BINARY VARIABLE EXTRACTION USING NONLINEAR PRINCIPAL COMPONENT ANALYSIS IN CLASSICAL LOCATION MODEL

LONG MEI MEI

MASTER IN 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)

LONG MEI MEI (817093)

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):

"BINARY VARIABLE EXTRACTION USING NONLINEAR PRINCIPAL COMPONENT ANALYSIS IN CLASSICAL 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 : ***24 Februari 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: ***February 24, 2016.***

Pengerusi Viva: *(Chairman for VIVA)*	**Assoc. Prof. Dr. Maznah Mat Kasim**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Dr. Norhaiza Ahmad**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Dr. Nor Idayu Mahat**	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) ***February 22, 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

Model lokasi ialah model klasifikasi ramalan yang menentukan kumpulan objek yang mengandungi campuran pembolehubah berkategori dan selanjar. Model lokasi paling ringkas dikenali sebagai model lokasi klasik, yang boleh dibina dengan mudah menggunakan penganggaran kebolehjadian maksimum. Model ini berprestasi secara ideal dengan beberapa pembolehubah binari. Walau bagaimanapun, terdapat isu banyak sel kosong apabila melibatkan sejumlah besar pembolehubah binari, b disebabkan oleh pertumbuhan sel multinomial secara eksponen dengan 2^b. Isu ini memberi kesan buruk kepada ketepatan klasifikasi apabila tiada maklumat yang boleh diperolehi daripada sel kosong untuk menganggar parameter yang diperlukan. Isu ini boleh diselesaikan dengan menggunakan pendekatan pengurangan dimensi ke dalam model lokasi klasik. Oleh itu, objektif kajian ini adalah untuk mencadangkan satu strategi klasifikasi baharu untuk mengurangkan pembolehubah binari yang besar. Ini boleh dilakukan dengan mengintegrasikan model lokasi klasik dan analisis komponen utama tak linear yang mana pengurangan pembolehubah binari adalah berdasarkan kepada *variance accounted for*, *VAF*. Model lokasi yang dicadang telah diuji dan dibanding dengan model lokasi klasik menggunakan kaedah *leave-one-out*. Keputusan membuktikan bahawa model lokasi yang dicadang boleh mengurangkan bilangan sel kosong dan mempunyai prestasi yang lebih baik dari segi kadar salah klasifikasi daripada model lokasi klasik. Model yang dicadang juga telah disahkan dengan menggunakan data sebenar. Dapatan kajian menunjukkan bahawa model ini adalah setanding atau lebih baik daripada kaedah-kaedah klasifikasi yang sedia ada. Kesimpulannya, kajian ini menunjukkan bahawa model lokasi cadangan yang baharu boleh menjadi satu kaedah alternatif dalam menyelesaikan masalah klasifikasi pembolehubah campuran, terutamanya apabila berhadapan dengan sejumlah besar pembolehubah binari.

Kata kunci: Pengurangan dimensi, Model lokasi, Kadar salah klasifikasi, Pembolehubah campuran, Analisis komponen utama tak linear.

Abstract

Location model is a predictive classification model that determines the groups of objects which contain mixed categorical and continuous variables. The simplest location model is known as classical location model, which can be constructed easily using maximum likelihood estimation. This model performs ideally with few binary variables. However, there is an issue of many empty cells when it involves a large number of binary variables, b due to the exponential growth of multinomial cells by 2^b. This issue affects the classification accuracy badly when no information can be obtained from the empty cells to estimate the required parameters. This issue can be solved by implementing the dimensionality reduction approach into the classical location model. Thus, the objective of this study is to propose a new classification strategy to reduce the large binary variables. This can be done by integrating classical location model and nonlinear principal component analysis where the binary variables reduction is based on variance accounted for, VAF. The proposed location model was tested and compared to the classical location model using leave-one-out method. The results proved that the proposed location model could reduce the number of empty cells and has better performance in term of misclassification rate than the classical location model. The proposed model was also validated using a real data. The findings showed that this model was comparable or even better than the existing classification methods. In conclusion, this study demonstrated that the new proposed location model can be an alternative method in solving the mixed variable classification problem, mainly when facing with a large number of binary variables.

Keywords: Dimensionality reduction, Location model, Misclassification rate, Mixed variables, Nonlinear principal component analysis.

Acknowledgement

Praise to God, Father, Lord of heaven and earth for all His mighty works. The Lord is my strength and my shield. My heart trusted in Him, who helped me along this long term journey. Thank you and praise Your glorious name.

