Moments, Factor Scores and Limiting Distributions of Individual Mahalanobis Distances

Deliang Dai

Licentiate Dissertation
Department of Economics and Statistics
Linnaeus University
Box 451
351 06 Växjö

ISBN: 978-91-87427-91-6

Acknowledgement

I would like to express my deepest appreciation to those people who provided me their help to finish this thesis.

First and foremost, the greatest gratitude goes to my supervisor Prof. Thomas Holgersson for all the comments, discussions and patience for leading me to the right direction on my research. Your unlimited knowledge and generous guidance have been invaluable to me throughout this amazing research journey. Thank you for such an interesting research area to me.

I an also very grateful to my assistant supervisor Prof. Ghazi Shukur for all the supports. Many thanks to Dr. Peter Karlsson who helps me a lot both on my academics and my “träning”. Thank you for showing me the real meanings of humble and kind. Thanks go to Dr. Hyunjoo Karlsson, for all the interesting conversations and foods.

Many thanks to Prof. Fan Yang Wallentin and Prof. Adam Taube for bringing me to the world of statistics. Thanks go to Prof. Dietrich von Rosen and Tatjana von Rosen for their kindly help and valuable conversations.

Thanks go to my dear roommates. Thank you Aziz. It is always very interesting to chat with you. I’ve learned many useful knowledge from you. From research to practical tips of living in Sweden. Thank you Chizheng for the badminton games and all the Chinese foods. Thank you Abdulaziz for all the interesting chats on football and casual life. You amazing guys make our office a fantastic place.

Thanks to all the colleagues at the Department of Economics and Statistics for making a good circumstance for working.

Last but not least, thanks to my wife Yuli for her support and patience to me.

Deliang Dai
2014.05.14

Content

1. Introduction

1.1. Background

In multivariate analysis, Mahalanobis distance (hereafter MD) is a fundamental statistic which has been investigated by many researchers in different areas. It was proposed by Mahalanobis (1930). The MD is used for measuring the distance between random vector and its hypothesis mean vector (Rao, 1945; Hotelling, 1933). Based on this idea, MD is developed into different forms of definitions. MDs with different forms are referred to some literatures (Gower, 1966; Khatri, 1968; Diciccio and Romano, 1988).

The MDs are widely implemented in many directions. Firstly, one is the measure between two random vectors such as discriminant analysis on the linear and quadratic discriminations (Fisher, 1936; Srivastava and Khatri, 1979; Fisher, 1940; Hastie et al., 1995; Fujikoshi, 2002; Pavlenko, 2003; McLachlan, 2004) and classification with covariates (Anderson, 1951; Friedman et al., 2001; Berger, 1980; Blackwell, 1979; Leung and Srivastava, 1983a,b). Secondly, one is detection for the multivariate outliers (Mardia et al., 1980; Wilks, 1963). Thirdly, due to the connections to Hotelling T^2, MD is also studied in hypothesis testing (Fujikoshi et al., 2011; Mardia et al., 1980). Fourthly, Mardia (1974); Mardia et al. (1980); Mitchell and Krzanowski (1985); Holgersson and Shukur (2001) use skewness and kurtosis as the criteria statistics for assessing the assumption of multivariate normality.

As we known, the crucial part in MD is the inverse covariance matrix. By taking into account of the covariance matrix, the MD has some advantages. It behaviors better than the other distances measures such as Euclidean distance. However, there are a few problems on derivation of the inverse covariance matrix. One problem is the increasing dimension asymptotic situation that the sample size is proportional to the dimension of variables (Serdobolskii, 2010; Girko, 2010; Ledoit and Wolf, 2004; Jonsson, 1982; Marčenko and Pastur, 1967). With this kind of dataset, the complexity of the structure of covariance matrix increases dramatically. Another problem is when the sample size (n) is much less than the dimensions of variables (p) (Bai and Silverstein, 2009). The computational expense will be rising sharply. All the issues above require further studies to improve the calculation of covariance matrix.

So far, more and more studies of the inverse covariance matrix are developed on the non-classical dataset. Here, the classical data means the sample size (n) is much larger than the dimensions of variables (p). There are several ways for analysis the dataset which has a proportional ratio p/n. One is the additional way while another is the subtraction way. The Shrinkage is a typical method as the additional way. It is implemented by imposing some minor positive numbers on the diagonal of covariance matrix. Therefore, we can get an estimation of inverse covariance matrix. The estimator of inverse covariance matrix has much smaller variation by giving up its unbiased property. The subtraction way is reduction of dimensions. Factor analysis and principal component analysis are two methods for the dimension reduction. They could both maintain the necessary information and reduce the dimension of the variables.

