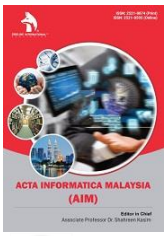
ZIBELINE INTERNATIONAL PUBLISHING

ISSN: 2521-0874 (Print)
ISSN: 2590-4043 (Online)
CODEN: AEMCDV

Acta Informatica Malaysia (AIM)

DOI: http://doi.org/10.26480/aim.02.2023.92.96



RESEARCH ARTICLE

OPTIMIZATION OF QUANTITATIVE RESEARCH METHODS IN SOCIAL SCIENCES IN THE ERA OF BIG DATA

Yirui Song[a*]

[a]*School of Statistics and Data Science, Nankai University, Weijin Road, Nankai District, Tianjin, P.R.China*
[*]*Corresponding Author Email: panyirui@126.com*

ARTICLE DETAILS

Article History:

Received 21 March 2023
Revised 24 April 2023
Accepted 27 May 2023
Available online 02 June 2023

ABSTRACT

As information technology continues to advance, the complexity of data is ever-increasing. Traditional quantitative research methods in the social sciences, such as basic visualization and traditional statistical models, are gradually becoming inadequate in meeting the demands of modern data analysis. Despite the challenges that big data presents, it also brings new opportunities - through its usage, the optimization of traditional methods can be achieved. More intricate graphing techniques such as mosaic plots, alluvial plots, slope charts, and area charts, alongside machine learning algorithms that are better adapted for big data analysis such as decision trees, random forests, and K-Means algorithm, are opening new avenues for quantitative analysis in social sciences. This will ultimately foster further development of the field by allowing new methods and ideas to emerge.

KEYWORDS

Social sciences; big data; data visualization; machine learning

1. INTRODUCTION

Quantitative research is one of the essential methods in the field of social sciences. The data analysis methods primarily used for this research are mainly two categories: basic visualization and traditional statistical models. Basic visualization includes various fundamental graphics, such as line graphs, bar charts, and pie charts, that can provide descriptive analysis of the data. On the other hand, traditional statistical models are based on basic statistical knowledge, where algorithms are pre-set by fixing a particular process to follow. Common methods include linear regression, logistic regression, and correlation analysis. The main idea of this approach is to propose a null hypothesis and alternative hypothesis and validate them by utilizing statistical knowledge (Babones, 2016).

While both of these methods can achieve good results when the data structure is relatively simple, in recent years, with the continuous development of information technology, the structure and relationships between newly generated data have become more complex. Consequently, traditional data analysis methods have become inadequate to meet the research demands. However, both data visualization methods and algorithmic models can emerge from Big Data technology, leading to further development and alteration. Big Data not only poses significant challenges but also offers new opportunities for development. (Cowls and Schroeder, 2015)

2. OPTIMIZATION OF DATA VISUALIZATION METHODS

Visualization is a technique that converts abstract data into tangible graphics to support data comprehension and perception. It is one of the essential tools in data analysis, particularly in big data analysis and has been used in various fields such as data mining, clustering, and decision-making (Shakeel et al., 2022). Through data visualization, users can comprehend data more visibly, enhancing their understanding (Qin et al., 2020). Consequently, this enables more accurate analysis and judgment.

Traditional visualization methods can only work and be effective when the data structure is relatively simple. In cases where data presentation is characterized by more dimensions, more groups, and higher data analysis requirements, these methods prove inadequate. In such scenarios, more sophisticated visualization techniques are required to analyze the data. Presently, Power BI, Tableau, among other software and websites, offer comprehensive data visualization services that require only data input without the need for coding, thereby reducing their technical difficulty and the barrier to entry. On the other hand, traditional data analysis tools such as Python, R, MATLAB, and others can achieve complex visualization demands while providing greater opportunities for personalized requirements.

3. MOSAIC PLOT

The mosaic plot evolved from the stacked bar chart (Theus, 2012; Theus and Lauer, 1999) and is typically used to display categorical data visually. It works well for showing the frequency of multiple categorical variables when the variable type is a categorical variable and the number of categories is no less than three. A mosaic plot consists of multiple rectangles nested according to different categorical variables, with the area proportional to the corresponding value (Hofmann, 2000; Hofmann, 2001). The color represents the residual of the fitted model. In R, a mosaic plot can be drawn using the mosaic() function in the vcd package (Meyer et al., 2006).