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Hashibah binti Hamid for the continuous support of Master study and related research, for her patience and motivation. Her guidance helped me in all the time of programming and writing of this thesis. I could not have completed my study without her expertise and knowledge.

Special appreciation goes to my co-supervisor, Dr Nazrina binti Aziz for her knowledge teaching and motivational advice. There is no profession that is more important, yet underappreciated than teaching. Thanks for teaching me, educating me and empowering me caringly in my learning process with explanation and demonstration.

No one who achieves success does so without the help of others. My sincere appreciation goes to Dr. Nor Idayu binti Mahat, Associate Professor Dr. Azlina Murad Sani, Professor Dr. Zulikha binti Jamaludin and Dr Adyda Ibrahim for widen my research from various perspectives and strengthen my skills in academic writing as well as research methodology. I am also thankful for their willingness in sharing knowledge with me.

Besides, my appreciation goes to Universiti Utara Malaysia and the committee for their financial support. This thesis would not have been done without the financial aid. Thank you for giving me this opportunity to gain knowledge and experiences in a learning environment.

My sincere thanks and apologies also goes to my dear family, who provided me an opportunity to further my study. Thank you for always stand by my side and support me continuously. The loving ways of family is the best support that lead me. Without

their unlimited love and precious support it would not be possible to conduct this research.

Last but not least, I would like to thank my senior, Mr. C'hng Chee Keong for enlightening me with his insightful comments and helpful information. In particular, I am grateful to Ms. Irene Yong for her caring and kindness and for the sleepless nights we were discussing together. Also I thank the rest of my friends, especially Mr. Kowit and Ms. Penny for their friendships and all the fun we have had throughout my journey. Thank you for accompany and motivate me always.

Thank you very much. To all of them, I dedicate this work.

Table of Contents

List of Tables

List of Figures

Glossary of Terms

Binary variable: Variables which only take two values. It can be coded as 0 or 1, for yes or no, male or female and true or false respectively. Categorical types of data can be converted into binary structure as many variables are naturally binary.

Case study: In-depth studies of a phenomenon with cases and solutions presented. It can provides a deeper understanding to assist a person in gaining experience about a certain historical situation.

Categorical variable: Variable that can take on one of a limited or fixed number of possible values, then each individual can be assigned into a particular category as it has two or more categories.

Continuous variable: Variable that can take on any value between its minimum and maximum values. It is a quantity that has a changing value. Thus, it has an infinite number of possible values.

Dimensionality reduction: Process of reducing the number of variables under consideration. It can be divided into variable extraction and variable selection.

High dimensional data: Data that has many measurements from each sample concurrently.

Homogeneous covariance matrix: Formed of covariance matrix across groups that is all same.

Misclassification rate: A prediction error used in a classification problem for evaluation purposes. It is determined with a confusion matrix. A good prediction is able to identify true positive and true negative, otherwise, it is a bad prediction.

Mixed variables classification: Process of classifying an object into one of several populations based on data consisting a mixture of categorical and continuous variables.

Monte Carlo study: A statistical evaluation of mathematical functions using random samples. It is a simulation that uses repeated random sampling to obtain numerical results.

Principal component: A set of linearly uncorrected underlying variables that are extracted from a set of possibly correlated variables based on total variance explained through an orthogonal transformation. The first principal component has the largest possible variance. Thus, the number of principal components is less than or equal to the number of original variables.

Supervised classification: Classifying an object into one of few predefined groups. The group structures are known a priori.

Variable extraction: Reduce a large number of measured variables by extracting a small number of new variates that contain maximum variance explained.

Variable selection: Reduce irrelevant variables by choosing a subset of the original variables.

Variance accounted for: Explained variation measures the proportion to which a mathematical model accounted for the variation of a given dataset.

List of Abbreviations

CART	Classification and Regression Tree
LDA	Linear Discriminant Analysis
LM	Location Model
LOO	Leave-One-Out
MCA	Multiple Corresponding Analysis
NPCA	Nonlinear Principal Component Analysis
PCA	Principal Component Analysis
QDA	Quadratic Discriminant Analysis
R	Statistical Software with R Programming Language
SAS	Statistical Analysis Software
SPSS	Statistical Package for Social Science
VAF	Variance Accounted For

List of Publications

Long, M. M., Hamid, H. & Aziz, N. (2015). *Variables Extraction of Large Binary Variables in Discriminant Analysis based on the Location Model for Mixed Variables*. Paper Presented at the 2nd Innovation and Analytics Conference & Exhibition (IACE 2015), 29 September - 1 October 2015, Alor Setar, Kedah, Malaysia.

