# A Simulation Study: Cluster Analysis in High Dimensional Space in Psychology Research

Jiayin Qu

Dr. Aaron Schmidt

UNIVERSITY OF MINNESOTA

Driven to Discover®

# Study Summary

- A massive volume of data have been collected in I/O field, resulting in numerous variables available for each object (e.g., participants, teams, orgs).
- For researchers who'd like to apply cluster analysis to datasets, the concern about the "curse of dimensionality" comes with an increasing number of variables available to be measured and analyzed in psychology research.
- While multiple mechanisms can lead to identical or similar peaking effect, the current simulation study performed in R aimed to tell them apart.

- The results suggest that the "curse of dimensionality" is likely to be due to the presence of irrelevant features in the high-dimensional space, rather than overfitting to the training sample.
- The result implies that researchers should still act as gatekeepers to make informative decisions when collecting and analyzing data in order to prevent the "Garbage in, garbage out" situation.

Corresponding Author can be reached at quxxx179@umn.edu

# Background: Cluster Analysis

- **Cluster analysis** is a machine learning technique using measures of similarity to separating *objects* into clusters in order to explore, confirm, or simplify the data.
  - **When the major purpose is exploration:**
    People can be clustered into different leadership styles.
    Teams can be clustered based on their communication styles and team functions.
    Organizations can be clustered based on their strategies.
  - **When the major purpose is confirmation:**
    Predicting personality traits by comparing one's most frequently used word in social media with a training set whose personality traits are known (Pratama & Sama, 2015).
    Recognizing emotions in automated assessments and video interviews by comparing the facial expression with priori classification (Khan, Goskula, Nasiruddin, & Quazi, 2011)
  - **When the major purpose is simplification:**
    Predict *g* using trace data from games (Landers & Schmidt, 2016).
    Predict employee performance from social media interactions (Auer et al., 2019).

- **Objects** are vectors in a multi-dimensional space, where each dimension represents a distinct variable of the object (Kaufman & Rousseeuw, 2009).
- In cluster analysis, **a set of m items that have n variables** (e.g., a set of teams that have variables like communication styles, cohesions, functions, etc.) can be described as a m by n pattern matrix describing:

$$m \text{ objects} \overbrace{\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}}^{n \text{ variables}}$$

- **Similarity of objects** is measured by their distances. One of the most commonly used proximity measures is the *Euclidean Distance* (Steinbach, Ertöz, & Kumar, 2004). It is calculated by:

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

(n is the dimensionality of the data object, and p and q are vectors, starting from the origin of the space to their terminal points in a Euclidean n-space)

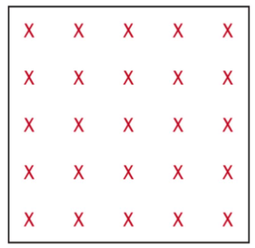
Corresponding Author can be reached at quxxx179@umn.edu

# Background: Curse of Dimensionality

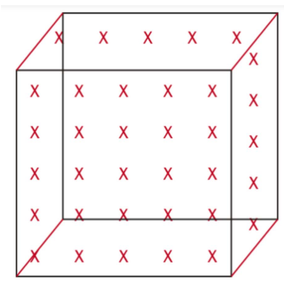
- **The curse of dimensionality**
  - The variable space becomes more sparse as the number of variables increases for a given number of observations (m). To maintain the average distance between data points, the number of observations needs to grow exponentially ( Steinbach et al., 2004).
  - That is, $10^2$ data points are needed in two-dimensional space and $10^3$ data points are needed in three-dimensional space to maintain the average distance.

$$m \text{ objects} \overbrace{\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}}^{n \text{ variables}}$$