For the Titanic data set included in R, which includes variables such as “Sex”, “Cabin Class”, “Age”, and “Survival”, the multidimensional contingency table cannot capture the data contrasts effectively. In such cases, mosaic plots can be useful in exhibiting the survival status of different groups, as shown in Figure 1.

Quick Response Code	Access this article online	
	Website: www.actainformaticamalaysia.com	**DOI:** 10.26480/aim.02.2023.92.96

Cite the Article: Yirui Song (2023). Optimization of Quantitative Research Methods in Social Sciences in the Era of Big Data. *Acta Informatica Malaysia, 7(2): 92-96.*

The mosaic plot clearly shows the survival status of different groups, such as the lower survival rate of males, especially for males in second and third-class cabins. The color above the right axis line indicates that the survival rate of this group exceeded expectations when survival was independent of other variables while below means the opposite, indicating that the survival rate was lower than expected.

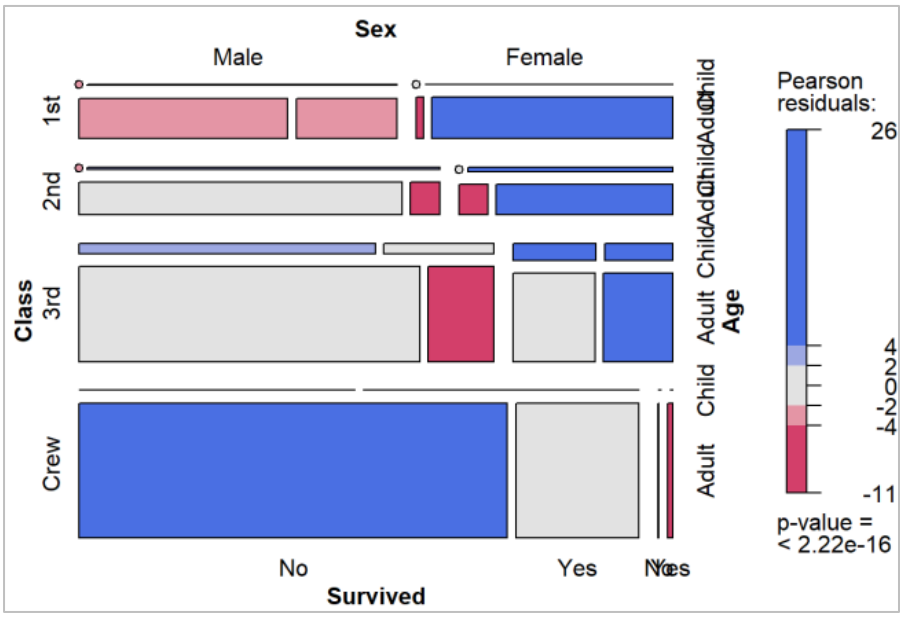


Figure 1: Mosaic plot drawn using Titanic data set

4. Alluvial Plot

The alluvial plot is a variation of a flowchart and is typically used to display the distribution of groups across categories. It works well in exhibiting the relationship between different characteristics of groups. The alluvial plot comprises multiple parallel vertical axes, with each axis representing a categorical variable segmented proportionally, similar to a stacked bar chart (Brunson, 2020). The width of the branches between the axes corresponds to the proportion. At the same time, using special colors can track a group's characteristics and its distribution in other variables. Unlike the more common Sankey diagram, the axes of an alluvial plot can be arranged arbitrarily without any order. In R, an alluvial plot can be drawn using the geom_alluvium() function in the ggalluvial package.

Again, using the Titanic data set included in R, if we want to emphasize the characteristics of survivors, we can create an alluvial plot, as shown in Figure 2.

From the alluvial plot, the distribution of survivors with different colors can be seen easily, with more females surviving and the survival rate increasing with higher cabin class.

