Mahalanobis distances on factor structured data

Deliang Dai[1]

[1]Department of Economics and Statistics
Linnaeus University, Växjö

August 24, 2014

Outline

Outline

Background of Mahalanobis distance

Mahalanobis distance (MD) was proposed by Mahalanobis (1936). It is used for measuring the distance between two random vectors or one random vector and its mean vector. It is widely used in many areas such as cluster analysis; discriminant analysis (Fisher; 1940); assessing multivariate normality (Mardia; 1974; Mardia et al.; 1980; Mitchell and Krzanowski; 1985; Holgersson and Shukur; 2001); hypothesis testing and outlier detecting (Wilks; 1963; Mardia et al.; 1980).

Definition of Mahalanobis distance

The definition of Mahalanobis distance is given as follows,

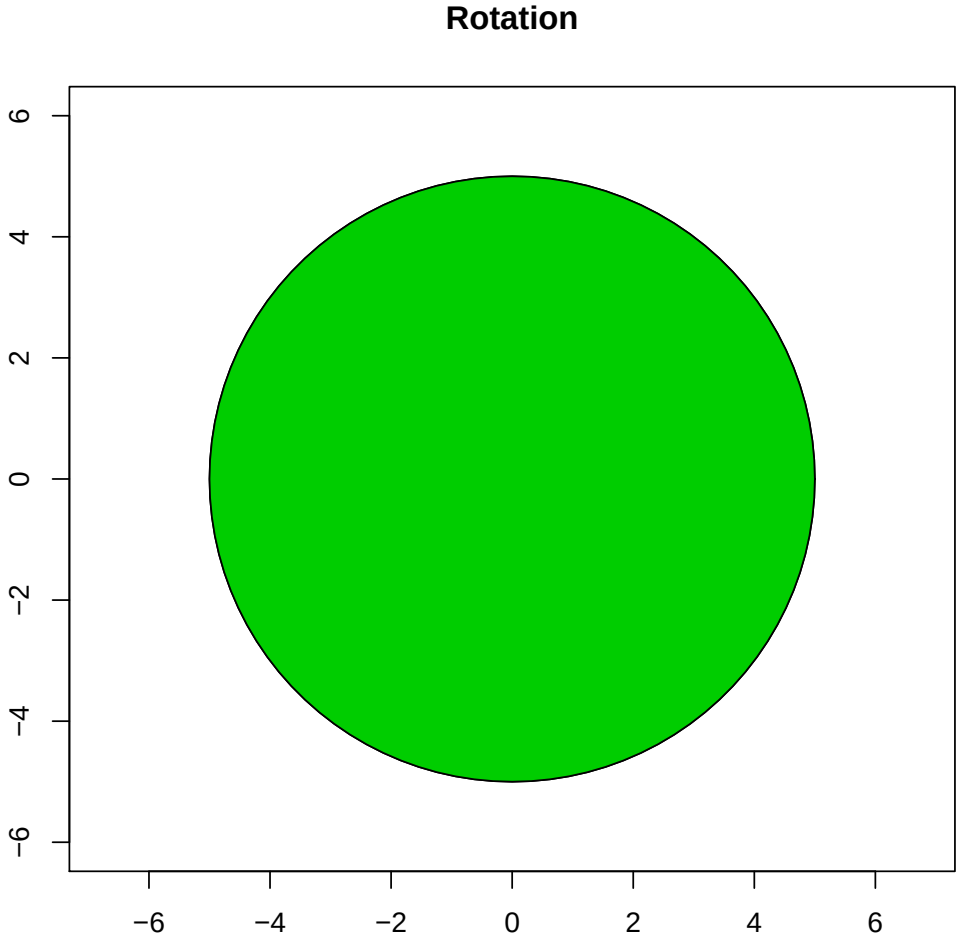
Definition 1

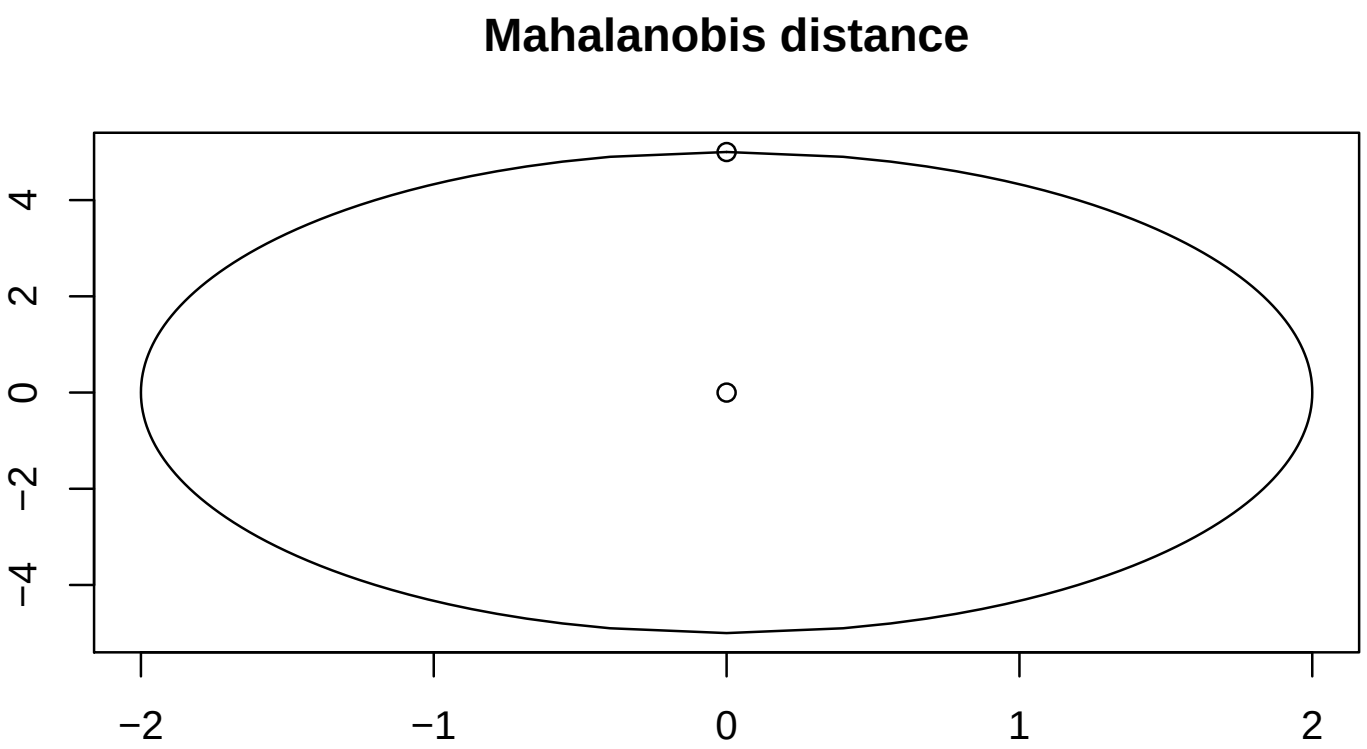
Let $\mathbf{X}_i : p \times 1$ be a random vector such that $E[\mathbf{X}_i] = \boldsymbol{\mu}_{p\times 1}$ and $E\left[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})'\right] = \boldsymbol{\Sigma}_{p\times p}$ for $i = 1, ..., n$. Then we make the following definition:

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

Advantages of MD

The Mahalanobis distance which includes the covariance matrix is linear transformation invariant.


Rotation
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6


Mahalanobis distance
4
2
0
−2
−4
−2
−1
0
1
2

Background of factor model