5 data points in one dimension



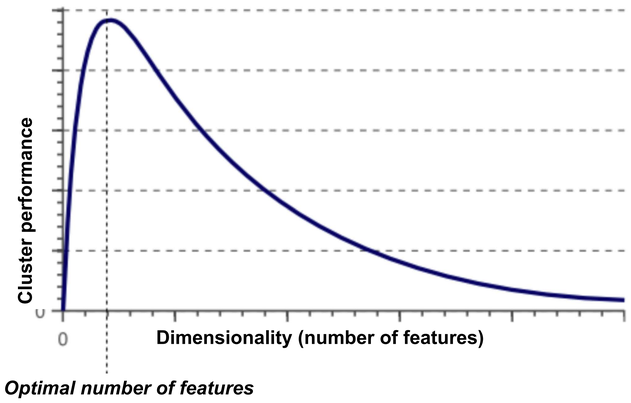
25 data points in two dimensions



125 data points in three dimensions

- **The curse of dimensionality results in the peaking phenomenon**
  - As the number of variables increases, the classifier's performance—the percentage of accurate classification obtained on testing subset after being trained on the training subset—increases until the optimal number of variables is reached.
  - Researchers have proposed *two major underlying mechanisms* of the peaking phenomenon:

More variables can offer more information, resulting in better clustering (Hughes, 1968; Kotsiantis et al., 2006).




**Potential mechanism 1:**
Overfitting of the classifier to the training sample due to a finite number of samples in a high dimensional space. (Hsieh & Landgrebe, 1998).

**Potential mechanism 2:**
The amount of noise is likely to increase when dimensions increase (Sima & Dougherty, 2008).

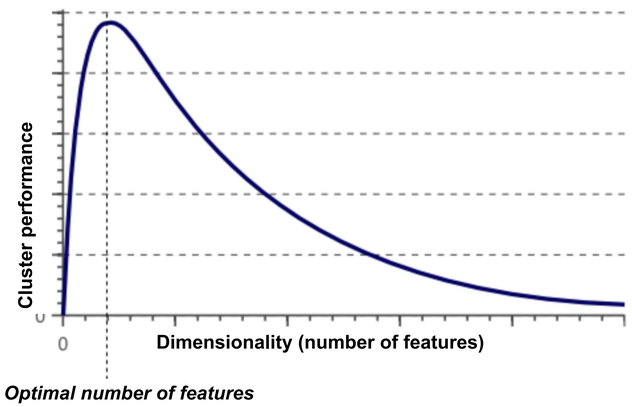
Corresponding Author can be reached at quxxx179@umn.edu

# Background: Supervised vs. Unsupervised Analysis

- As **multiple mechanisms** can be involved in data analysis in high dimensional space and lead to identical or similar peaking effect, it is important to tell them apart in research to have a clearer understanding for researchers employing the technique.

- **In the current study**, we made use of the similarity and difference between supervised k-NN and unsupervised K-means algorithms to examine the mechanisms:

| | Supervised (k-NN) | Unsupervised (K-means) |
|---|---|---|
| **Similarity** | | |
| | Both divide objects into well separable groups based on their affiliated characteristics using Euclidean Distance | |
| **Difference** | | |
| Goal: | Establish generalizable rules for classifying future data (Kotsiantis et al., 2006) | Discover the natural groupings of a whole set of data (Friedman, et al., 2001). |
| Training: | Train classifier on training subsets, test on testing subsets | It does not separate testing and training subsets |
| Performance: | Performance = percentage of accurate classification obtained on testing subset | Performance = minimizing the sum of the squared error over all k clusters. |




**Potential mechanism 1:**
Overfitting of the classifier to the training sample due to a finite number of samples in a high dimensional space. (Hsieh & Landgrebe, 1998).
**→ k-NN (with the training sample) would suffer from overfitting, *but* K-means would benefit from more information.**

**Potential mechanism 2:**
The amount of noise is likely to increase when dimensions increase (Sima & Dougherty, 2008).
**→ K-means and k-NN algorithms would both suffer from too much noise.**

Corresponding Author can be reached at quxxx179@umn.edu

# Current Problem

- **Current problems**
  - The ”Curse of Dimensionality” coming with an increasing number of variables available to be measured and analyzed in psychology research, has become crucial for researchers who want to apply machine learning to research.
  - We created a high-dimensional space to exam the effect of overfitting of the classifier to the training set, and then introduced irrelevant features to exam the effect of noise on the peaking phenomenon.

