UNIVERSITI PUTRA MALAYSIA

IMPROVED CLUSTERING USING ROBUST AND CLASSICAL PRINCIPAL COMPONENT

AHMED KADOM HASSN

FS 2017 47

IMPROVED CLUSTERING USING ROBUST AND CLASSICAL PRINCIPAL COMPONENT

By

AHMED KADOM HASSN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Degree of Master of Science

June 2017

DEDICATION

- *TO my respectful father and lovely mother who taught me the meaning of courage and always had confidence in me.*
- *To my wife for all his contribution, patience, and understanding throughout my master studies. He supported me a lot and made it all possible for me.*
- *To my kids, who accompanied me through the different parts of my study. Their love has always been my greatest inspiration.*

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the Degree of Master of Science

IMPROVED CLUSTERING USING ROBUST AND CLASSICAL PRINCIPAL COMPONENT

By

AHMED KADOM HASSN

June 2017

Chairman : Anwar Fitrianto, PhD
Faculty : Science

k-means algorithm is a popular data clustering algorithm. k-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Finding the appropriate number of clusters for a given data set is generally a trial-and-error process which made more difficult by the subjective nature of deciding what constitutes ‘correct’ clustering. When dimension of data is large it is often difficult to apply k-means clustering algorithm since it needs lots of computational times.

To remedy this problem, we propose to integrate Principal Component analysis (PCA) which is useful for dimensionality reduction of a dataset with the k-means clustering algorithm. We call our propose method as k-means by principal components (pc1). In this study, the kernels that are created by using the k-means method are replaced with kernels which are created by using PCA method where the PCA method reduces the dimensionality of a data. The results of the study show that the k-means by PCA is faster and more efficient than the classical k-means algorithm.

The classical k-means algorithm and the k-means by PCA algorithm are very sensitive to the presence of outlier. Hence the k-means by robust PCA is developed to rectify the problem of outliers in the dataset.

The findings indicate that in the absence of outliers, the performances of both methods; the k-means by PCA and the k-means by robust PCA are equally good. Nonetheless, the k-means by robust PCA is not much affected by outliers compared to the k-means by classical PCA.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk Ijazah Sarjana Sains

PENAMBAHBAIKAN PENGELOMPOKAN DENGAN MENGGUNAKAN ANALISIS KOMPONEN UTAMA TEGUH DAN KLASIK

Oleh

AHMED KADOM HASSN

Jun 2017

Pengerusi : Anwar Fitrianto, PhD
Fakulti : Sains

Algoritma *k-means* ialah algoritma data kluster yang popular. Matlamat pengelompokan *k-means* adalah untuk membahagi cerapan n ke dalam kluster *k* dengan setiap cerapan adalah kepunyaan kluster dengan min yang terdekat, ianya berfungsi sebagai prototaip kluster. Mencari bilangan kluster yang sesuai untuk sesuatu set data adalah secara amnya suatu proses percubaan yang menjadi lebih sukar disebabkan sifat subjektif dalam menentukan apa yang merupakan pengelompokan yang 'betul'. Apabila dimensi data besar biasanya sukar untuk menggunakan algoritma pengelompokan k-means, kerana ianya memerlukan banyak masa pengkomputeran.

Untuk membetulkan masalah ini, kami mencadangkan untuk mengintegrasikan analisis komponen utama (PCA), di mana ianya berguna untuk pengurangan dimensi set data dengan algoritma pengelompokan *k-means*. Kami namakan kaedah yang dicadangkan sebagai *k-means* dari komponen utama (pc1). Dalam kajian ini, kernel-kernel yang dicipta dengan menggunakan kaedah *k-means* telah digantikan dengan kernel-kernel yang dicipta menggunakan kaedah PCA di mana kaedah PCA telah mengurangkan dimensi pada data tersebut. Keputusan dari kajian ini menunjukkan bahawa *k-means* dengan PCA adalah lebih cepat dan cekap daripada algoritma *k-means* klasik.