Factor model is widely applied in many researches on the unobserved measurements. For example, the arbitrage pricing theory (Ross; 1976) based on the assumption that asset returns follow a factor structure. In consumer theory, factor analysis shows that utility maximizing could be described by at most three factors (Gorman et al.; 1981; Lewbel; 1991). In desegregate business cycle analysis, the factor model is used to identify the common and specific shocks that drive a country's economy (Gregory and Head; 1999; Bai; 2003).

The combination of the factor model and Mahalanobis distance offers an option for both approximation of the structure of covariance and measuring the distance between observations at the same time. The balance between simplicity of structure and information offers a good method for the dependent structured data.
In this presentation, we only concern the detection of outlier.

Outline

Definition of factor model

Definition 2

Let $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, be the $p \times 1$ random observations with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The factor model could be represented as

$$\mathbf{x} - \boldsymbol{\mu}_{(p\times 1)} = \mathbf{L}_{(p\times m)}\mathbf{F}_{(m\times 1)} + \boldsymbol{\varepsilon}_{(p\times 1)},$$

where $\mathbf{x}$ is the observed date ($p > m$), $\boldsymbol{\mu}$ the corresponding expectation , $\mathbf{L}$ the factor loadings and $\mathbf{F}$ is a $m \times 1$ vector of factor scores.