- **Research Question: Is the peaking phenomenon occurred in high dimensional space caused by the effect of overfitting of the classifier to the training sample, the effect of a higher likelihood of including noise, or a conjoint effect?**
  - *Hypothesis 1:* If the peaking phenomenon is due to overfitting to the training sample, K-means algorithm would not show the peaking phenomenon in the high-dimensional space, but k-NN would show such phenomenon.
  - *Hypothesis 2:* If the peaking phenomenon is due to a higher likelihood of including noise (irrelevant features), both K-means and k-NN algorithms would suffer from the curse of dimensionality.

Corresponding Author can be reached at quxxx179@umn.edu

# Simulation and Analysis

- **Clustered data were simulated under the R package fungible with the function monte() (Waller, 2019).**

- **To test hypothesis 1:** If the peaking phenomenon is due to overfitting to the training sample, K-means algorithm would not show the peaking phenomenon in the high-dimensional space, but k-NN would show such phenomenon.
  - For each dataset all observations were sampled from a mixture of k multivariate normal distributions.
  - These simulated samples differed along five parameters: (a) indicator validity, (b) number of variables, (c) number of clusters, (d) number of objects in each cluster.
  - The simulated samples were first used to perform K-means analysis and then were used in 10-fold cross-validation (they were partitioned into 10 equal size subsamples and used one as the testing data and the remaining as training data) for *k*-NN analysis.

- **To test hypothesis 2:** If the peaking phenomenon is due to a higher likelihood of including noise (irrelevant features), both K-means and k-NN algorithms would suffer from the curse of dimensionality.
  - Irrelevant variables were then generated by a random number generator that followed the same distribution as the simulated cluster data with a mean of 0.00 and a standard deviation of 1.00.
  - Those variables were then combined to the simulated clustered data to evaluate the performance of K-means and k-NN algorithms.

- **To assess cluster analysis' performance:**
  - In K-means, Adjusted Rand Index (ADI): compare the pre-labeled dataset and the clustering performed by K-means algorithm; ADI closer to 1 is more desirable as it signaled fewer mis-classification.
  - In k-NN, k-NN accuracy (%): use 10-fold cross-validation and calculate the number of accurate classifications divided by the number of total objects in the testing samples. k-NN accuracy closer to 100 is more desirable.

Corresponding Author can be reached at quxxx179@umn.edu

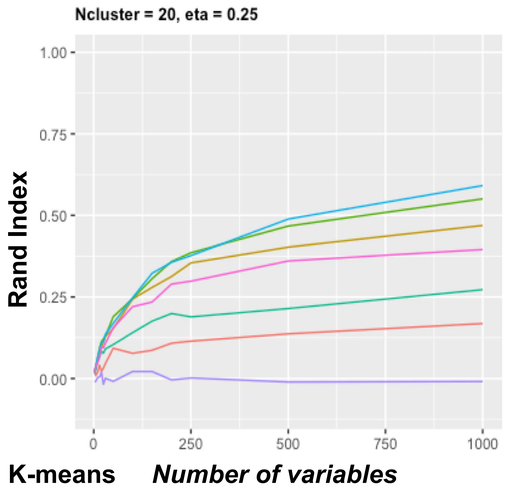
# General Findings

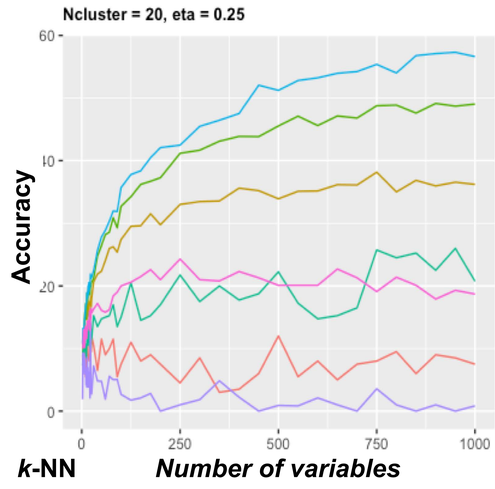
- If all variables are meaningful and relevant to the clustering, **K-means and k-NN should not result in the peaking phenomenon** in the high-dimensional space as long as the sample size is sufficient (a sample size of 50 should be sufficient).
- **The "Curse of Dimensionality" is likely to be due to** the presence of irrelevant features in the dataset, rather than overfitting to the training sample.