Hamid, H., Long, M. M. & Syed Yahaya, S. Y. (2015). New Discrimination Procedure of Classical Location Model for Large Categorical Variables. *Sains Malaysiana* (under review).

CHAPTER ONE
INTRODUCTION

1.1 Background

Classification problems abound in both theory and practical applications concerning the group memberships which in turn assign a new entity (e.g. a company, people, plant) into some predefined groups (e.g. category, department, class) (Olosunde & Soyinka, 2013). This process of discrimination is defined as a supervised classification (Hand, 2006). One of the earliest methods of classification is discriminant analysis (Crook, Edelman, & Thomas, 2007). The focus of discriminant analysis is to find a predictive classification model that can be used to classify an entity correctly to the predetermined groups (Banerjee & Pawar, 2013; Birzer & Craig-Moreland, 2008). As a matter of fact, discriminant analysis has been widely used for the classification problems to predict a group for future entities or events (Guo, Hastie, & Tibshirani, 2007).

Classification is a worth study area to be explored because it helps support major of the decision making. Volumes have been written about predictive discriminant analysis to solve classification problems in our real life. For example, classification has been applied in business and finance to predict the bankruptcy of a corporate in order to maximize the profit gained in future (Alrawashdeh, Sabri, & Ismail, 2012; Altman, 1968; Eisenbeis, 1977). Classification also has been employed in medical sciences to provide diagnostic information such as the prediction of the patients' future condition (Carakostas, Gossett, Church, & Cleghorn, 1986; Goulermas, Findlow, Nester, Howard, & Bowker, 2005; Maclaren, 1985; Poon, 2004; Takane,

The contents of the thesis is for internal user only

REFERENCES

Al-Ani, A., & Deriche, M. (2002). A new technique for combining multiple classifiers using the dempster-shafer theory of evidence. *Journal of Artificial Intelligence Research*, *17*, 333–361.

Albanis, G. T., & Batchelor, R. A. (2007). Combining heterogeneous classifiers for stock selection. *Intelligent Systems in Accounting, Finance and Management*, *15*(1-2), 1–21. doi:10.1002/isaf

Alrawashdeh, M. J., Sabri, S. R. M., & Ismail, M. T. (2012). Robust linear discriminant analysis with financial ratios in special interval. *Applied Mathematical Sciences*, *6*(121), 6021–6034.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. doi:10.1111/j.1540-6261.1968.tb00843.x

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, *59*(1), 19–35.

Asparoukhov, O., & Krzanowski, W. J. (2000). Non-parametric smoothing of the location model in mixed variable discrimination. *Statistics and Computing*, *10*, 289–297.

Banerjee, S., & Pawar, S. (2013). Predicting consumer purchase intention: A discriminant analysis approach. *NMIMS Management Review*, *XXIII*, 113–129.

Bar-Hen, A., & Daudin, J. J. (2007). Discriminant analysis based on continuous and discrete variables. In *Statistical Methods for Biostatistics and Related Fields* (pp. 3–27). Springer Berlin Heidelberg. doi:10.1007/978-3-540-32691-5_1

Basu, A., Bose, S., & Purkayastha, S. (2004). Robust discriminant analysis using weighted likelihood estimators. *Journal of Statistical Computation & Simulation*, *74*(6), 445–460.

Berardi, V. L., & Zhang, G. P. (1999). The effect of misclassificaton costs on neural network classifiers. *Decision Sciences*, *30*(3), 659–682.

Berchuck, A., Iversen, E. S., Luo, J., Clarke, J., Horne, H., Levine, D. A., … Lancaster, J. M. (2009). Microarray analysis of early stage serous ovarian cancers shows profiles predictive of favorable outcome. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *15*(7), 2448–2455. doi:10.1158/1078-0432.CCR-08-2430

Betz, N. E. (1987). Use of discriminant analysis in counseling psychology research. *Journal of Counseling Psychology*, *34*(4), 393–403. doi:10.1037/0022-0167.34.4.393

Birzer, M. L., & Craig-Moreland, D. E. (2008). Using discriminant analysis in policing research. *Professional Issues in Criminal Justice*, *3*(2), 33–48.

