PRINCIPAL COMPONENT AND MULTIPLE CORRESPONDENCE ANALYSIS FOR HANDLING MIXED VARIABLES IN THE SMOOTHED LOCATION MODEL

PENNY NGU AI HUONG

MASTER OF SCIENCE (STATISTICS)
UNIVERSITI UTARA MALAYSIA
2016

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

PENNY NGU AI HUONG

calon untuk Ijazah **MASTER**
(candidate for the degree of)

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"PRINCIPAL COMPONENT AND MULTIPLE CORRESPONDENCE ANALYSIS FOR HANDLING MIXED VARIABLES IN THE SMOOTHED LOCATION MODEL"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***27 Julai 2016.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: ***July 27, 2016.***

Pengerusi Viva: *(Chairman for VIVA)*	**Assoc. Prof. Dr. Mohd Kamal Mohd Nawawi**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Dr. Safwati Ibrahim**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Dr. Shamshuritawati Sharif**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Dr. Hashibah Hamid**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Dr. Nazrina Aziz**	Tandatangan *(Signature)*

Tarikh:
(Date) ***July 27, 2016***

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to :

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Isu pengelasan objek ke dalam kumpulan apabila pembolehubah yang diukur adalah campuran pembolehubah selanjar dan pembolehubah binari telah menarik perhatian ahli statistik. Antara kaedah-kaedah diskriminan dalam pengelasan, Model Lokasi Terlicin (SLM) digunakan untuk mengendalikan data yang mengandungi kedua-dua pembolehubah selanjar dan binari secara serentak. Namun, model ini adalah tidak tersaur jika data mengandungi pembolehubah binari yang besar bilangannya. Kehadiran pembolehubah binary yang besar akan mewujudkan sel multinomial yang banyak, yang akhirnya mengakibatkan wujudnya banyak bilangan sel kosong. Kajian lepas telah menunjukkan bahawa kewujudan sel kosong yang banyak berupaya menjejaskan prestasi model lokasi terlicin yang dibina. Dalam usaha untuk mengatasi masalah sel kosong yang banyak disebabkan oleh banyak pembolehubah terukur (terutamanya binari), kajian ini mencadangkan empat model SLM yang baharu melalui penggabungan SLM sedia ada dengan Analisis Komponen Utama (PCA) dan empat jenis analisis kesepadanan berganda (MCA). PCA digunakan untuk menguruskan bilangan pembolehubah selanjar yang besar manakala MCA digunakan untuk mengendalikan pembolehubah binari yang banyak. Prestasi empat model yang dicadangkan, SLM+PCA+MCA Indikator, SLM+PCA+MCA Burt, SLM+PCA+Analisis Kesepadanan Tercantum (JCA), dan SLM+PCA+MCA Terlaras dibandingkan berdasarkan kadar kesilapan pengelasan. Keputusan kajian simulasi menunjukkan model SLM+PCA+JCA berprestasi terbaik dalam semua keadaan yang diuji kerana ia berjaya mengekstrak jumlah komponen binari terkecil dan masa pelaksanaannya paling singkat. Siasatan pada set data sebenar barah payudara penuh juga menunjukkan bahawa model ini menghasilkan kadar kesilapan pengelasan terendah. Kadar kesilapan pengelasan terendah yang berikutnya diperolehi oleh SLM+PCA+MCA Terlaras diikuti SLM+PCA+MCA Burt dan SLM+PCA+MCA Indikator. Walaupun model SLM+PCA+MCA Indikator memberi prestasi yang paling lemah tetapi model ini masih lebih baik daripada beberapa kaedah pengelasan sedia ada. Keseluruhannya, model-model lokasi terlicin yang dibina boleh dianggap sebagai kaedah alternatif untuk tugas-tugas pengelasan dalam mengendalikan pembolehubah campuran yang banyak, terutamanya pembolehubah binari.

Kata Kunci: Model Lokasi Terlicin, Analisis Komponen Utama, Analisis Kesepadanan Berganda, Pembolehubah binary besar, Pembolehubah campuran