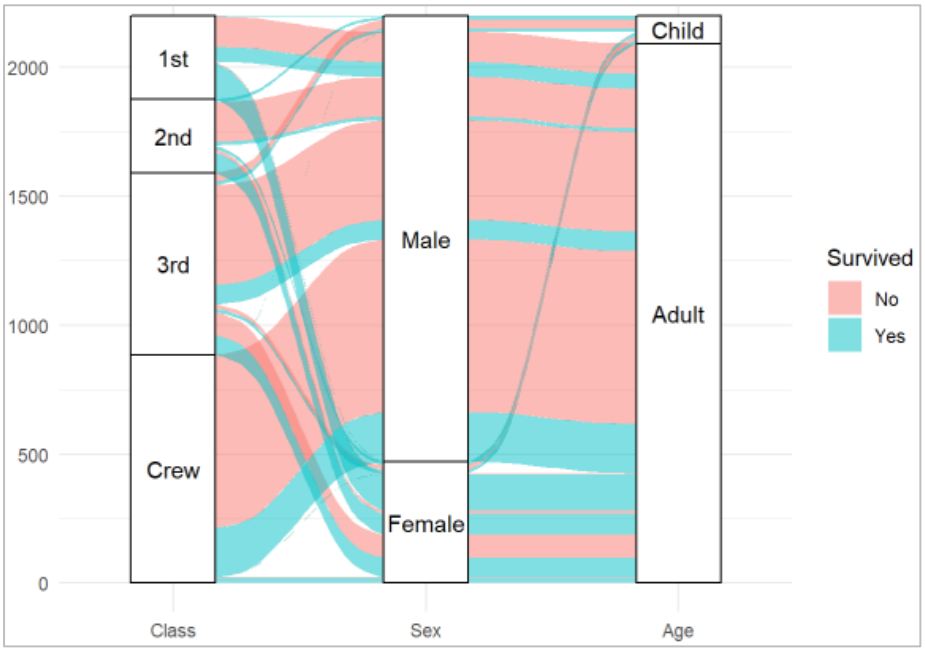


Figure 2: Alluvial plot drawn using the Titanic data set

5. Slope Chart

The slope chart is a special type of line chart that is used to display the changes in multiple groups of data between two time points or categories. The advantage is that it can clearly show the changes in each group of data between the two points and make comparisons. Meanwhile, selecting appropriate colors can highlight more interesting data or data that shows an upward or downward trend. In R, a slope chart can be drawn using the newggslopegraph() function in the CGPfunctions package.

Taking the newgdp data set included in R as an example, a grouped bar chart is drawn, but it cannot intuitively compare the relationship between the changes in trends among different groups. At this time, a slope chart should be used, and the interesting group (US in this case) can be highlighted using color, as shown in Figure 3.

6. Area Chart

The area chart, also known as the stacked area chart, is a visualization that

combines a line chart and a stacked bar chart to display the trend of multiple groups of data over time or other variables. The area chart is based on the line chart and fills the area below the line, and the lines are stacked on top of each other similarly to a stacked bar chart. The plotting of each line is based on the previous line, similar to a stacked bar chart. The area chart can intuitively display the overall trend and the contribution rate of each subgroup to the overall trend. If only the latter needs to be highlighted, a percentage stacked area chart can be used. In R, an area chart can be drawn using the geom_area() function in the ggplot package.

Taking the economics_long data set included in R as an example, when plotting the basic line chart data, it can only show the trend of two individual parts separately, as shown in Figure 4. If we want to display the overall trend and the trend of the contribution of two parts, we can use an area chart, or we can use a percentage stacked area chart to only highlight the contribution, as shown in Figure 5.

The combination of the two chats shows that the overall trend shows a fluctuating upward trend, the "unemploy" part accounts for a fluctuating downward trend in the whole, and maintains a relatively stable trend in the later period.