Assumptions of factor model

Beside the definition of factor model above, we need some more assumptions as follows.

Definition 3

Let $E(\mathbf{F}) = \mathbf{0}$, $Cov(\mathbf{F}) = \mathbf{I}_{m\times m}$ and $E(\varepsilon) = \mathbf{0}$. As defined above, assume $\varepsilon \sim N(\mathbf{0}_{p\times 1}, \mathbf{\Psi})$, where $\mathbf{\Psi}$ is a diagonal matrix and $\mathbf{F} \sim N(\mathbf{0}_{m\times 1}, \mathbf{I})$ are distributed independently ($Cov(\varepsilon, \mathbf{F}) = \mathbf{0}$).

Definitions of MDs on factor model

Definition 4

The MD on $\mathbf{F}$ part is

$$D(\mathbf{F}_i) = \mathbf{F}_i' \mathrm{cov}(\mathbf{F}_i)^{-1} \mathbf{F}_i = \mathbf{F}_i' \mathbf{F}_i,$$

while for the ε part, the MD is

$$D(\varepsilon_i) = \varepsilon_i' \mathbf{\Psi}^{-1} \varepsilon_i.$$

Definitions of MDs on factor model

Definition 5

The MDs for the estimated factor scores and error are:

$$D\left(\hat{\mathbf{F}}_i\right) = \hat{\mathbf{F}}_i' cov(\hat{\mathbf{F}}_i)^{-1}\hat{\mathbf{F}}_i \text{ and } D\left(\hat{\varepsilon}_i\right) = \hat{\varepsilon}_i'\hat{\boldsymbol{\Psi}}^{-1}\hat{\varepsilon}_i.$$

where $\hat{\mathbf{F}}_i = \hat{\mathbf{L}}'\mathbf{S}^{-1}(\mathbf{x_i} - \bar{\mathbf{x}})$ is the estimated factor scores from the regression method (Johnson and Wichern; 2002).

Outline

Joint distribution of seperated MDs

Proposition 1

The joint distribution of $D(\mathbf{F}_i)$ *and* $D(\varepsilon_i)$ *is,*

$$Z = \begin{bmatrix} D(\mathbf{F}_i) \\ D(\varepsilon_i) \end{bmatrix} \sim \begin{bmatrix} \chi^2_{(m)} \\ \chi^2_{(p)} \end{bmatrix}, \operatorname{cov}(Z) = \begin{bmatrix} 2m & 0 \\ 0 & 2p \end{bmatrix}.$$

Joint distribution of seperated MDs

Proposition 2

The asymptotic distributions of the MD under the factor model with an unknown factor structure is

$$\hat{\mathbf{Y}} = \begin{bmatrix} D\left(\hat{\mathbf{F}}_i\right) \\ D\left(\hat{\varepsilon}_i\right) \end{bmatrix} \xrightarrow{\ell} \begin{bmatrix} \chi^2_{(m)} \\ \chi^2_{(p)} \end{bmatrix}, \operatorname{cov}\left(\hat{\mathbf{Y}}\right) = \begin{bmatrix} 2m & 0 \\ 0 & 2p \end{bmatrix}.$$

Outline

Definitions of contaminated data

Definition 6

Assuming two kinds of outliers are interested: (i) an additive outlier in ε_i and (ii) an additive outlier in $\mathbf{F}_i : m \times 1$. The corresponding definitions are given as follows (assuming known mean):

(i) $\mathbf{X}_i = \mathbf{L}\mathbf{F}_i + \varepsilon_i + \boldsymbol{\delta}_i, \boldsymbol{\delta}_i = \mathbf{0}, \forall_{i \neq k}, \boldsymbol{\delta}_i : p \times 1$ and $\boldsymbol{\delta}_k{}' \boldsymbol{\delta}_k = 1$.

(ii) $\mathbf{X}_i = \mathbf{L}\left(\mathbf{F}_i + \boldsymbol{\delta}_i\right) + \varepsilon_i, \boldsymbol{\delta}_i = \mathbf{0}, \forall_{i \neq k}, \boldsymbol{\delta}_k : m \times 1$.