Abstract

The issue of classifying objects into groups when the measured variables are mixtures of continuous and binary variables has attracted the attention of statisticians. Among the discriminant methods in classification, Smoothed Location Model (SLM) is used to handle data that contains both continuous and binary variables simultaneously. However, this model is infeasible if the data is having a large number of binary variables. The presence of huge binary variables will create numerous multinomial cells that will later cause the occurrence of large number of empty cells. Past studies have shown that the occurrence of many empty cells affected the performance of the constructed smoothed location model. In order to overcome the problem of many empty cells due to large number of measured variables (mainly binary), this study proposes four new SLMs by combining the existing SLM with Principal Component Analysis (PCA) and four types of Multiple Correspondence Analysis (MCA). PCA is used to handle large continuous variables whereas MCA is used to deal with huge binary variables. The performance of the four proposed models, SLM+PCA+Indicator MCA, SLM+PCA+Burt MCA, SLM+PCA+Joint Correspondence Analysis (JCA), and SLM+PCA+Adjusted MCA are compared based on the misclassification rate. Results of a simulation study show that SLM+PCA+JCA model performs the best in all tested conditions since it successfully extracted the smallest amount of binary components and executed with the shortest computational time. Investigations on a real data set of full breast cancer also showed that this model produces the lowest misclassification rate. The next lowest misclassification rate is obtained by SLM+PCA+Adjusted MCA followed by SLM+PCA+Burt MCA and SLM+PCA+Indicator MCA models. Although SLM+PCA+Indicator MCA model gives the poorest performance but it is still better than a few existing classification methods. Overall, the developed smoothed location models can be considered as alternative methods for classification tasks in handling large number of mixed variables, mainly the binary.

Keywords: Smoothed Location Model, Principal Component Analysis, Multiple Correspondence Analysis, Large binary variables, Mixed variables

Acknowledgement

I would have never been able to finish my dissertation without the guidance of my supervisors, help from friends and full support from my precious family.

Firstly, I would like to express my deepest gratitude to my main supervisor, Dr. Hashibah binti Hamid, for her excellent supervision, care and patience in guiding me throughout the research. I am very thankful for her generosity in sharing precious knowledge with me especially in computer programming that enables me to finish my data analysis on time. I would also like to express my appreciation to my second supervisor, Dr. Nazrina binti Aziz for guiding my research and helping to develop the structure of the thesis especially in my writing.

I would like to thank every staff at the School of Quantitative Sciences and Awang Had Salleh Graduate School of Universiti Utara Malaysia for their valuable support and assistance.

This study would not have been possible without financial aid. I would like to thank the government of Malaysia for the financial support through Mybrain 15 and also University Utara Malaysia for the financial support and facilities provided to assist me in completing my studies.

I would also like to thank all my friends for their direct and indirect help in completing my research.

Last but not least, I am very grateful to my beloved family especially my parents, Mr. Ngu Kuong Hui and Mrs. Chiam Leh Choo for their endless love and continuous support throughout the two years of my studies. I also wish to express my appreciation to my brothers, Mr. William Ngu and Mr. Mathew Ngu for their encouragement towards the completion of this research.

To all of them, I dedicate this work.

Table of Contents

Permission to Use i
Abstrak ii
Abstract iii
Acknowledgement iv
Table of Contents v
List of Tables viii
List of Figures ix
List of Publications x

CHAPTER ONE INTRODUCTION 1
1.1 Background 1
1.2 Problem Statement 8
1.3 Research Objectives 10
1.4 Research Contributions 11
1.5 Research Scopes 11
1.6 Thesis Organization 12

CHAPTER TWO LITERATURE REVIEW 13
2.1 Introduction 13
2.2 The Evolution of Location Model 13
2.2.1 Smoothed Location Model 15
2.2.2 Non-parametric Smoothing Estimation 15
2.3 Data Reduction Techniques for High Dimensional Data 18
2.3.1 Variable Selection 19
2.3.2 Variable Extraction 19
2.4 Principal Component Analysis (PCA) 21
2.4.1 Principal Component Scores (PCs) 23
2.4.2 Determining the Number of PCs to Retain 25
2.5 Multiple Correspondence Analysis (MCA) 25
2.5.1 Types of MCA 27
2.5.2 Determining the Number of Components to Retain 31