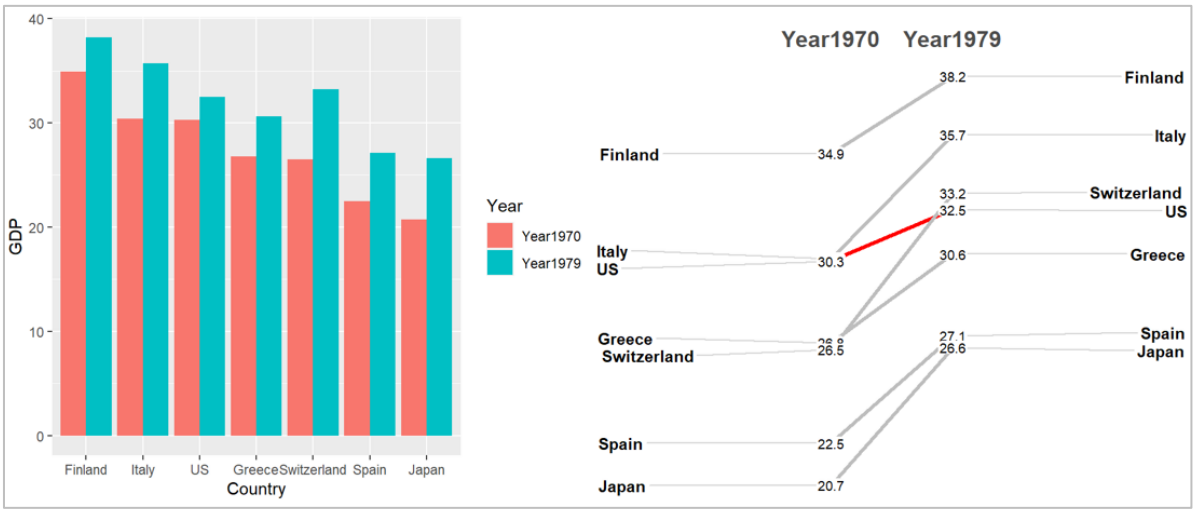


Figure 3: Comparison of a bar chart and slope chart drawn using the newgdp data set (left is a bar chart, and right is a slope chart)