Blasius, J., & Gower, J. C. (2005). Multivariate prediction with nonlinear principal components analysis: Application. *Quality and Quantity*, *39*, 373–390. doi:10.1007/s11135-005-3006-0

Braga-Neto, U. M., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, *20*(3), 374–380. doi:10.1093/bioinformatics/btg419

Braga-Neto, U. M., Hashimoto, R., Dougherty, E. R., Nguyen, D. V, & Carroll, R. J. (2004). Is cross-validation better than resubstitution for ranking genes? *Bioinformatics*, *20*(2), 253–258. doi:10.1093/bioinformatics/btg399

Brito, I., Celeux, G., & Ferreira, A. S. (2006). Combining methods in supervised classification: A comparative study on discrete and continuous problems. *Revstat - Statistical Journal*, *4*(3), 201–225. Retrieved from http://www.ine.pt/revstat/pdf/rs060302.pdf

Carakostas, M. C., Gossett, K. A., Church, G. E., & Cleghorn, B. L. (1986). Veterinary Pathology Online. *Veterinary Pathology*, *23*, 254–269. doi:10.1177/030098588602300306

Chang, P. C., & Afifi, A. A. (1974). Classification based on dichotomous and continuous variables. *Journal of the American Statistical Association*, *69*(346), 336–339.

Chen, K., Wang, L., & Chi, H. (1997). Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification. *International Journal of Pattern Recognition and Artificial Intelligence*, *11*(3), 417–445. doi:10.1142/S0218001497000196

Cochran, W. G., & Hopkins, C. E. (1961). Some classification problems with multivariate qualitative data. *Biometrics*, *17*(1), 10–32.

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. doi:10.1155/2013/302163

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*, 1447–1465. doi:10.1016/j.ejor.2006.09.100

Daudin, J. J. (1986). Selection of variables in mixed-variable discriminant analysis. *Biometrics*, *42*(3), 473–481.

Daudin, J. J., & Bar-Hen, A. (1999). Selection in discriminant analysis with continuous and discrete variables. *Computational Statistics and Data Analysis*,

32, 161–175. doi:10.1016/S0167-9473(99)00027-4

Day, N. E., & Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics*, *23*(2), 313–323.

De Leeuw, J. (2006). Nonlinear principal component analysis and related techniques. In J. B. M Greenacre (Ed.), *Multiple Correspondence Analysis and Related Methods* (pp. 107–133). Chapman and Hall, Boca Raton, FA. doi:10.1109/IJCNN.2003.1223477

De Leeuw, J. (2011). *Nonlinear principal component analysis and related techniques*. Department of Statistics, UCLA. Retrieved from https://escholarship.org/uc/item/7bt7j6nk

De Leeuw, J. (2013). History of nonlinear principal component analysis, 1–14.

De Leeuw, J., & Mair, P. (2009). Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, *31*(4), 1–21. Retrieved from http://www.jstatsoft.org/

de 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. doi:10.1080/02664761003758976

Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accountin Research*, *10*(1), 167–179.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1–33. Retrieved from http://mlo.cs.man.ac.uk/resources/Curses.pdf

Dray, S. (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational Statistics and Data Analysis*, *52*, 2228–2237. doi:10.1016/j.csda.2007.07.015

Eisenbeis, R. A. (1977). Pitfalls in the application of discriminant analysis in Business, Finance, and Economics. *The Journal of Finance*, *32*(3), 875–900.

Everitt, B. S. & Merette, C. (1990). The Clustering of Mixed-mode Data: A Comparison of Possible Approaches. *Journal of Applied Statistics, 17,* 283-297.

Fan, J., & Fan, Y. (2008). High dimensional classification using features annealed indepence rules. *Annals of Statistics*, *36*(6), 2605-2637. doi:10.1214/07-AOS504.High

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101-148. doi:10.1063/1.3520482

Ferrari, P. A., & Manzi, G. (2010). Nonlinear principal component analysis as a tool for the evaluation of customer satisfaction. *Quality Technology and*

Quantitative Management, *7*(2), 117–132. Retrieved from http://air.unimi.it/handle/2434/141402\nhttp://web2.cc.nctu.edu.tw/~qtqm/qtqmpapers/2010V7N2/2010V7N2_F2.pdf

Ferrari, P. A., & Salini, S. (2011). Complementary use of rasch models and nonlinear principal components analysis in the assessment of the opinion of Europeans about utilities. *Journal of Classification*, *28*, 53–69. doi:10.1007/s00357-011-9081-0

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, England: Wiley.