2.6 Model Evaluation 31

CHAPTER THREE RESEARCH METHODOLOGY 34

3.1 Introduction 34

3.2 Procedure Design for Classification 34

3.3 Algorithms for Variable Extraction 36

3.4 Construction of the Smoothed Location Model 41

3.4.1 Classification Model and Nonparametric Smoothing Estimation 41

3.4.2 Weight for Smoothing Parameter 43

3.5 Monte Carlo Study 47

3.5.1 Generation of Multivariate Data with Mixed Variables 47

3.5.2 Generation of Normal Mixed Data 48

3.6 Model Evaluation 49

3.7 Application of the Proposed Models on Real Data Set 50

CHAPTER FOUR FINDINGS AND DISCUSSION 51

4.1 Introduction 51

4.2 Classification Performance of the Constructed Smoothed Location Model 51

4.2.1 Results of the Constructed Smoothed Location Models with PCA and Indicator MCA 52

4.2.2 Results of the Constructed Smoothed Location Models with PCA and Burt MCA 56

4.2.3 Results of the Constructed Smoothed Location Models with PCA and JCA 57

4.2.4 Results of the Constructed Smoothed Location Models with PCA and Adjusted MCA 58

4.2.5 Comparison of All Results based on Proposed Models 59

4.3 The Average Execution Time 64

4.4 Application of the Proposed Models on Real Data Set 66

4.4.1 Full Breast Cancer Data 66

4.4.2 Comparison among Classification Methods 68

CHAPTER FIVE CONCLUSION AND FUTURE WORK 71

5.1 Introduction 71

5.2 Discussion and Conclusion 71
5.3 Future Work Direction 74
REFERENCES 75

List of Tables

Table 3.1 Data Conditions and Data Labelling 49
Table 4.1 Performance of the Constructed Smoothed Location Models with PCA and Indicator MCA for All Simulated Data Sets 52
Table 4.2 Performance of the Constructed Smoothed Location Models with PCA and Burt MCA for All Simulated Data Sets 58
Table 4.3 Performance of the Constructed Smoothed Location Models with PCA and JCA for All Simulated Data Sets 57
Table 4.4 Performance of the Constructed Smoothed Location Models with PCA and Adjusted MCA for All Simulated Data Sets 58
Table 4.5 Performance of Constructed Smoothed Location Models with PCA and All Four Types of MCA for b=20 under All Sample Sizes Considered 59
Table 4.6 Performance of Constructed Smoothed Location Models with PCA and All Four Types of MCA under n=120 for all Binary Sizes Considered 63
Table 4.7 Empirical Results of the Constructed Smoothed Location Models with PCA and All For Types of MCA for b=20 under n=60 64
Table 4.8 Average Computational Time (in seconds) for the Whole Estimation Process of Smoothed Location Model with PCA and All Four Types of MCA 65
Table 4.9 Results of Full Breast Cancer Data for Eight Classification Methods 69

List of Figures

Figure 4.1 Performance of Constructed Smoothed Location Model based on KL Distance 55
Figure 4.2 Performance of the Constructed Smoothed Location Model based on Number of Binary Retained 55

List of Publications

Ngu, P.A.H., Hamid, H. & Aziz, N. (2015). *Multiple Correspondence Analysis for Handling Large Binary Variables in Smoothed Location Model.* Paper Presented at 2nd Innovation and Analytics Conference & Exhibition (IACE 2015), 29 September - 1 October 2015, Alor Setar, Kedah, Malaysia

Ngu, P.A.H., Hamid, H. & Aziz, N. (2015). *The Performance of Smoothed Location Model with PCA+Indicator MCA and PCA+Adjusted MCA.* Paper Presented at 4th International Conference on Quantitative Science and Its Applications (ICOQSIA 2016), 16 August - 18 August 2016, Putrajaya, Malaysia

CHAPTER ONE
INTRODUCTION

1.1 Background

Classification is a procedure of grouping objects or individual into their groups according to some common characteristics (Hunter, 2009). Classification tasks are found in different areas of studies such as in medical where it involves classification of breast tumors, in financial where it involves classification of bankruptcy and classification of students' performance based on their grades in education. (Veer et al., 2002; Hauser & Booth, 2011). These classifications are called standard classification while standard statistical classifications represent a subset for statistical use (Hoffman, 1999).

By using statistics, classification can be done in many ways. One of them is through discriminant analysis. Discriminant analysis is a statistical analysis method used to classify an object into one of several populations (Lachenbruch, 1975; Hand, 1981; Pyryt, 2004; Jombart, Devillard, & Balloux, 2010). There are many successful applications of discriminant analysis based on variables that have been collected and used in various fields such as economy, environmental sciences and humanistic as well as social behavior and geographical ecology (William, 1983; Chanda & Murthy, 2008; Taniguchi, Hirukawa, & Tamaki, 2008; Soni & Shrivastave, 2010).

Discriminant analysis has been used for classification not only on single type of variables but also mixture type (Krzanowski, 1980; Knoke, 1982; Daudin, 1986). Data with single type of variable refers to the data set containing only the continuous

The contents of the thesis is for internal user only

REFERENCES

Abdi, H., & Valentin, D. (2007). *Multiple Correspondence Analysis.* Encyclopedia of Measurement and Statistics: USA: Sage.

Adler, N. & Golany, B. (2002). Including Principal Component Weight to Improve Discrimination in Data Envelopment Analysis. *Journal of the Operational Research Society*, *53*, 985–991.

Afifi, A. A., & Elashoff, R. M. (1969). Multivariate Two Sample Tests with Dichotomous and Continuous Variables: The Location Model. *The Annals of Mathematical Statistics*, *40*(1), 290–298.

Akturk, D., Gun, S. & Kumuk, T. (2007). Multiple Correspondence Analysis Technique Used in Analyzing the Categorical Data in Social Sciences. *Journal of Applied Sciences*, *7*(4), 585–588.