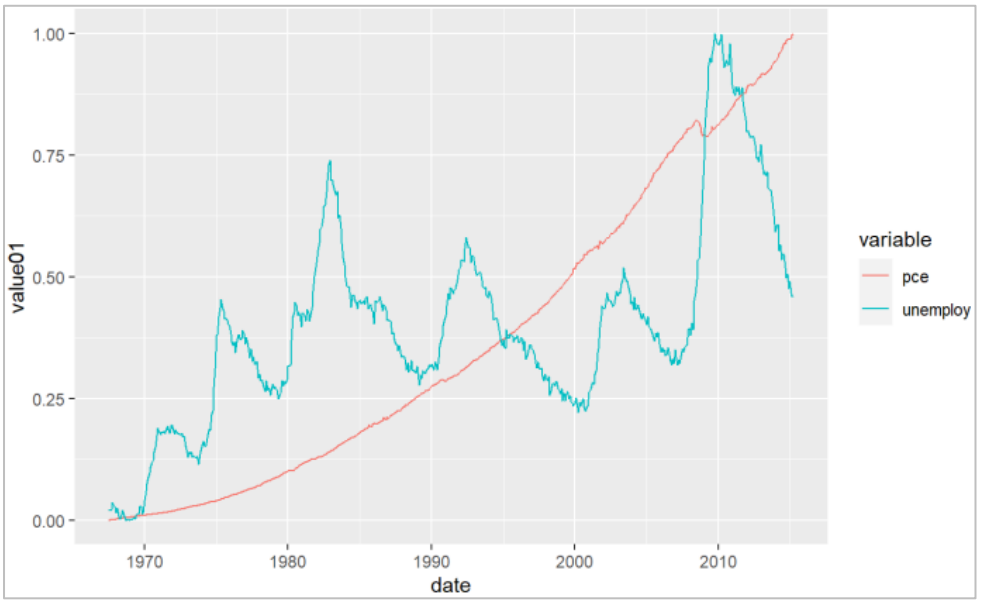


Figure 4: Line chart drawn using the economics_long data set

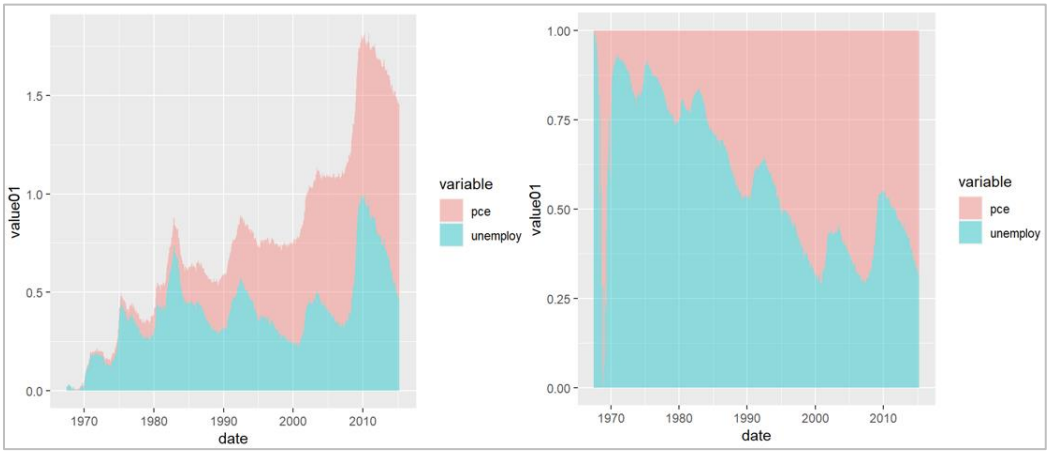


Figure 5: Area charts drawn using the economics_long data set (left is a stacked area chart, and right is a percentage stacked area chart).

7. Optimization of Algorithm Models

Machine learning algorithms can autonomously learn and improve from given data (Kersting, 2018), without the need for explicit programming, demonstrating superior performance to traditional algorithms when dealing with complex data. In recent years, machine learning algorithms have been applied in various fields such as social media analysis (Balaji et al., 2021) showcasing their powerful data analysis capabilities.

Although traditional statistical models such as linear regression and logistic regression belong to machine learning and have the advantage of high interpretability, they can only handle linear problems with relatively simple data structures. In the context of the information age, where data structures are becoming more complex and data relationships are becoming more concealed, their performance can often be unsatisfactory. On the other hand, machine learning algorithms such as decision trees, random forests, and K-Means are more suitable for big data analysis, and generally perform better in various scenarios.

8. Decision Trees and Random Forest

A decision tree is a tree-like structure consisting of nodes and directed edges. Internal nodes represent decision conditions, leaf nodes represent output objects and branches represent possible values after a decision, as shown in Figure 6. Decision trees achieve optimal segmentation and establish prediction models by dividing data into smaller subsets and pruning. This includes both classification and regression prediction models. (Loh, 2011)

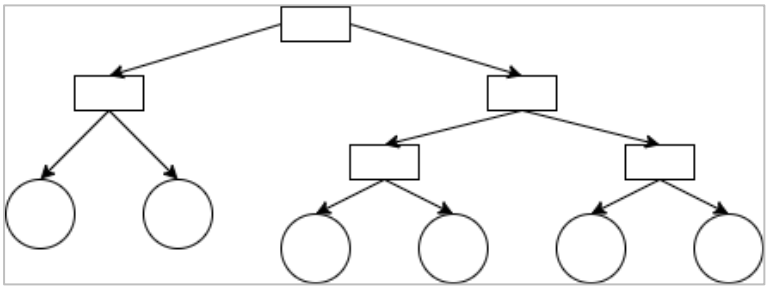


Figure 6: Schematic diagram of basic form of decision tree

In classification prediction problems, commonly used decision tree models include ID3 decision trees, C4.5 decision trees and C5.0 decision trees. Among them, C5.0 decision trees are capable of handling both discrete and continuous attributes, and have fast computational speed, with a smaller decision tree generating similar results and precision as other decision tree algorithms. Therefore, C5.0 decision trees are commonly used in practical applications. For regression prediction problems, decision trees can be divided into regression trees and model trees. Regression trees predict based on the average value when reaching the leaf node. Model trees build multiple linear regression models at the leaf node for prediction. Both methods do not require any assumptions about the data and are applicable to situations with complex data.

Random forest is an ensemble algorithm (Breiman, 2001) that combines multiple decision trees and selects the final prediction result through voting (Pal, 2005). It is regarded as one of the most popular nonlinear modeling tools available today. In comparison with decision trees, random forest has exhibited better performance when dealing with high-dimensional feature space data (Biau and Scornet, 2016). This is due to its ability to reduce the impact of overfitting and interference caused by outliers and white noise.

9. K-Means Clustering Algorithm

The clustering algorithm is a typical unsupervised learning method, that is, there is no given label in the data, and the data is classified solely through the internal relationships. Data with high similarity is classified into the same cluster (Jain et al., 1999). The application of cluster algorithms has a wide range of functions, including constructing user profiles, improving information processing efficiency, and integrating resources. Among these algorithms, K-Means is the most classical algorithm for clustering, as it demonstrates ease of understanding, high level of interpretability in results, and low algorithmic complexity (Kanungo et al, 2002; Xu and Wunsch, 2005). K-Means has opened up new avenues for quantitative research in social sciences.