1.2. Contributions of thesis

This thesis concerns on the properties of the MDs under different circumstances. The distributional properties, first moments of different types of MDs and central limit theorem (CLT) are derived in the thesis. Several different directions on deriving the inverse covariance matrix are investigated. A new estimator of the inverse covariance matrix is proposed. The corresponding MDs on the new estimator of inverse covariance matrix and the limiting of its moments are also studied. Some of the methods are implemented with empirical dataset.

1.3. Outline

First, a brief introduction of the background is given. Section 2 introduces some basic definitions of MDs. The summaries of papers will be presented in Section 3. Three papers are appended: Paper I with the title "Estimating individual Mahalanobis distance in high-dimensional data"; Paper II with the title "High-dimensional CLTs for individual Mahalanobis distances"; Paper III with the title "Mahalanobis distances of factor structure data".

2. Mahalanobis distances

The classical definition of MD (Mahalanobis, 1936) is

$$D_{ij} = (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' \mathbf{\Sigma}^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) .$$

It measures the difference between two mean vectors. For different considerations, there are several types of MDs. In this thesis, we consider three types of MDs as follows,

$$D_{ii} = p^{-1}(\mathbf{X}_i - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1} (\mathbf{X}_i - \boldsymbol{\mu}) , \tag{1}$$

$$d_{ii} = p^{-1}\big(\mathbf{X}_i - \bar{\mathbf{X}}\big)'\mathbf{S}^{-1} \big(\mathbf{X}_i - \bar{\mathbf{X}}\big) , \tag{2}$$

$$d_{(ii)} = p^{-1}\big(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)}\big)'\mathbf{S}_{(i)}^{-1} \big(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)}\big) . \tag{3}$$

(1) is the classical definition of MD with known mean and variance. (2) is the MD with sample mean and covariance matrix. (3) concerns the so-called "leave-one-out" and "leave-two-out" (De Maesschalck et al., 2000; Mardia, 1977) random vectors. By leaving the ith observation out, we get the independence between the sample covariance matrix and the centralized vector. Moreover, it will not contaminate the sample mean and covariance matrix especially for high dimensional dataset.

3. Summary of papers

3.1. Paper I: Estimating individual Mahalanobis distance in high-dimensional data

In Paper I, several different types of MDs are defined. They are mainly built on different kinds of inverse covariance matrices. The corresponding first moments are derived. The limitings of the first moments reveal some unexpected results. The reason is confirmed that the sample covariance matrix is not an appropriate estimator for the high dimensional dataset. Therefore, we propose a new estimator of the inverse covariance matrix. The new estimator is developed basing on some earlier results (Efron and Morris, 1976). The properties of the corresponding new MDs basing on the new inverse covariance matrix are investigated.

3.2. Paper II: High-dimensional CLTs for individual Mahalanobis distances

In Paper II, the central limit theorems (CLT) for different types of MDs are derived. Different from the classical theorem, we assume that the sample size n and dimension of variables p go to infinite simultaneously. The CLTs for different types of MDs have different convergence rates and convergence to different points under high dimensional situation.

3.3. Paper III: Mahalanobis distances of factor structure data

In Paper III, we use factor model to reduce the dimensions of the data set and build a factor based inverse covariance matrix. The inverse of covariance matrix behaviors well in the new types of MDs. The detections of the source of outliers are also studied on the additive type of outliers. In the last section, the methods are implemented in an empirical study.

4. Summary in Swedish

I avhandlingen studeras olika egenskaper hos estimatorer av olika Mahalanobis avstånd.

I det första pappret härleds grundläggande egenskaper hos olika Mahalanobis avstånden. Vidare, så föreslås en ny estimator av inverskovariansmatrisen, vilken ingår i de olika Mahalanobis avstånden.

I det andra pappret, så antas det att den datagenerande processen följer en faktor struktur, och baserat på detta antagande så härleds estimatorer av olika Mahalanobis avstånd. Dessutom utvecklas en ny metod för upptäcka extremvärden och en empirisk applikation av metoden ges i pappret.

I sista pappret så undersöks Centrala gränsvärdessatser för olika estimatorer av de olika Mahalanobis avstånden i en högdimensionel kontext.