Anderson, T. W. (1963). Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics*, *34*(1), 122–148.

Asparoukhov, O. & Krazanowski, W. J. (2000). Non-parametric Smoothing of the Location Model in Mixed Variables Discrimination. *Statistics and Computing*, 10(4), 289-297.

Ayele, D., Zewotir, T. & Mwambi, H. (2013). Multiple Correspondence Analysis as a Tool for Analysis of Large Health Surveys in African Settings. *African Health Sciences*, *14*(4), 1036–1045.

Bakar, Z. A., Mohemad, R., Ahmad, A., & Deris, M. M. (2006). A Comparative Study for Outlier Detection Techniques in Data Mining. *IEEE Conference on Cybernetics and Intelligent Systems*, 1–6.

Bar-Hen, A. (2002).). Generalized Principal Component Analysis of Continuous and Discrete Variables. *Interstat Statistics*, *8*(6), 1–26.

Baxter, M. J. (1995). Standardization and Transformation in Principal Component Analysis, With Applications to Archaemotry. *Journal of the Royal Statistical Society. Series C*, *44*(4), 513–527.

Belhumeur, P. N., Hespanha, J. P. & Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *19*(7), 711–720.

Bellman, R. (1961). *Adaptive Control Process: A Guided Tour*. Princeton: Princeton University Press.

Benzecri, J. P. (1992). *Correspondence Analysis Handbook*. New York, NY: Marcel Dekker, Inc.

Bishop, C. M. (1995). *Neural Network for Pattern Recognition*. New York, NY: Oxford University Press.

Bittencourt, H. R. & Clarke, R. T. (2003). Logistic Discriminant between Classes with Nearly Equal Spectral Response in High Dimensionality. In *Proceedings of the IEEE International of Geoscience and Remote Sensing Symposiums, 6*, 21-25 July 2003, Toulouse, France (pp.3748-3750). Piscataway, NJ: IEEE Operations Center.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*, 123–140.

Buhlmann. P. & Geer, S. (2011). *Statistics for High Dimensional Data: Methods, Theory and Applications*. Berlin, Heidelberg: Springer-Verlag.

Buttrey, S. (1998). Nearest-neighbour classification with categorical variables. *Computational Statistics and Data Analysis*, *28*, 157–169.

Camiz, S. & Gomes, G. C. (2013). *Joint Correspondence Analysis versus Multiple Correspondence Analysis: A Solution to an Undetected Problem. In A. Giusti, G. Ritter & M. Vichi (Eds), Classification and Data Mining: Studies in Classification, Data Analysis and Knowledge Organization*. (pp. 11-19). Berlin, Heidelberg: Springer-Verlag.

Cateni, S., Vannucci, M. & Colla, V. (2013). *Multivariate Analysis in Management, Engineering and the Sciences*. *Classification and Ordination Methods as a Tool for Analyzing of Plant Communities.* Retrieved from http://www.intechopen.com/books/multivariate-analysis-in-management-engineering-and-the-sciences/classification-and-ordination-methods-as-a-tool-for-analyzing-of-plant-communities#SEC6

Chanda, B. & Murthy, C. A. (2008). *Advance in Intelligent Information Processing*. Singapore: World Science Publishing.

Chang, P. C. & Afifi, A. A. (1974). Classification Based on Dichotomous and Continuous Variables. *Journal of the American Statistical Association*, *69*(346), 336–339.

Choi, S. C. (1986). Discrimination and Classificaton: Overview. *Computers and*

Mathematics with Applications, *12A*(2), 173–177.

Chou, Y., & Wang, W. (2010). Checking Dimensionality in Item Response Models With Principal Component Analysis on Standardized Residuals. *Educational and Psychological Measurement*, *70*(5), 717–731.

Costa, P. S., Santos, N. C., Cunha, P., Cotter, J., & Sousa, N. (2013). The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. *Journal of Aging Research*, 1–12. Retrieved from http://dx.doi.org/10.1155/2013/302163

Cox, D. R. (1966). *Some procedures associated with the logistic qualitative response curve. In Research Papers in Statistics: Festschrift for J. Neyman, Ed. F. N. David*. New York, NY: Wiley.

Cox, T. F., & Pearce, K. F. (1997). A robust logistic discrimination model. *Statistics and Computing*, *7*(3), 155–161.

D'Enza, A. I. & Greenacre, M. J. (2012). Multiple Correspondence Analysis for the Quantification and Visualization of Large Categorical Data Sets. In A. Di Ciaccio, M. Coli & J. M. A. Ibanez (Eds.), *Advanced Statistical Methods for the Analysis of Large Data-Sets: Studies in Theoretical and Applied Statistics* (pp. 453-463). Berlin, Heideberg: Springer-Verlag.