Algoritma *k-means* klasik dan *k-means* dengan algoritma PCA adalah lebih sensitif dengan kehadiran titik terpencil. Oleh itu, *k-means* dengan PCA teguh telah dicadangkan untuk membetulkan masalah titik terpencil di dalam set data.

Keputusan menunjukkan bahawa pencapaian kedua-dua kaedah dengan kehadiran titik terpencil; *k-means* dengan PCA dan *k-means* dengan PCA teguh adalah sama bagus. Walaubagaimanapun, *k-means* dengan PCA teguh tidak banyak terjejas dengan titik terpencil berbanding dengan *k-means* dengan PCA klasik.

ACKNOWLEDGEMENTS

First of all, I wish to thank God who always supported me in all difficulties of my study life.

To have successful children has been one of my parent's dreams. I tried as much as I can afford to fulfil their dreams in order to thank them sincerely for scarifying their life to grow up with me.

I would like to express my deep gratitude to my master thesis supervisor, Prof.Habshah and Dr.Anwar Fitrianto. I have learned many things since I became Prof.Habshah and Dr.Anwar Fitrianto student. They spent a lot of time teaching and guiding me how to do this research.

I could not possibly forget all the wonderful people that have offered me their friendship and have enriched my life during these master studies duration. My acknowledgement would be incomplete without mentioning my friends Hassan, Mohammed, Shelan, Ahmed and all the others who made wonderful memories for me. Thanks you all. Lastly, my special thanks to my kids and my wife whose patience is admirable for me. Without her undoubting faith, my thesis would never have been completed. My sincerely regards to my sisters, specially my elder one, my brother, who encouraged me not to miss my hope in doing my research and supported me a lot mentally.

My master studies wouldn't be possible without the scholarship granted to me by my country. Much gratitude is also due to the entire faculty of science members who created an environment in which master and PhD students can flourish. I am lucky to have the chance to be graduated from this faculty.

I certify that a Thesis Examination Committee has met on 2 June 2017 to conduct the final examination of Ahmed Kadom Hassn on his thesis entitled "Improved Clustering Using Robust and Classical Principal Component" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Master of Science.

Members of the Thesis Examination Committee were as follows:

Fudziah binti Ismail, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Abdul Ghapor bin Hussin, PhD
Professor
National Defence University of Malaysia
Universiti Putra Malaysia
(External Examiner)

Yong Zulina Zubairi, PhD
Associate Professor
University of Malaya
Malaysia
(External Examiner)

NOR AINI AB. SHUKOR, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 4 September 2017

This thesis was submitted to the Senate of the Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Master of Science.The members of the Supervisory Committee were as follows:

Anwar Fitrianto, PhD
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Habshah, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

____

ROBIAH BINTI YUNUS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software

Signature: ________________________________ Date: ________________

Name and Matric No: Ahmed Kadom Hassn, GS42479

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) were adhered to.

Signature: ______________________
Name of Chairman of Supervisory Committee: Dr. Anwar Fitrianto

Signature: ______________________
Name of Member of Supervisory Committee: Professor Dr. Habshah

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CPCA	Classical Principle component analysis
E	Matrix error
FA	Factor analysis
MCD	Minimum covariance determinant
n_k	The number of points in C_k
PCA	Principle component analysis
ROBPCA	Robust Principal component analysis
SVD	singular value decomposition
v_k	Eigenvectors satisfying
W	The loadings or weight matrix
$\hat{\Sigma}_0$	The covariance matrix

CHAPTER 1

INTRODUCTION

1.1 Introduction and Background of the study

Searching for ‘‘natural’’ groups of objects is an important exploratory technique for understanding complex data. The origins of clustering can be traced back to taxonomy where it is necessary that different people assign similar objects to the same group. Clustering or grouping were traditionally done by taxonomists who picked important grouping variables based on their rich knowledge of species. Nowadays, the principal function of clustering is to name, display, summarize and to elicit an explanation for the resulting partitions of the dataset (Hartigan, 1975).

