(1), 95–104.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2002), 'Cluster validity methods: part i', *ACM Sigmod Record* **31**(2), 40–45.

Jain, A., Murty, M. & Flynn, P. (1999), 'Data clustering: a review', *ACM computing surveys (CSUR)* **31**(3), 264–323.

Kryszczuk, K. & Hurley, P. (2010), 'Estimation of the number of clusters using multiple clustering validity indices', *Multiple Classifier Systems* pp. 114–123.

Lanza, S., Flaherty, B. & Collins, L. (2003), 'Latent class and latent transition analysis'.

Likas, A., Vlassis, N. & J Verbeek, J. (2003), 'The global k-means clustering algorithm', *Pattern recognition* **36**(2), 451–461.

Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. (2010), Understanding of internal clustering validation measures, *in* 'Data Mining (ICDM), 2010 IEEE 10th International Conference on', IEEE, pp. 911–916.

Magidson, J. & Vermunt, J. (2002), 'Latent class models for clustering: A comparison with k-means', *Canadian Journal of Marketing Research* **20**(1), 36–43.

of Health, A. I. & Welfare(AIHW) (2003), 'The active australia survey: A guide and manual for implementation, analysis and reporting', **Catalogue no. CVD 22. Canberra: AIHW**.

Rendón, E., Abundez, I., Gutierrez, C. & DÍAZ, S. (2011), A comparison of internal and external cluster validation indexes, *in* 'Proceedings of the 2011 American conference', pp. 158–163.

Rousseeuw, P. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65.

Salzberg, S. (1991), 'Distance metrics for instance-based learning', *Methodologies for Intelligent Systems* pp. 399–408.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Study Overview (2011).
URL: *http://www.45andup.org.au/*

Wang, K., Wang, B. & Peng, L. (2009), 'Cvap: Validation for cluster analyses', *Data Science Journal* (0), 904220071.

Welcome to Statistical Innovations Inc. (2011).
URL: *http://www.statisticalinnovations.com/*