Das, K. (2007). *Feature Extraction and Classification for Large-scale Data Analysis (Unpublished doctoral dissertation).* Department of Electrical and Computer Engineering, University of California.

Das, K., Meyer, J. & Nenadic, Z (2006). Analysis of Large-Scale Brain Data for Brain-Computer Interfaces. In *Proceeding of the 28th IEEE Annual International Conference of Engineering in Medicine and Biology Society*, 30 August-3 September 2006, New York (pp. 5731-5734). Retrieved from http://cbmspc.eng.uci.edu/PUBLICATIONS/zn:06b.pdf.

Daudin, J. J. (1986). Selection of Variables in Mixed-Variable Discriminant Analysis. *Biometrics*, *42*(3), 473–481.

Day, N. E. & Kerridge, D. F. (1967). A General Maximum Likelihood Discriminant. *Biometrics*, *23*, 313–323.

de Leeuw, J. (2006). Here's Looking at Multivariable. In J. Blasius & M. J. Greenacre (Eds.), *Visualization of Categorical Data* (pp. 1-11). San Diego: Academic Press.

Deng, H., Jin, L., Zhen, L., & Huang, J. (2005). A New Facial Expression Recognition Method on Local Gabor Filter Bank and PCA plus LDA. *Journal of Information Technology*, *11*(11), 86–96.

Desikachar, P. & Viswanathan, B. (2011). *Patterns of Labour Market Insecurity in Rural India: A Multidimensional and Multivariate Analysis* (No. Working Paper No. 62/2011). Madras School of Economics.

Doey, L., & Kurta, J. (2011). Correspondence analysis applied to psychological research. *Tutorials in Quantitative Methods for Psychology*, *7*(1), 5–14.

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross Validation. *Journal of the American Statistical Association*, *78*, 316–331.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*(1), 179–188.

Fix, E. & Hodges, J. L. (1951). *Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties* (No. Report No. 51-4). Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf

Garcia, T., & Grande, I. (2003). A model for the valuation of farmland in Spain: The case for the use of multivariate analysis. *Journal of Property Investment & Finance*, *21*(2), 136–153.

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of American Statistical Association*, *70*, 320–328.

Gervini, D. & Rousson, V. (2004). Criteria for Evaluating Dimension-Reducing Components for Multivariate Data. *The American Statistician*, *58*(1), 72–76.

Ghosh, A. & Barman, S. (2013). Prediction of Prostate Cancer Cells Based on Principal Component Analysis Techinque. *Procedia Technology*, *10*, 37–44.

Ghosh, A. (2011). Forecasting BSE Sensex under Optimal Conditions: An Investigation Post Factor Analysis. *Journal of Business Studies Quarterly*, *3*(2), 52–73.

Giri, N. C. (2004). *Multivariate Statistical Analysis*. New York, NY: Marcel Dekker Inc.

Glynn, D. (2012). Correspondence Analysis: Exploring Data and Identifying Patterns. In. D. Glynn & J. Robinson (Eds), *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics* (pp. 133-179). Amsterdam:

John Benjamins.

Gnanadesikan, R., Roger, K., Breiman, L., Dunn, O. J., Friedman, J. H., Fu, K. S., Hartigan, J. A., Kettenring, J. R., Lachenbruch, P. A., Olshen, R. A. & Rohlf, F. J. (1989). Discriminant analysis and clustering. Panel on Discriminant Analysis, Classification and Clustering. *Statistical Science*, *4*(1), 34–69.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York, NY: John Wiley & Sons, Inc.

Gottumukkal, R. & Asari, K. (2004). An Improved Face Recognition Technique Based on Modular PCA Approach. *Patern Recognit Lett*, *25*, 429–436.

Green, P. E., Krieger, A. M. & Carroll, J. D. (1987). Multidimensional Scaling: S Complementary Approach. *Journal of Advertisement Research*, 21–27.

Greenacre, M. J. & Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods*. London: Taylor and Francis Group.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M. (2007). *Correspondence Analysis in Practice (2nd ed.)*. Boco Raton: Chapman & Hall.

Greenacre, M. J. (2006). *Typing Up the Loose Ends in Simple, Multiple and Joint Correspondence Analysis. In A. Rizzi & M. Vichi (Eds.),*. Berlin, Heidelberg: Physica-Verlag.

Griebel, m. & Hullmann, A. (2013). *Dimensionality Reduction of High Dimensional Data with a Non-Linear Principal Component Aligned Generative Topographic Mapping*. Institute for Numerical Simulation.

Guttmann, L. (1941). *The Quantification of a Class of Attributes: A Theory and Method of Scale Construction. In P. Horst, P. Wallin & L. Gutmann (Eds.)*. New York, NY: Social Science Research Council.