Clustering defines as programmed grouping of similar circumstances created several dissimilarity amount in Statistics and Computer Science. Clustering is sometimes referred to as ‘‘numerical taxonomy’’.

For example, a DNA microarray is a type of dataset to which clustering algorithms are applied. A microarray is a rectangular array of N rows, one for each case (e.g. a patient or a tumor) and p columns, one for each feature (e.g. genes, SNP’s).

The dependable, precise plus robust arrangement of growths is vital for prosperous analysis of cancer. However, in medical a clinical presentation of microarray-established is to identify and diagnose cancer, the explanation of new growth sessions would be centered on the partitions produced by grouping. The clusters can formerly be applied to build forecasters for fresh tumor models (Dudoit et al., 2002).

Current applications of clustering algorithms often include a large number of features and visualizing such datasets is difficult. Typically, simply a moderately few numbers of feature is important to determine the class memberships of the cases.

If thousands of potential clustering features must be considered, the traditional taxonomists’ approach of hand picking important features becomes difficult and impractical. Instead, we need a method that automatically chooses the important clustering variables. Furthermore, large datasets may contain outliers, which are defined as cases that do not belong to any of the given clusters. In this situation, one may wish to use a wise algorithm that identifies the important features and outliers together with the clusters.

Principal component analysis is known as a general statistical technique which gives much insight and attempts to describe the covariance organization of data by values of a minor number of constituents. Therefore, these constituents are linear arrangements of the unique variables, and frequently agree for an explanation and a enhanced thoughtful of the dissimilar causes of variant. For the reason that PCA is disturbed with the data reduction, yet, it is commonly applied for the investigation of high-dimensional data which are normally come across in chemometrics, computer vision, engineering, genetics, and other fields. PCA is formerly and regularly the first approach of the data analysis, go alone with discriminant analysis, cluster analysis, or other multivariate techniques. However, as a result, it is essential to discover individuals main constituents that comprise maximum of the information. In the conventional technique, the leading constituent relates to the trend in which the projected observations recorded have the major sum of variance. The second constituent is then orthogonal to which the first and over again take full advantage of the variance of the data points projected on it. Persistently, in this method its yield entirely the principal constituents, whereby, relate to the eigenvectors of the experimental covariance matrix. Regrettably, both the conventional variance, which is being made best use of it and the conventional covariance matrix, which is being disintegrated are very delicate to abnormal interpretations. Accordingly, the first constituents are frequently involved in the direction of distant points, and it may not make use of the much difference of regular observations. Hence, reduction of data centered on classical PCA (CPCA) turn into undependable if outliers are existing in the data (Mia Hubert et al., 2005)

Almost the entire of the PCA algorithms stated previously are created on the expectations that data have not being damaged by outliers. The procedure is that, actual data regularly comprise various outliers and commonly they are not simply to be disjointed from the real data set (Chen, 2002) .

The major aim of robust PCA approaches is to attain principal constituents that are completely may not be affected considerably by outliers. The first set of techniques group is attained by swapping the conventional covariance matrix by a robust covariance estimator. (Campbell, 1980; Maronna et al., 1976) suggested to apply affine equi-variant M-estimators of scatter for this aim, nevertheless, these may not fight many outliers. Furthermore, recently (Croux et al., 2000) applied positive-breakdown estimators such as the minimum covariance determinant (MCD) technique (Edelsbrunner et al., 1990) and S-estimators (Davies, 1987; Mia Hubert et al., 2005; Leroy et al., 1987).