Contaminated data

Let $d_{\delta j}$, $j = 1, 2$ be the MDs for the two kinds of outliers, and let $\mathbf{S} = n^{-1} \sum_{i=1}^{n} (\mathbf{L}\mathbf{F}_i + \boldsymbol{\varepsilon}_i)(\mathbf{L}\mathbf{F}_i + \boldsymbol{\varepsilon}_i)'$ be the sample covariance matrix with known mean. To simplify calculations, let $\boldsymbol{\delta}_k \perp \mathbf{F}_i$ and $\boldsymbol{\delta}_k \perp \boldsymbol{\varepsilon}_i$ by assumption, so that they are not only uncorrelated but also orthogonal. The following expressions separate the contaminated and un-contaminated parts of the MD:

Contaminated data

Proposition 3

For case (i),

$$d_{\delta 1} = \begin{cases} d_{ii} & i \neq k \\ d_{kk} - \dfrac{n\left(1 - \mathbf{x}_k{}' n^{-1}\mathbf{S}^{-1}\boldsymbol{\delta}_k\right)^2}{1 + \boldsymbol{\delta}'_k n^{-1}\mathbf{S}^{-1}\boldsymbol{\delta}_k} + n & i = k \end{cases}$$

For case (ii),

$$d_{\delta 2} = \begin{cases} d_{ii} & i \neq k \\ d_{kk} - \dfrac{n\left(1 - \mathbf{x}_k{}' n^{-1}\mathbf{S}^{-1}\mathbf{L}\boldsymbol{\delta}_k\right)^2}{1 + \boldsymbol{\delta}_k{}'\mathbf{L}' n^{-1}\mathbf{S}^{-1}\mathbf{L}\boldsymbol{\delta}_k} + n. & i = k \end{cases}$$

where $d_{ii} = n\mathbf{x}_i{}'\mathbf{W}^{-1}\mathbf{x}_i = \mathbf{x}_i{}'\mathbf{S}^{-1}\mathbf{x}_i$.

Outline

Dataset

We implicate the results on a data setting which is consisted by the monthly closing stock prices from ten companies.
The companies are Exxon Mobil Corp., Chevron Corp., Apple Inc., General Electric Co., Wells Fargo Co., Procter Gamble Co., Microsoft Corp., Johnson & Johnson, Google Inc., Johnson Control Inc. They follow by the order of variables.

Dataset

We modified the closing prices into the monthly return rate (r),

$$r = \frac{y_t - y_{t-1}}{y_{t-1}},$$

where y_t stands for the closing price in the tth month. The whole analysis based on the returns on the stock prices.