**Potential mechanism 1:**
Overfitting of the classifier to the training sample

→ k-NN would suffer from overfitting, *but* K-means would benefit from more information (H1).

**Result:**
As the number of variables available increased, K-means and k-NN both performed better and did not suffer from overfitting.




**Hypothesis 1 was not supported:**
When all variables provided equally meaningful information to the clustering, more variables only provided more information for better performance.

**Potential mechanism 2:**
The amount of noise is likely to increase when dimensions increase.

→ K-means and k-NN algorithms would both suffer from too much noise (H2).

**Result:**
The more a dataset included irrelevant variables, the worse K-means and k-NN perform.

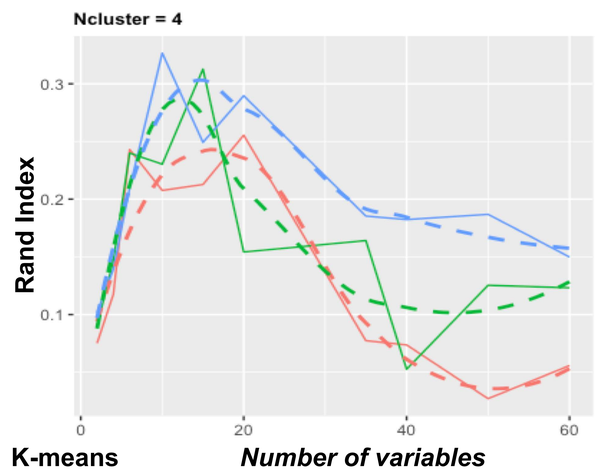
| Manipulated Variables | *k*-NN Accuracy (%) | K-means Adjusted Rand Index |
|---|---|---|
| % of noise | | |
| low (25%) | 30.33 | 0.25 |
| middle (25-75%) | 23.6 | 0.11 |
| high (75%) | 17.4 | 0.017 |

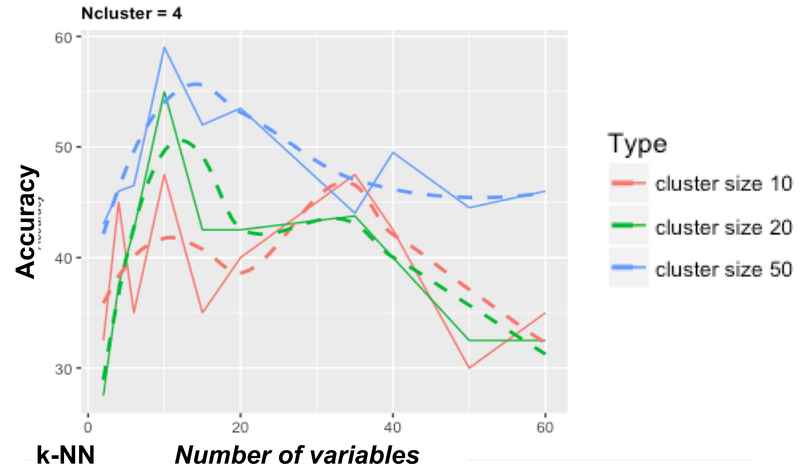
**Hypothesis 2 was supported:**
The peaking phenomenon resulted from more irrelevant variables in high dimensional space.

Corresponding Author can be reached at quxxx179@umn.edu

# Discussion

- **To illustrate the idea of hypothesis 2:**
  - We simulated a dataset with 10 relevant variables and 50 irrelevant variables.
  - We performed k-NN and K-means algorithm by specifying the number of variables that the algorithms could randomly choose among total variables.
  - The probability of including more irrelevant variables thus rose as the dimensions (the number of variables) increase.
  - The peaking effect was shown in both K-means and k-NN analysis.