1.2 Statement of the problem

Data exploration techniques are very important for studying enormous quantity of high dimensional data. Principal component analysis (PCA) is a generally applied statistical method in non-parametric dimension reduction. The k-means cluster analysis is usually applied in data clustering for non-parametric learning responsibilities. On the other perspective, clustering examinations (Duda et al., 2012;

Friedman et al., 2001; Jain et al., 1988) and tries to give permission by which data pass quickly to achieve accessibility by first demand understanding and also by separating data points into disconnect groups so that similar data points be in the right place to same cluster, while data points which are not the same be in the right place to different clusters. The utmost common and capable clustering techniques is the k-means method (Hartigan et al., 1979; Lloyd, 1957; MacQueen, 1967) which uses models as centers to signify clusters by improving the squared cost function (detail explanation on k-means and associated ISODATA techniques, can be seen in (Jain & Dubes, 1988), and (Wallace, 1989)). On the other perspective, high dimensional data are frequently changed into lower dimensional data through the principal component analysis (PCA) where logical arrangements can be identified more obviously(Jolliffe, 2002b). This kind of unsupervised dimension reduction is applied in actual extensive fields such as meteorology, image processing, genomic analysis, and information retrieval. Therefore, it may also be general that PCA is applied to project data to lessen the dimensional subspace and k-means is formerly applied in the subspace (Zha et al., 2001). Considering other circumstances, data are inserted in a low-dimensional space like the Eigen space of the graph Laplacian, and k-means at that point used (A. Y. Ng et al., 2001). The major sources of PCA-based dimension reduction is that PCA choices up the magnitudes with the maximum variances. Mathematically, this is an alternative to seeking the paramount low rank estimation of the data through the singular value decomposition (SVD) (Eckart et al., 1936). Though, this distortion of anomaly reduction property only is insufficient to describe the helpfulness of PCA(Ding et al., 2004b).

In consideration of the classical approach to principal component analysis, the first constituent relates to the trend in which the projected interpretations have the biggest variance. The second constituent is therefore the orthogonal to the first constituent and yield better when using the variance of the data arguments projected on it. Persistently, in another perspective, in this manner it gives almost all the principal constituents, whereby its relates to the eigenvectors of the experimental covariance matrix. Regrettably, both the conventional variance is being used as the best and the conventional covariance matrix is being disintegrated and are very complex to abnormal explanations. Accordingly, the first constituents are regularly fascinated in the direction of faraway distant points, and would not point the discrepancy of the consistent observations. As a result, reduction of data established on classical PCA (CPCA) turn out to be undependable if outliers are existing in the data(Mia Hubert et al., 2005).

1.3 Objectives of the study

The principal components are essentially the continuous explanation of the cluster affiliation pointers in the k-means cluster analysis method. The PCA measurement is repeatedly reduce the executed data clustering agreeing to the k-means cost function. This however, affords an essential validation of PCA-based reduction of data. The outcomes also make available operational methods to explain the k- means cluster analysis issues. k-means approach applies k models, the centers of clusters, which exactly describe the data. (Ding & He, 2004b).

Usually, when considering the first components it generally and frequently fascinated in the direction of faraway distant points, and which possibly would not give the precise difference in variation of the systematic observations. Consequently, reduction of data is centered on classical PCA (CPCA) which develops as undependable if outliers are existing in the data (Mia Hubert et al., 2005).

The research problem can be outlined as follows:

- Clustering based on k-means method is very popular. However, as soon as the measurement of the data is big it may often difficult to apply k-mean cluster, because it needs lots of computational times. Therefore, computationally k-mean is very expensive for large dimension of data.
- Both PCA & k-mean clustering algorithm are affected by outliers. In this situation the use of robust PCA is recommended for clustering the data.

Based on statements of problem, the present study tries to arrive the following objectives:-

i) To develop k-mean clustering algorithm based on PCA data reduction technique.
ii) To formulate k-mean clustering algorithm based on Robust PCA in the presence of outliers.

1.4 Thesis Outline

In line with the objectives and scope of this research, the subjects of the thesis are arranged in five sections. After the introduction, the various sections are ordered such that research goals are clearly presented in the outlined sequence.