Boxplot

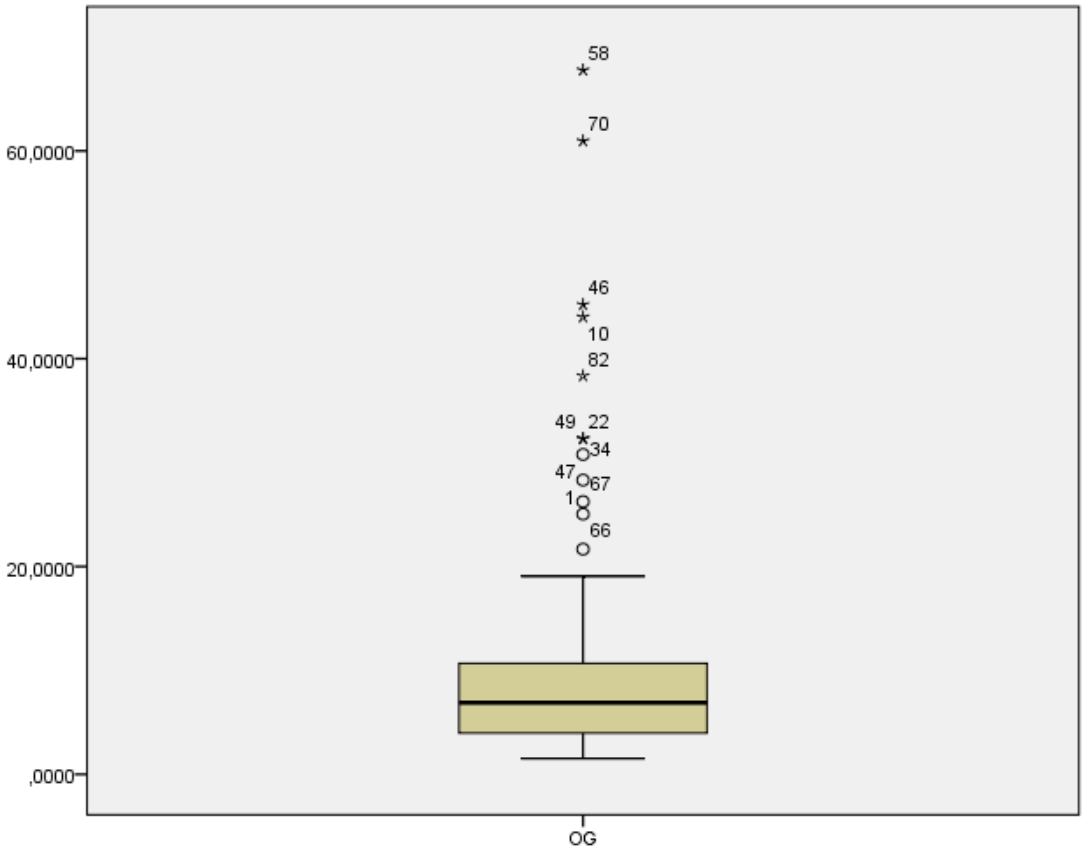


Figure: The boxplot of the original MDs.

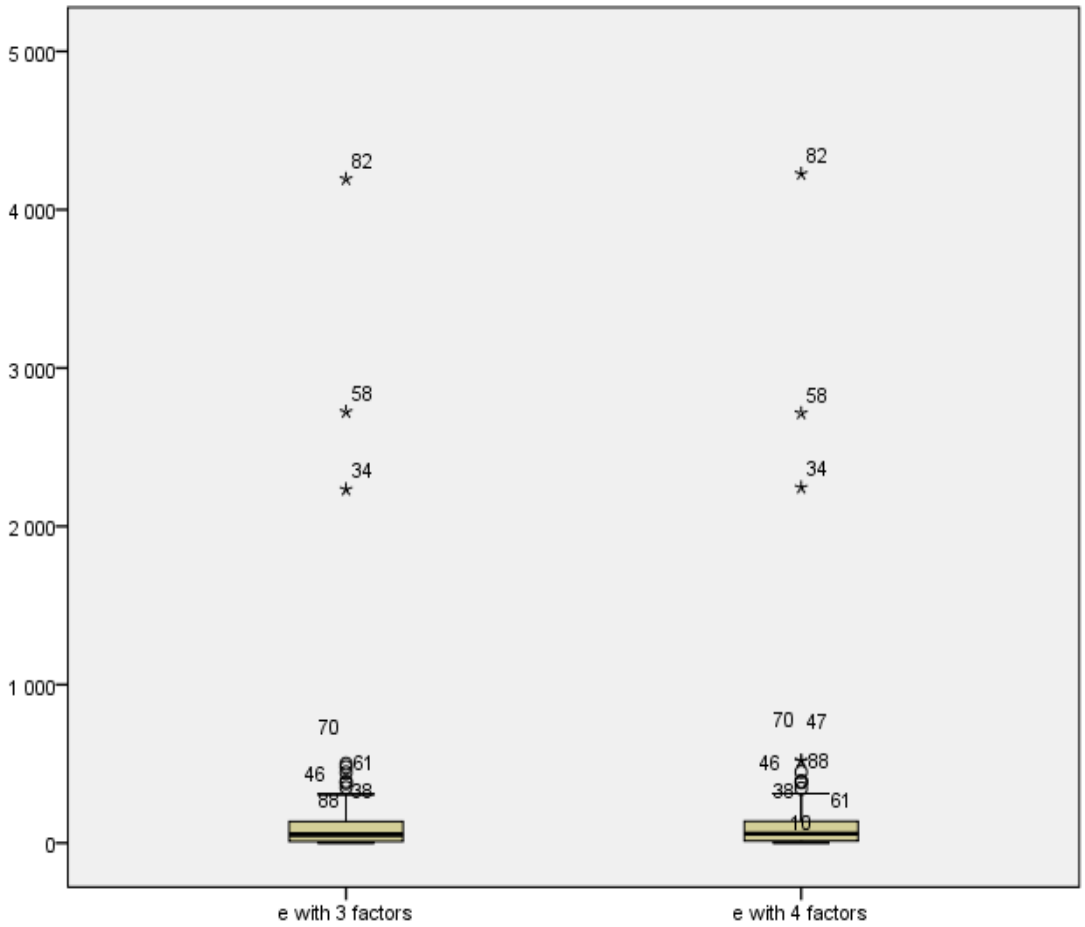


Figure: The boxplot of error terms.