Glick, N. (1973). Sample-based multinomial classification. *Biometrics*, *29*(2), 241–256.

Goulermas, J. Y., Findlow, A. H., Nester, C. J., Howard, D., & Bowker, P. (2005). Automated design of robust discriminant analysis classifier for foot pressure lesions using kinematic data. *IEEE Transactions on Biomedical Engineering*, *52*(9), 1549–1562. doi:10.1109/TBME.2005.851519

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, *7*(7), 745–757. doi:10.1002/sim.4780070704

Guo, Y., Hastie, T., & Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, *8*(1), 86–100. doi:10.1093/biostatistics/kxj035

Gupta, V. (2013). *Exploring data generated by pocket devices*. London. Retrieved from http://files.howtolivewiki.com/SMART_CITIES/The_Smart_City.To_Whos_Advantage.Pocket_Devices_and_Data_Trails.Vinay_Gupta.pdf

Hamid, H. (2010). A new approach for classifying large number of mixed variables. *International Scholarly and Scientific Research and Innovation*, *4*(10), 120–125. doi:14621

Hamid, H. (2014). *Integrated smoothed location model and data reduction approaches for multi variables classification*. Unpublished Doctoral Dissertation. Universiti Utara Malaysia.

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.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, *21*(1), 1–15. doi:10.1214/088342306000000079

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two-group classification methods. *Educational and Psychological Measurement*, *71*(5), 870–901. doi:10.1177/0013164411398357

Holden, J. E., & Kelley, K. (2010). The effects of initially misclassified data on the effectiveness of discriminant function analysis and finite mixture modeling. *Educational and Psychological Measurement*, *70*(1), 36–55. doi:10.1177/0013164409344533

Hothorn, T., & Lausen, B. (2003). Double-bagging: Combining classifiers by bootstrap aggregation. *Pattern Recognition*, *36*, 1303–1309. doi:10.1016/S0031-3203(02)00169-3

Hsiao, C., & Chen, H. (2010). On classification from the view of outliers. *IAENG International Journal of Computer Science*, *37*(4), 1–9. Retrieved from http://arxiv.org/abs/0907.5155

Jin, H., & Kim, S. (2015). Performance evaluations of diagnostic prediction with neural networks with data filters in different types. *International Journal of Bio-Science and Bio-Technology*, *7*(1), 61–70. doi:http://dx.doi.org/10.14257/ijbsbt.2015.7.1.07

Katz, M. H. (2011). *Multivariate analysis: A practical guide for clinicians and public health researchers*. Cambridge: Cambridge University Press.

Kim, K., Aronov, P., Zakharkin, S. O., Anderson, D., Perroud, B., Thompson, I. M., & Weiss, R. H. (2009). Urine metabolomics analysis for kidney cancer detection and biomarker discovery. *Molecular & Cellular Proteomics: MCP*, *8*(3), 558–570. doi:10.1074/mcp.M800165-MCP200

Knoke, J. D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics*, *38*(1), 191–200.

Kristensen, P. (1992). Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology*, *3*(3), 210–215. Retrieved from http://www.jstor.org/stable/3703154

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of American Statistical Association*, *70*(352), 782–790.

Krzanowski, W. J. (1979). Some linear transformations for mixtures of binary and continuous variables, with particular reference to linear discriminant analysis. *Biometrika*, *66*(1), 33–39. doi:10.1093/biomet/66.1.33

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, *36*(3), 493–499.

Krzanowski, W. J. (1982). Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis testing approach. *Biometrics*, *38*(4), 991-

1002.

Krzanowski, W. J. (1983). Stepwise location model choice in mixed-variable discrimination. *Journal of the Royal Statistical Society. Series C (Applied Statisitcs)*, *32*(3), 260–266.

Krzanowski, W. J. (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification*, *10*(1), 25–49. doi:10.1007/BF02638452

Krzanowski, W. J. (1995). Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis. *Computational Statistics & Data Analysis*, *19*, 419–431. doi:10.1016/0167-9473(94)00011-7

Lachenbruch, P. A., & Goldstein, M. (1979). Discriminant analysis. *Biometrics*, *35*(1), 69–85.

LeBlanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, *91*(436), 1641–1650. doi:10.1080/01621459.1996.10476733

Lee, S., Huang, J. Z., & Hu, J. (2010). Sparse logistic principal components analysis for binary data. *Annals of Applied Statistics*, *4*(3), 1579–1601. doi:10.1016/j.biotechadv.2011.08.021.Secreted

Li, Q. (2006). *An integrated framework of feature selection and extraction for appearance-based recognition*. Unpublished Doctoral Dissertation. University of Delaware Newark, DE, USA.

Li, Q., & Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, *86*, 266–292. doi:10.1016/S0047-259X(02)00025-8

Li, X., & Ye, N. (2006). A supervised clustering and classification algorithm for mining data with mixed variables. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *36*(2), 396–406. doi:10.1109/TSMCA.2005.853501

Lillvist, A. (2009). Observations of social competence of children in need of special support based on traditional disability categories versus a functional approach. *Early Child Development and Care*, *180*(9), 1129–1142. doi:10.1080/03004430902830297

Linting, M. (2007). *Nonparametric inference in nonlinear principal components analysis: Exploration and beyond*. Doctoral Thesis, Leiden University. Retrieved from http://hdl.handle.net/1887/12386

Linting, M., Meulman, J. J., Groenen, P. J. F., & van der Kooij, A. J. (2007a). Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, *12*(3), 336–358. doi:10.1037/1082-989X.12.3.336

Linting, M., Meulman, J. J., Groenen, P. J. F., & van der Kooij, A. J. (2007b). Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological Methods, 12*(3), 359–379. doi:10.1037/1082-989X.12.3.359

Linting, M., & van der Kooij, A. J. (2012). Nonlinear principal components analysis with CATPCA: A tutorial. *Journal of Personality Assessment, 94*(1), 12–25. doi:10.1080/00223891.2011.627965

Linting, M., van Os, B. J., & Meulman, J. J. (2011). Statistical significance of the contribution of variables to the PCA solution: An alternative premutation strategy. *Psychometrika, 76*(3), 440–460.

Little, R. J. A., & Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, *72*(3), 497–512.

Lombardo, R., & Meulman, J. J. (2010). Multiple correspondence analysis via polynomial transformations of ordered categorical variables. *Journal of Classification, 27,* 191–210. doi:10.1007/s00357-010-

Maclaren, W. M. (1985). Using discriminant analysis to predict attacks of complicated pneumoconiosis in coalworkers. *Journal of the Royal Statistical Society, Series D (The Statistician), 34*(2), 197–208.

Mahat, N. I. (2006). *Some investigations in discriminant analysis with mixed variables*. Unpublished Doctoral Dissertation. University of Exeter, London, UK.

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*(2), 105–122. doi:10.1007/s11634-007-0009-9

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.

Mair, P., & Leeuw, J. De. (2008). Rank and set restrictions for homogeneity analysis in R: The "homals" package. In *JSM 2008 Proceedings, Statistical Computing Section.* (pp. 2142–2149).

Manisera, M., Dusseldorp, E., & van der Kooij, A. J. (2005). Component structure of job satisfaction based on Herzberg's theory. Retrieved May 9, 2015, from http://www.datatheory.nl/pages/fullmanuscript_final_epm.pdf

Manisera, M., van der Kooij, A. J., & Dusseldorp, E. (2010). Identifying the component structure of job satisfaction by nonlinear principal components analysis. *Quality Technology and Quantitative Management, 7*, 97–115. Retrieved from http://elisedusseldorp.nl/pdf/Manisera_QTQM2010.pdf

Marian, N., Villarroya, A., & Oller, J. M. (2003). Minimum distance probability discriminant analysis for mixed variables. *Biometrics*, *59*, 248–253.

Markos, A. I., Vozalis, M. G., & Margaritis, K. G. (2010). An optimal scaling approach to collaborative filtering using categorical principal component analysis and neighborhood formation. In *Artificial Intelligence Applications and Innovations* (pp. 22-29). Springer Berlin Heidelberg.

Meulman, J. J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations. *Psychometrika, 57*(4), 539–565.

Meulman, J. J. (2003). Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, *68*(4), 493–517.