Chapter Two. This segment presents a concise survey of the literature which basically considering the cluster analysis on k-means to identify group data into homogeneous gatherings based on similarities through a set of attributes. The clustering analysis and principal component analysis (PCA) are highlighted in this chapter.

Chapter Three. In this chapter, the k-means clustering algorithm and the PCA are discussed. The k-mean clustering based on PCA is proposed to increase the efficiency of the clustering algorithm and at the same time reduces computational times. Monte carlo simulation study and numerical example are presented.

Chapter Four. This chapter described the proposed k-means clustering algorithm based on robust PCA to reduce the effect of outliers on determining the number of clusters. To evaluate the performance of the proposed method, monte carlo simulation study and real data applications are carried out.

Chapter Five. Finally**,** the chapter offers complete summarized and detailed discussion of some results, contributions, and recommendations for further research.

REFERENCES

Al-Daoud, M. D. B. (2005). *A new algorithm for cluster initialization.* Paper presented at the WEC'05: The Second World Enformatika Conference.

Aldenderfer, M. S., & Blashfield, R. K. (1984). Cluster analysis. Sage University paper series on quantitative applications in the social sciences 07-044.

Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm.

Balcan, M.-F. F., Ehrlich, S., & Liang, Y. (2013). *Distributed $ k $-means and $ k $-median Clustering on General Topologies.* Paper presented at the Advances in Neural Information Processing Systems.

Blashfield, R. K., & Aldenderfer, M. S. (1978). The literature on cluster analysis. *Multivariate Behavioral Research, 13*(3), 271-295.

Boente, G., Pires, A. M., & Rodrigues, I. M. (2002). Influence functions and outlier detection under the common principal components model: a robust approach. *Biometrika*, 861-875.

Brusco, M. J., & Cradit, J. D. (2001). A variable-selection heuristic for K-means clustering. *Psychometrika, 66*(2), 249-270.

Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied statistics*, 231-237.

Chen, H. (2002). Principal component analysis with missing data and outliers. *URL: http://www. cmlab. csie. ntu. edu. tw/~ cyy/learning/papers/PCA_Tutorial. pdf*.

Considine, J., Li, F., Kollios, G., & Byers, J. (2004). *Approximate aggregation techniques for sensor databases.* Paper presented at the Data Engineering, 2004. Proceedings. 20th International Conference on.

Corbett, J. C., Dean, J., Epstein, M., Fikes, A., Frost, C., Furman, J. J., . . . Hochschild, P. (2013). Spanner: Google's globally distributed database. *ACM Transactions on Computer Systems (TOCS), 31*(3), 8.

Croux, C., & Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 603-618.

Croux, C., & Ruiz-Gazen, A. (1996). *A fast algorithm for robust principal components based on projection pursuit.* Paper presented at the Compstat.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 1269-1292.

Davy, M., & Luz, S. (2007). *Dimensionality reduction for active learning with nearest neighbour classifier in text categorisation problems.* Paper presented at the Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on.

De La Torre, F., & Black, M. J. (2003). A framework for robust subspace learning. *International Journal of Computer Vision, 54*(1-3), 117-142.

Debruyne, M., & Hubert, M. (2005). The Influence Function of Stahel-Donoho Type Methods for Robust PCA.

Ding, C., & He, X. (2004a). *Cluster structure of k-means clustering via principal component analysis.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.

Ding, C., & He, X. (2004b). *K-means clustering via principal component analysis.* Paper presented at the Proceedings of the twenty-first international conference on Machine learning.

Duda, R. O., Hart, P. E., & Stork, D. G. (1973). *Pattern classification*: Wiley, New York.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*: John Wiley & Sons.

Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology, 3*(7), research0036. 0031.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*(3), 211-218.

Edelsbrunner, H., & Souvaine, D. L. (1990). Computing least median of squares regression lines and guided topological sweep. *Journal of the American Statistical Association, 85*(409), 115-119.