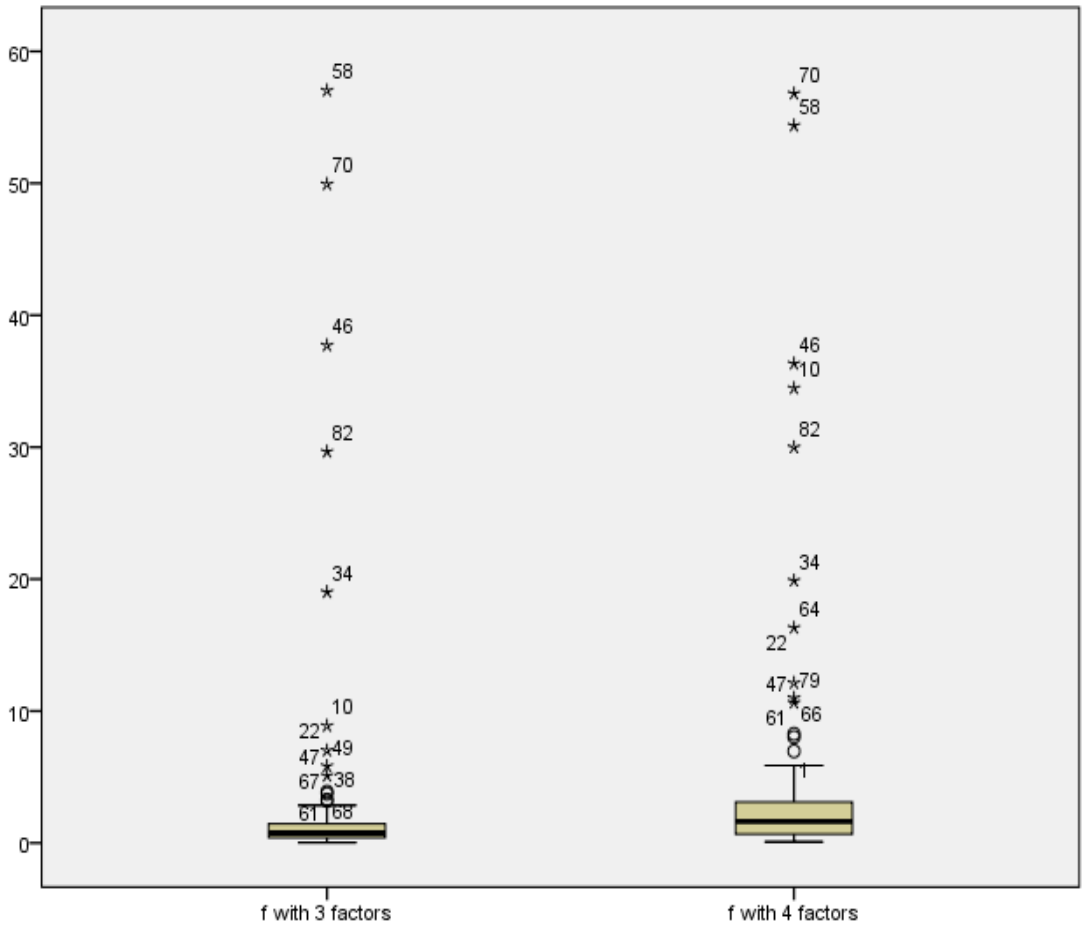


Figure: The boxplot of factor scores.

Bibliography I

Bai, J. (2003). Inferential theory for factor models of large dimensions, *Econometrica* **71**(1): 135–171.

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

Gorman, N., McConnell, I. and Lachmann, P. (1981). Characterisation of the third component of canine and feline complement, *Veterinary immunology and immunopathology* **2**(4): 309–320.

Gregory, A. W. and Head, A. C. (1999). Common and country-specific fluctuations in productivity, investment, and the current account, *Journal of Monetary Economics* **44**(3): 423–451.

Bibliography II

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.

Johnson, R. A. and Wichern, D. W. (2002). *Applied multivariate statistical analysis*, Vol. 5, Prentice hall Upper Saddle River, NJ.

Lewbel, A. (1991). The rank of demand systems: theory and nonparametric estimation, *Econometrica: Journal of the Econometric Society* pp. 711–730.

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.

Bibliography III

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., Kent, J. and Bibby, J. (1980). Multivariate analysis.

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

Ross, S. A. (1976). The arbitrage theory of capital asset pricing, *Journal of economic theory* **13**(3): 341–360.

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