Hamid, H. & Mahat, N. I. (2013). Using Principal Component Analysis to extract mixed variables for smoothed location model. *Far East Journal of Mathematical Sciences (FJMS)*, *80*(1), 33–54.

Hamid, H. (2014). *Integrated Smoothed Location Model and Data Reduction Approaches for Multi Variables Classification*. Doctor of Philosophy. Universiti Utara Malaysia.

Hand, D. J. (1981). *Discrimination and Classification*. Chichester: John Wiley & Sons.

Hauser, R. P., & Booth, D. (2011). Predicting Bankruptcy with Robust Logistic Regression. *Journal of Data Science, 9*, 565–584.

Hinneburg, A., & Keim, D. A. (1999). Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. *International Conference on Very Large Database*, *1*, 506–517.

Hirschfeld, H. O. (1935). A Connection Between Correlation and Contingency. *Proceedings of the Cambridge Philosophical Society, 31*, 520–524.

Hirsh, O., Bosner, S., Hullermeier, E., Senge, R., Dembczynski, K. & D.-, & Banzhoff, N. (2011). Multivariate Modeling to Identify Patterns in Clinical Data: The Example of Chest Pain. *BMC Medical Research Methodology, 11*, 155–164.

Hoffman, D. L., & Batra, R. (1991). Viewer Response to Programs: Dimensionality and Concurrent Behavior. *Journal of Advertising Research, 23*, 45–46.

Hoffman, D. L., & Franke, G. R. (1986). Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Research, 23*, 213–227.

Hoffman, E. (1999). Standard Statistical Classification: Basic Principles. *Statistical Commission, 8*, 1–30.

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two-group classification methods. *Educational and Psychological Measurement, 71*(5), 870–901.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology, 24*, 417–441.

Hunter, E. J. (2009). *Classification Made Simple: An Introduction to Knowledge Organisation and Informative Retrieval (3rd ed)*. England: Ashgate Publishing Company.

Hyvarinen, A. & Oja, E. (2000). ndependent Component Analysis: Algorithms and Applications. *Neural Network, 13*(4-5), 411–430.

Jackson, D. A. (1993). Stopping Rules in Principal Components Analysis : A Comparison of Heuristical and Statistical Approaches. *Ecology, 74*(8), 2204–

2214.

Jackson, J. E. (1991). Stopping Rules in Principal Components Analysis: A Camparison Heuristical and Statistical Approach. *Ecology*, *74*(8), 2204–2214.

Jeffers, J. N. R. (1967). Two Case Studies in the Application of Principal Component Analysis. *Journal of the Royal Statistical Society*, *16*(3), 225–236.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York, NY: Springer-Verlag.

Jolliffe, I. T. (2002). *Principal Component Analysis (2nd ed.)*. New York, NY: Springer-Verlag.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(94), 1–15.

Kaiser, H. F. (1961). A note on Gutman's Lower Bound for the Number of Common Factors. *British Journal of Mathematical and Statistical Psycology*, *14*, 1–2.

Kant, I. (1968). *The Critique of Pure Reason*. New York, NY: Hackett Publishing Co.

Kemsley, E. K. (1996). Discriminant Analysis of High-Dimensional Data: A Comparison of Principal Component Analysis and Partial Least Squares Data Reduction Methods. *Chemometrics and Intelligent Systems*, *33*, 47–61.

Knoke, J. D. (1982). Discriminant Analysis with Discrete and Continuous Variables. *Biometrics*, *38*(1), 191–200.

Kolenikov, S., & Angeles, G. (2009). Socioeconomic Status Measurement With Discrete Proxy Variables: Is Principal Component Analysis a Reliable Answer? *Review of Income and Wealth*, *55*(1), 128–165.

Krzanowski, W. J. (1975). Discrimination and Classification Using both Binary and Continuous Variables. *Journal of American Statistical Association*, *70*, 782–790.

Krzanowski, W. J. (1977). The Performance of Fisher's Linear Discriminant Function under Non-optimal Conditions. *Technometrics*, *19*, 191–200.

Krzanowski, W. J. (1979). Some Linear Transformation for Mixtures of Binary and

Continuous Variables with Particular Reference to Linear Discriminant Analysis. *Biometrika*, *66*(1), 33–39.

Krzanowski, W. J. (1980). Mixtures of Continuous and Categorical Variables in Discriminant Analysis. *Biometrics, 36,* 493-499.

Krzanowski, W. J. (1982). Mixtures of Continuous and Categorical Variables in Discriminant Analysis : A Hypothesis- Testing Approach. *Biometric, 38*, 991–1002.

Krzanowski, W. J. (1983a). Stepwise Location Model Choice in Mixed Variables in Discriminant Analysis. *Applied Statistics*, *32*(3), 260–266.