Meulman, J. J., van der Kooij, A. J., & Heiser, W. J. (2004). Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In D. Kaplan (Ed.), *Handbook of Quantitative Methods in the Social Sciences* (pp. 49–70). Newbury Park, CA: Sage Publications. doi:10.4135/9781412986311

Moustaki, I., & Papageorgiou, I. (2005). Latent class models for mixed variables with applications in Archaeometry. *Computational Statistics and Data Analysis*, *48*(3), 659–675. doi:10.1016/j.csda.2004.03.001

Nasios, N., & Bors, A. G. (2007). Kernel-based classification using quantum mechanics. *Pattern Recognition*, *40*, 875–889. doi:10.1016/j.patcog.2006.08.011

Nishisato, S., & Arri, P. S. (1975). Nonlinear programming approach to optimal scaling of partially ordered categories. *Psychometrika*, *40*(4), 525–548. doi:10.1007/BF02291554

Olkin, I., & Tate, R. F. (1961). Multivariate correlation models with discrete and continuous variables. *The Annals of Mathematical Statistics 32*, 448–465.

Olosunde, A. A., & Soyinka, A. T. (2013). Discrimination and classification of poultry feeds data. *International Journal of Mathematical Research*, *2*(5), 37–41. doi:10.1080/00207390600819003

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, *49*, 974–997. doi:10.1016/j.csda.2004.06.015

Poon, W.-Y. (2004). Identifying influence observations in discriminant analysis. *Statistical Methods in Medical Reserach*, *13*, 291–308. doi:10.1191/0962280204sm367ra

Prokop, M., & Řezanková, H. (2011). Data dimensionality reduction methods for

ordinal data. *International Days of Statistics and Economics, Prague*, 523–533.

Prokop, M., & Řezanková, H. (2013). Comparison of dimensionality reduction methods applied to ordinal data. *The Seventh International Days of Statistics and Economics, Prague*, 1150–1159.

Ramadevi, G. N., & Usharaani, K. (2013). Study on dimensionality reduction techniques and applications. *Publications of Problems & Application in Engineering Research*, *4*(1), 134–140.

Russom, P. (2013). *Managing big data*. Washington.

Schein, a I., Saul, L. K., & Ungar, L. H. (2003). A generalized linear model for principal component analysis of binary data. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Retrieved from http://research.microsoft.com/conferences/aistats2003/proceedings/papers.htm

Schmitz, P. I. M., Habbema, J. D. F., & Hermans, J. (1983). The performance of logistic discrimination on myocardial infarction data, in comparison with some other discriminant analysis methods. *Statistics in Medicine*, *2*(2), 199–205.

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 the National Cancer Institute*, *95*(1), 14–18.

Solanas, A., Manolov, R., Leiva, D., & Richard's, M. M. (2011). Retaining principal components for discrete variables. *Anuario de Psicologia*, *41*(1-3), 33–50.

Takane, Y., Bozdogan, H., & Shibayama, T. (1987). Ideal point dicriminant analysis. *Psychometrika*, *52*(3), 371–392.

Titterington, D. M., Murray, G. D., Murray, L. S., Speigelhalter, D. J., Skene, A. M., Habbema, J. D. F., & Gelpke, G. J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society A*, *144*(2), 145–175. Retrieved from http://www.jstor.org/stable/2981918

van Heerden, C., Barnard, E., Davel, M., van der Walt, C., van Dyk, E., Feld, M., & Muller, C. (2010). Combining regression and classification methods for improving automatic speaker age recognition. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference* (pp. 5174–5177). IEEE. doi:10.1109/ICASSP.2010.5495006

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.

Wernecke, K.-D. (1992). A coupling procedure for the discrimination of mixed data. *Biometrics*, *48*(2), 497–506.

Wernecke, K.-D., Unger, S., & Kalb, G. (1986). The use of combined classifiers in medical functional diagnostics. *Biometrical Journal*, *28*(1), 81–88. doi:10.1002/bimj.4710280116

Xu, L., Krzyzak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, *22*(3), 418–435. doi:10.1109/21.155943

Yang, Y. (2005). Can the strangths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, *92*(4), 937–950. Retrieved from http://www.jstor.org/stable/20441246

Young, P. D. (2009). *Dimension reduction and missing data in statistical discrimination*. Unpublished Doctoral Dissertation. USA Baylor University.

Zhang, M. Q. (2000). Discriminant analysis and its application in DNA sequence motif recognition. *Briefings in Bioinformatics*, *1*(4), 1–12. doi:10.1093/bib/1.4.331

Zheng, H., & Zhang, Y. (2008). Feature selection for high-dimensional data in astronomy. *Advances in Space Research*, *41*(12), 1960–1964. doi:10.1016/j.asr.2007.08.033