- **Future research** should:
  - Evaluate datasets that, for example, have unequal size of clusters, do not follow normal distribution, and include redundant variables that can be represented by other variables.
  - Conduct in-depth analyses to examine feature selection algorithms such as subspace clustering, random mapping, etc.

- Over the last decade, there is an apparent trend in psychology to generate and collect a massive volume of data and collect as many information as one can obtain. **The result from the current study implies** that researchers should still act as gatekeepers in the academic world and make informative decisions when collecting and analyzing data in order to prevent the “Garbage in, garbage out” situation.

Corresponding Author can be reached at quxxx179@umn.edu

# References


Alpaydin, E. (2009). Introduction to machine learning. MIT press.
Auer, E. M., Marin, S., Landers, R. N., Collmus, A. B., Armstrong, M. B., Mujcic, S., & Blaik, J. A. (2019, April). Predicting g with Trace Data: Evidence from a Game-based Assessment. In J. F. Capman (Chair), Looking Under the Hood: Making Use of Trace Data. Symposium presented at the 34th Annual Conference of the Society for Industrial and Organizational Psychology, National Harbor, MD.
Bellman, Richard (1961) Adaptive Control Processes: A Guided Tour. Princeton University Press.
Borgen, F. H., & Barnett, D. C. (1987). Applying cluster analysis in counseling psychology research. Journal of Counseling Psychology, 34(4), 456.
Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52(3), 345-370.
Domeniconi, C., Papadopoulos, D., Gunopulos, D., & Ma, S. (2004, April). Subspace clustering of high dimensional data. In Proceedings of the 2004 SIAM international conference on data mining (pp. 517-521). Society for Industrial and Applied Mathematics.
Domingos, P. M. (2012). A few useful things to know about machine learning. Commun. acm, 55(10), 78-87.
Eys, M. A., Loughead, T. M., & Hardy, J. (2007). Athlete leadership dispersion and satisfaction in interactive sport teams. Psychology of Sport and Exercise, 8(3), 281-296.
Facca, T. M., & Allen, S. J. (2011). Using cluster analysis to segment students based on self-reported emotionally intelligent leadership behaviors. Journal of Leadership Education, 10(2), 72-96.
Fahim, A. M., Saake, G., Salem, A. M., Torkey, F. A., & Ramadan, M. A. (2008). K-means for spherical clusters with large variance in sizes. Journal of World Academy of Science, Engineering and Technology.
Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.
Guo G, Wang H, Bell D, Bi Y, Greer K (2003) KNN model-based approach in classification. Lect Notes Comput Sci 2888:986–996
Hassanat, A. B., Abbadi, M. A., Altarawneh, G. A., & Alhasanat, A. A. (2014). Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. arXiv preprint arXiv:1409.0919.
Hatten, K. J., Schendel, D. E., & Cooper, A. C. (1978). A strategic model of the US brewing industry: 1952-1971. Academy of Management journal, 21(4), 592-610.
Henry, D. B., Tolan, P. H., & Gorman-Smith, D. (2005). Cluster analysis in family psychology research. Journal of Family Psychology, 19(1), 121.
Hinneburg, A., & Keim, D. A. (1999). Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering.
Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of classification, 2(1), 193-218.
Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. IEEE transactions on information theory, 14(1), 55-63.
Hsieh, P. F., & Landgrebe, D. (1998). Classification of high dimensional data. ECE Technical Reports, 52.
Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.
Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data (Vol. 6). Englewood Cliffs: Prentice hall.
Jiřina, M., & Jiřina jr, M. (2008). Classifier based on inverted indexes of neighbors (Vol. 1034). Technical Report No.
Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.
Kaski, S. (1998, May). Dimensionality reduction by random mapping: Fast similarity computation for clustering. In 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227) (Vol. 1, pp. 413-418). IEEE.
Khan, M., Goskula, T., Nasiruddin, M., & Quazi, R. (2011). Comparison between k-nn and svm method for speech emotion recognition. International Journal on Computer Science and Engineering, 3(2), 607-611.
Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190.
Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24.
Landers, R. N., & Schmidt, G. B. (2016). Social Media in Employee Selection and Recruitment. Theory, Practice, and Current Challenges. Cham: Springer International Publishing AG.
Okamoto, S., & Yugami, N. (2003). Effects of domain characteristics on instance-based learning algorithms. Theoretical Computer Science, 298(1), 207-233.
Maitra, R., Peterson, A. D., & Ghosh, A. P. (2011). A systematic evaluation of different methods for initializing the k-means clustering algorithm. A separability index for clustering and classification problems with applications to cluster merging and systematic evaluation of clustering algorithms, 41.
Massart, D. L. (1983). The interpretation of analytical chemical data by the use of cluster analysis (No. 04; QD75. 4. S8, M3.).
McLachlan, G. J., Bean, R. W., & Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. Bioinformatics, 18(3), 413-422.
Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003, November). Predicting student performance: an application of data mining methods with an educational web-based system. In 33rd Annual Frontiers in Education, 2003. FIE 2003. (Vol. 1, pp. T2A-13). IEEE.
Ng, A. (2012). Clustering with the k-means algorithm. Machine Learning.
Noiva, K., Fernández, J. E., & Wescoat Jr, J. L. (2016). Cluster analysis of urban water supply and demand: Toward large-scale comparative sustainability planning. Sustainable cities and society, 27, 484-496.
Perry, J. C. (2008). School engagement among urban youth of color: Criterion pattern effects of vocational exploration and racial identity. Journal of Career Development, 34(4), 397-422.
Pratama, B. Y., & Sarno, R. (2015, November). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In 2015 International Conference on Data and Software Engineering (ICoDSE) (pp. 170-174). IEEE.
Rousseeuw, P. J., & Kaufman, L. (1990). Finding groups in data. Hoboken: Wiley Online Library.
R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1, pp. 205-233
Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., ... & Bauer, Z. (2018). The AI Index 2018 Annual Report.
Sima, C., & Dougherty, E. R. (2008). The peaking phenomenon in the presence of feature-selection. Pattern Recognition Letters, 29(11), 1667-1674.
Smyth, P. (1996, August). Clustering Using Monte Carlo Cross-Validation. In Kdd (Vol. 1, pp. 26-133).
Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In New directions in statistical physics (pp. 273-309). Springer, Berlin, Heidelberg.
Steinley, D. (2006). K-means clustering: a half-century synthesis. British Journal of Mathematical and Statistical Psychology, 59(1), 1-34.
Steinley, D. (2006). Profiling local optima in K-means clustering: developing a diagnostic technique. Psychological methods, 11(2), 178.
Tadjudin, S., & Landgrebe, D. (1998). Classification of high dimensional data with limited training samples.
Tarca, A. L., Carey, V. J., Chen, X. W., Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. PLoS computational biology, 3(6), e116.
Verleysen, M., & François, D. (2005, June). The curse of dimensionality in data mining and time series prediction. In International Work-Conference on Artificial Neural Networks (pp. 758-770). Springer, Berlin, Heidelberg.
Waller, N. G. (2019). fungible: Psychometric Functions from the Waller Lab. R package version 1.80.
Waller, N. G., Kaiser, H. A., Illian, J. B., & Manry, M. (1998). A comparison of the classification capabilities of the 1-dimensional kohonen neural network with two pratitioning and three hierarchical cluster analysis algorithms. Psychometrika, 63(1), 5-22.
Waller, N. G., Underhill, J. M., & Kaiser, H. A. (1999). A method for generating simulated plasmodes and artificial test clusters with user-defined shape, size, and orientation. Multivariate Behavioral Research, 34, 123–142.
Yang, L. R., Huang, C. F., & Wu, K. S. (2011). The association among project manager's leadership style, teamwork and project success. International journal of project management, 29(3), 258-267.


Corresponding Author can be reached at quxxx179@umn.edu