Engelen, Hubert, M., & Branden, K. V. (2016). A comparison of three procedures for robust PCA in high dimensions. *Austrian Journal of Statistics, 34*(2), 117-126.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* Paper presented at the Kdd.

Ester, M., Kriegel, H.-P., & Xu, X. (1995). *Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification.* Paper presented at the International Symposium on Spatial Databases.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Fearon, J. D. (1999). What is identity (as we now use the word). *Unpublished manuscript, Stanford University, Stanford, Calif*.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM, 24*(6), 381-395.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1): Springer series in statistics Springer, Berlin.

Garcia, J., Fdez-Valdivia, J., Cortijo, F., & Molina, R. (1995). A dynamic approach for clustering data. *Signal Processing, 44*(2), 181-196.

Gerbrands, J. J. (1981). On the relationships between SVD, KLT and PCA. *Pattern recognition, 14*(1), 375-381.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 81-124.

Greenhill, S., & Venkatesh, S. (2007). *Distributed query processing for mobile surveillance.* Paper presented at the Proceedings of the 15th ACM international conference on Multimedia.

Greenwald, M. B., & Khanna, S. (2004). *Power-conserving computation of order-statistics over sensor networks.* Paper presented at the Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.

Hannachi, A., Jolliffe, I., & Stephenson, D. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology, 27*(9), 1119-1152.

Hartigan, J. A. (1975). Clustering algorithms (probability & mathematical statistics): John Wiley & Sons Inc New York.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100-108.

Huber, P. J. (1981). Wiley series in probability and mathematics statistics. *Robust Statistics*, 309-312.

Hubert, M., & Engelen, S. (2004). Fast cross validation for high breakdown resampling algorithms: preparation.

Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics, 47*(1), 64-79.

Hubert, M., Rousseeuw, P. J., & Verboven, S. (2002). A fast method for robust principal components with applications to chemometrics. *Chemometrics and intelligent laboratory systems*, *60*(1), 101-111.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*: Prentice-Hall, Inc.

Jolliffe. (1986). Principal Component Analysis Springer New York Google Scholar.

Jolliffe. (2002a). *Principal component analysis*: Wiley Online Library.

Jolliffe. (2002b). Principal component analysis and factor analysis. *Principal component analysis*, 150-166.

Judd, D., McKinley, P. K., & Jain, A. K. (1996). *Large-scale parallel data clustering.* Paper presented at the Pattern Recognition, 1996., Proceedings of the 13th International Conference on.

Judd, D., McKinley, P. K., & Jain, A. K. (1998). Large-scale parallel data clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 871-876.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344): John Wiley & Sons.

Ke, Q., & Kanade, T. (2005). *Robust L/sub 1/norm factorization in the presence of outliers and missing data by alternative convex programming.* Paper presented at the Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.

Khaled, A., Sanjay, R., & Vineet, S. (1997). An efficient k-means clustering algorithm.

Kontos, D., & Megalooikonomou, V. (2005). Fast and effective characterization for classification and similarity searches of 2D and 3D spatial region data. *Pattern recognition, 38*(11), 1831-1846.

Krieger, A. M., & Green, P. E. (1996). Modifying cluster-based segments to enhance agreement with an exogenous response variable. *Journal of Marketing Research*, 351-363.

La Grange, A., Le Roux, N., & Gardner-Lubbe, S. (2009). BiplotGUI: interactive biplots in R. *Journal of Statistical Software, 30*(12), 1-37.

Leroy, A. M., & Rousseeuw, P. J. (1987). Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*.

Li, G., & Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *Journal of the American Statistical Association, 80*(391), 759-766.

Liang, Y., Balcan, M.-F., & Kanchanapally, V. (2013). *Distributed PCA and k-means clustering.* Paper presented at the The Big Learning Workshop at NIPS.

Lloyd, S. (1957). Least squares quantization in PCM's Bell Telephone Labs.