Anderson, T. W. (1951). Classification by multivariate analysis, *Psychometrika* **16**(1): 31–50.

Bai, Z. and Silverstein, J. (2009). *Spectral analysis of large dimensional random matrices*, Springer Verlag.

Berger, J. (1980). *Statistical decision theory, foundations, concepts, and methods*, Springer series in statistics: Probability and its applications, Springer-Verlag.

Blackwell, D. (1979). *Theory of games and statistical decisions*, Courier Dover Publications.

De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D. L. (2000). The mahalanobis distance, *Chemometrics and Intelligent Laboratory Systems* **50**(1): 1–18.

Diciccio, T. and Romano, J. (1988). A review of bootstrap confidence intervals, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 338–354.

Efron, B. and Morris, C. (1976). Multivariate empirical bayes and estimation of covariance matrices, *The Annals of Statistics* pp. 22–32.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems, *Annals of Human Genetics* **7**(2): 179–188.

Fisher, R. A. (1940). The precision of discriminant functions, *Annals of Human Genetics* **10**(1): 422–429.

Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The elements of statistical learning*, Vol. 1, Springer Series in Statistics.

Fujikoshi, Y. (2002). Selection of variables for discriminant analysis in a high-dimensional case, *Sankhyā: The Indian Journal of Statistics, Series A* pp. 256–267.

Fujikoshi, Y., Ulyanov, V. and Shimizu, R. (2011). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*, Vol. 760, Wiley.

Girko, V. (2010). *Statistical analysis of observations of increasing dimension*, Vol. 28, Springer.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* **53**(3-4): 325–338.

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis, *The Annals of Statistics* **23**(1): 73–102.

Holgersson, H. and Shukur, G. (2001). Some aspects of non-normality tests in systems of regression equations, *Communications in Statistics-Simulation and Computation* **30**(2): 291–310.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components., *Journal of educational psychology* **24**(6): 417.

Jonsson, D. (1982). Some limit theorems for the eigenvalues of a sample covariance matrix, *Journal of Multivariate Analysis* **12**(1): 1–38.

Khatri, C. (1968). Some results for the singular normal multivariate regression models, *Sankhyā: The Indian Journal of Statistics, Series A* pp. 267–280.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices, *Journal of multivariate analysis* **88**(2): 365–411.

Leung, C. and Srivastava, M. (1983a). Asymptotic comparison of two discriminants used in normal covariate classification, *Communications in Statistics-Theory and Methods* **12**(14): 1637–1646.

Leung, C. and Srivastava, M. (1983b). Covariate classification for two correlated populations, *Communications in Statistics-Theory and Methods* **12**(2): 223–241.

Mahalanobis, P. (1930). On tests and measures of group divergence, *J. Asiat. Soc.*, Vol. 26, pp. 541–588.

Mahalanobis, P. (1936). On the generalized distance in statistics, *Proceedings of the National Institute of Sciences of India*, Vol. 2, New Delhi, pp. 49–55.

Marčenko, V. and Pastur, L. (1967). Distribution of eigenvalues for some sets of random matrices, *Sbornik: Mathematics* **1**(4): 457–483.

Mardia, K. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies, *Sankhyā: The Indian Journal of Statistics, Series B* pp. 115–128.

Mardia, K. (1977). Mahalanobis distances and angles, *Multivariate analysis IV* pp. 495–511.

Mardia, K., Kent, J. and Bibby, J. (1980). *Multivariate analysis*, Academic press.

McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition*, Vol. 544, John Wiley & Sons.

Mitchell, A. and Krzanowski, W. (1985). The mahalanobis distance and elliptic distributions, *Biometrika* **72**(2): 464–467.

Pavlenko, T. (2003). On feature selection, curse-of-dimensionality and error probability in discriminant analysis, *Journal of statistical planning and inference* **115**(2): 565–584.

Rao, C. R. (1945). Familial correlations or the multivariate generalisations of the intraclass correlations, *Current Science* **14**(3): P66–67.

Serdobolskii, V. (2010). *Multivariate statistical analysis: A high-dimensional approach*, Vol. 41, Springer.

Srivastava, S. and Khatri, C. (1979). *An introduction to multivariate statistics*, North-Holland/New York.
URL: *http://books.google.se/books?id=swbvAAAAMAAJ*

Wilks, S. (1963). Multivariate statistical outliers, *Sankhyā: The Indian Journal of Statistics, Series A* pp. 407–426.