Krzanowski, W. J. (1993). The Location Model for Mixtures of Categorical and Continuous Variables. *Journal of Classification*, *10*, 25–49.

Krzanowski, W. J. (1995). Selection of Variables, and Assessment of Their Performance, in Mixed Variable Discriminant Analysis. *Computational Statistics and Data Analysis*, *19*, 419-431.

Kuo, C.F, Syu, S. S, Lin, C. H. & Peng, K. C. (2012). Application of Principal Component Analysis and Gray Relational Method in the Optimization of the Met Spinning Process Using the Cooling Air System. *Textile Research Journal*, *83*(4), 371–380.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of Error Rates in Discriminant Analysis. *Technometrics, 10*, 1–11.

Lachenbruch, P. A., Sneeringer, C., & Revo, L. T. (1975). Robustness of the Linear and Quadratic Discrimination Function to Certain Types of Nonnormality. *Communications in Statistics*, *1*, 39–56.

Lachenbruch, P. A. (1975). *Discriminant Analysis*. New York, NY: Hafner.

Lee, S., Zou, F., & Wright, F. A. (2010). Convergence and Prediction of Principal Component Scores in High-Dimensional Settings. *Annals of Statistics*, *38*(6), 3605–3629.

Leon, A. R., Soo, A. & Williamson, T. (2011). Classification with Discrete and Continuous Variables via General Mixed-Data Models. *Journal of Applied Statistics, 38*(5), 1021–1032.

Li, X. & Xu, R. (2009). *High Dimensional Data Analysis in Cancer Research*. New York, NY: Springer-Verlag.

Li, Q. (2006). *An Integrated Framework of Feature Selection and Extraction for Appearance-based Recognition (Unpublished doctoral dissertation).* University of Delaware Newark, DE, USA.

Loslever, P. (2009). Using Multiple Correspondence Analysis with Membership Values when the System Study Yields Miscellaneous Datasets. *Cybernetics and Systems*, *40*(7), 633–652.

Mahat, N. I., Krzanowski, W. J., & Hernandez, A. (2007). Variable selection in discriminant analysis based on the location model for mixed variables. *Advances in Data Analysis and Classification*, *1*, 105–122.

Mahat, N. I., Krzanowski, W. J., & Hernandez, A. (2009). Strategies for Non-Parametric Smoothing of the Location Model in Mixed-Variable Discriminant Analysis. Modern Applied Science, 3(1), 151-163.

Massey, W. F. (1965). Principal Components Regression in Exploratory Statistical Research. *Journal of American Statistical Association*, *60*, 234–246.

McCulloch, W. S. & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5(4), 115–133.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York, NY: John Wiley & Sons, Inc.

Messaoud R. B., Boussaid, O. & Rabaseda, S. L. (2007). A Multiple Correspondence Analysis to Organize Data Cubes. In O. Vasilecas, J. Eder & A. Caplinskas (Eds.). In *Databases and Information System IV: Frontier in Artificial Intelligence and Applications* (pp. 133–146). Amsterdam: IOS Press.

Meulman, J.J., van Der Kooji, A. J. & Heiser, W. J. (2004). *Principal Component Analysis with Nonlinear Optimal Scaling Transformation for Ordinal and Nominal data. In D. Kaplan (Eds.), The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks: Sage.

Mitchell, T. M. (1997). *Machine Learning*. New York, NY: McGraw-Hill.

Nenadic, O. & Greenacre, M. (2007). Correspondence Analysis in R, with Two and Three Dimensional Graphics: The ca Package. *Journal of Statistical Software*, *20*(3), 1–13.

Olkin, I. & Tate, R. F. (1961). Multivariate Correlation Models with Discrete and Continuous Variables. *The Annals of Mathematical Statistics*, *32*, 448–465.

Panea, B., Casasús, I., Blanco, M., & Joy, M. (2009). The use of correspondence analysis in the study of beef quality: a case study. *Spanish Journal of Agricultural Research*, *7*(4), 876–885.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, *6*, 559–572.

Peter, G. M., Joop, T. & Charles, O. (1997). Multiple Correspondence Analysis as A Tool for Quantification or Classification of Career Data. *Journal of Educational and Behavioral Statistics*, *22*(4), 447–477.

Pyryt, M. C. (2004). Pegnato Revisited : Using Discriminant Analysis to Identify Gifted Children. *Psychology Science*, *46*(3), 342–347.

Quinn, G. P. & Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. New York, NY: Cambridge University Press.

Rao, C. R. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya A*, *26*, 329–358.

Reise, S. P., Waller, N. G. & Comrey, A. L. (2000). Factor Analysis and Scale Revision. *Psychological Assessment*, *12*, 287-297.