Lorr, M. (1983). *Cluster analysis for social scientists*: Jossey-Bass San Francisco.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations.* Paper presented at the Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.

Maronna, R. A., & Yohai, V. J. (1976). Robust estimation of multivariate location and scatter. *Wiley StatsRef: Statistics Reference Online*.

Mehrotra, K., Mohan, C. K., & Ranka, S. (1997). *Elements of artificial neural networks*: MIT press.

Michael, D., & Saturnine, L. (2007). *Dimensionality reduction for active learning with nearest neighbor classifier in text categorization problems.* Paper presented at the Sixth International Conference on Machine Learning and Applications.

Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. *Applied psychological measurement, 11*(4), 329-354.

Mitra, S., Agrawal, M., Yadav, A., Carlsson, N., Eager, D., & Mahanti, A. (2011). Characterizing web-based video sharing workloads. *ACM Transactions on the Web (TWEB), 5*(2), 8.

Morrison, D. F. (1990). Multivariate statistical methods. 3. *New York, NY. Mc*.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). *On spectral clustering: Analysis and an algorithm.* Paper presented at the NIPS.

Ng, G. Y., & Fan, A. C. (2001). Does elbow position affect strength and reproducibility of power grip measurements? *Physiotherapy, 87*(2), 68-72.

Ng, R. T., & Han, J. (1994). *E cient and E ective Clustering Methods for Spatial Data Mining.* Paper presented at the Proc. of.

Olston, C., Jiang, J., & Widom, J. (2003). *Adaptive filters for continuous queries over distributed data streams.* Paper presented at the Proceedings of the 2003 ACM SIGMOD international conference on Management of data.

Pham, D. T., Dimov, S. S., & Nguyen, C. (2004). An incremental K-means algorithm. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 218*(7), 783-795.

Pourkamali-Anaraki, F., & Becker, S. (2015). Preconditioned Data Sparsification for Big Data with Applications to PCA and K-means. *arXiv preprint arXiv:1511.00152*.

Ramasubramanian, V., & Paliwal, K. K. (1992). Fast k-dimensional tree algorithms for nearest neighbor search with application to vector quantization encoding. *IEEE Transactions on Signal Processing, 40*(3), 518-531.

Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212-223.

Schikuta, E. (1996). *Grid-clustering: An efficient hierarchical clustering method for very large data sets.* Paper presented at the Pattern Recognition, 1996., Proceedings of the 13th International Conference on.

Slater, S. F., & Olson, E. M. (2001). Marketing's contribution to the implementation of business strategy: An empirical analysis. *Strategic Management Journal, 22*(11), 1055-1067.

Steinley, D. (2006). K- means clustering: A half- century synthesis. *British Journal of Mathematical and Statistical Psychology, 59*(1), 1-34.

Tonidandel, S., & Overall, J. E. (2004). Determining the number of clusters by sampling with replacement. *Psychological Methods, 9*(2), 238.

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). *Constrained k-means clustering with background knowledge.* Paper presented at the ICML.

Wallace, R. S. (1989). Finding natural clusters through entropy minimization.

White, J. M., Faber, V., & Saltzman, J. S. (1995). Population attribute compression: Google Patents.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems, 2*(1-3), 37-52.

Woodruff, D. L., & Rocke, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association, 89*(427), 888-896.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks, 16*(3), 645-678.

Yeung, K. Y., & Ruzzo, W. L. (2000). An empirical study on principal component analysis for clustering gene expression data: Technical report, Department of Computer Science and Engineering, University of Washington.

Zha, H. (2002). *Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering.* Paper presented at the Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.

Zha, H., He, X., Ding, C., Gu, M., & Simon, H. D. (2001). *Spectral relaxation for k-means clustering.* Paper presented at the NIPS.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). *BIRCH: an efficient data clustering method for very large databases.* Paper presented at the ACM Sigmod Record.