10. Conclusions

The development of big data technology will continue to support quantitative research in social sciences. Big data analysis can break through the limitations of traditional methods in terms of sample size and data size, while avoiding error that can be introduced during traditional sampling methods. This provides social science researchers with higher data starting points and expanded avenues for research. Additionally, big data technology can mine data for previously overlooked details, providing new perspectives and ideas for research in social sciences.

The rapid development of information technology presents both challenges and opportunities for quantitative research in social sciences. Integrating new technology with traditional research methods and selecting appropriate data analysis methods based on different scenarios is crucial. Combining new developments in social and data sciences can further strengthen interdisciplinary cooperation, leading to common progress across multiple fields and disciplines.

References

Babones, S. 2016. Interpretive Quantitative Methods for the Social Sciences. Sociology, 50(3), 453–469, Jun, doi: 10.1177/0038038515583637.

Balaji, T.K., Annavarapu, C.S.R. and Bablani, A. 2021. Machine learning algorithms for social media analysis: A survey. Computer Science Review, 40, 100395, doi: 10.1016/j.cosrev.2021.100395

Biau, G. and Scornet, E. 2016. A random forest guided tour. TEST, 25(2), 197–227, Jun, doi: 10.1007/s11749-016-0481-7.

Breiman, L. 2001. Random Forests. Machine Learning, 45(1), 5–32, doi: 10.1023/A:1010933404324.

Brunson, J. C. 2020. "ggalluvial: Layered Grammar for Alluvial Plots," Journal of Open Source Software, 5(49), 2017, May, doi: 10.21105/joss.02017.

Cowls, J. and Schroeder, R. 2015. Causation, Correlation, and Big Data in Social Science Research. Policy & Internet, 7(4), 447–472, doi: 10.1002/poi3.100.

Hofmann, H. 2000. Exploring categorical data: interactive mosaic plots, Metrika, 51(1), 11–26, Jul, doi: 10.1007/s001840000041.

Hofmann, H. 2001. Generalized Odds Ratios for Visual Modeling. Journal of Computational and Graphical Statistics, 10(4), 628–640, Dec, doi: 10.1198/106186001317243368.

Jain, A.K., Murty, M.N., and Flynn, P. J. 1999. Data clustering: a review. ACM Comput. Surv., 31(3), 264–323, Sep, doi: 10.1145/331499.331504.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. and Wu, A. Y. 2002. An efficient k-means clustering algorithm: analysis and implementation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 881–892, Jul, doi: 10.1109/TPAMI.2002.1017616.

Kersting, K. 2018. Machine Learning and Artificial Intelligence: Two Fellow Travelers on the Quest for Intelligent Behavior in Machines. Front. Big Data, 1, 6, Nov, doi: 10.3389/fdata.2018.00006.

Loh, W.Y. 2011. Classification and regression trees. WIREs Data Mining

and Knowledge Discovery, 1(1): 14–23, doi: 10.1002/widm.8.

Meyer, D., Zeileis, A. and Hornik, K. 2006. The STRUCPLOT framework: Visualizing Multi-way contingency tables withvcd. Journal of Statistical Software, 17(3). doi:10.18637/jss.v017.i03

Pal, M. 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens., 26(1), 217–222, Jan, doi: 10.1080/01431160412331269698.

Qin, X., Luo, Y., Tang, N., and Li, G. 2020. Making data visualization more efficient and effective: a survey. The VLDB Journal, 29(1), 93–117, Jan, doi: 10.1007/s00778-019-00588-3.

Shakeel, H. M., Iram, S., Al-Aqrabi, H., Alsboui, T. and Hill, R. 2022. A Comprehensive State-of-the-Art Survey on Data Visualization Tools: Research Developments, Challenges and Future Domain Specific Visualization Framework. IEEE Access, 10, 96581–96601, doi: 10.1109/ACCESS.2022.3205115.

Theus, M. 2012. Mosaic plots. WIREs Computational Statistics, 4(2), 191–198, doi: 10.1002/wics.1192.

Theus, M. and Lauer, S. R. W. 1999. Visualizing Loglinear Models," Journal of Computational and Graphical Statistics, 8(3), 396–412, doi: 10.2307/1390864.

Xu, R. and Wunsch, D. 2005. Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645–678, May, doi: 10.1109/TNN.2005.845141.