Rencher, A. C. (2002). *Methods of Multivariate Analysis: Wiley Series in Probability and Statistics (2nd ed.).* New York, NY: John Wiley & Sons, Inc.

Saporta, G., & Niang, N. (2006). *). Correspondence Analysis and Classification. In: Michael Greenacre & Jorg Blasius (eds.) Multiple Correspondence Analysis and Related Methods.* Boca Raton: Chapman & Hall/CRC.

Schürks, M., Buring, J. E., & Kurth, T. (2011). Migraine features, associated symptoms and triggers: a principal component analysis in the Women's Health Study. *International Headache Society*, *31*(7), 861–869.

Simon, R., Radmacher, M. D., Dobbin, K. & McShane, L. M. (2003). Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *Journal of National Cancer Institute*, *95*(1), 14–18.

Siswadi, Muslim, A. & Bakhtiar, T. (2012). Variable Selection Using Principal Component and Procrustes Analyses and its Application in Educational Data. *Journal of Asian Scientific Research*, *2*(12), 856–865.

Smith, C. A. B. (1947). Some Examples of Discrimination. *Annals of Eugenics1*, *18*, 272–283.

Soni, S. & Shrivastave, S. (2010). Classification of Indian Stock Market Data Using Machine Learning Algorithms. *International Journal on Computer Science and Engineering, 2*(9), 2942–2946.

Stern, H. S. (1996). Neural Networks in Applied Statistics. *Technometrics, 38*, 217–225.

Taniguchi, M., Hirukawa, J. & Tamaki, K. (2008). *Optimal Statistical Inference in Financial Engineering*. Chapman: Boca Raton.

Tenenhaus, M. & Young, F. W. (1985). An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multivariate Data. *Psychometrika, 50*(1), 91–119.

Tian, Y., Guo, P. & Lyu, M. R. (2005). Comparative Studies on Feature Extraction Methods for Multispectral Remote Sensing Image Classification. In *In Proceedings of the International Conference on Systems, Man and Cybernetics* (pp. 1275–1279). Berlin, Heidelberg: Springer-Verlag.

Turk, M. A. & Pentland, A. P. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience, 3*(1), 71–86.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer-Verlag.

Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., … Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530–536.

Vlachonikolis, I. G., & Marriott, F. H. C. (1982). Discrimination with mixed binary and continuous data. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 31*(1), 23–31.

Wahl, P. W. & Kronmal, R. A. (1977). Discriminant Functions when Covariances are Unequal and Sample Sizes are Moderate. *Biometrics, 33*, 479–484.

Wakaki, H. (1990). Comparison of Linear and Quadratic Discriminant Funcitons. *Biometrika, 77*, 227–229.

Wang, X. & Tang, X. (2004). Experimental study on multiple LDA classifier combination for high dimensional data classification. In *Proceedings of the 5th International Workshop on Multiple Classifier Systems* (Vol. 3077, pp. 344–353). Cagliari, Italy: Springer-Verlag.

William, B. K. (1983). Some Observations on the Use of Discrimnant Analysis in Ecology. *Ecology Society of America*, *64*(8), 1283–1291.

Wold, H. (1966). Nonlinear Estimation by Iterative Least Squares Procedures. In F. N. David (Eds.). In *Research Papers in Statistics* (pp. 411–444). New York, NY: John Wiley & Sons, Inc.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1-3), 37–52.

Yang, C., Peng, H. & Wang, J. (2008). A New Feature Extraction Approach Based on Sentences Element Analysis. In *Proceedings of the International Conference on Computational Intelligence and Security*, *1*, 13-17 December 2008 (pp. 90-95). Washington, DC: IEEE Computer Society Press.

Yanqin, T., & Ping, G. (2005). Comparative Studies on Dimension Reduction Methods for Multispectral Remote Sensing Image. In *5 IEEE International Conference on Systems, Man and Cybernetics* (pp. 1275–1279). Berlin, Heidelberg: Springer-Verlag.

Young, P. D. (2009). *Dimension reduction and missing data in statistical discrimination. (Doctoral dissertation, USA Baylor University)*. Retrieved from http://beardocs.baylor.edu/xmlui/bitstream/handle/2104/5543/philip_young_phd.pdf.

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In *International Conference on Machine Learning (ICML)* (pp. 856–863). Washington, DC: Arizona State University.

Zhou, X. S. & Huang, T. S. (2001). Small sample learning during multimedia retrieval using BiasMap. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *1*, 11–17.

Zhu, M. (2001). *Feature extraction and dimension reduction with applications to classification and the analysis of co-occurrence data*. (Doctoral dissertation, Stanford University). Retrieved from http://search.proquest.com.ezproxy.cul.columbia.edu/docview/304728120/abstract/B59F30759A1641ABPQ/1?accountid=10226