

TESTS OF MULTINORMALITY BASED ON LOCATION AND SCATTER ESTIMATES

Kankainen, A., Taskinen, S. and Oja, H.

Department of Mathematics and Statistics, University of Jyväskylä

P.O. Box 35, Fin-40351 Jyväskylä, Finland.

February 9, 2004

Abstract

We consider tests of multinormality which are based on the Mahalanobis distance between two different location statistics and on a distance between two scatter matrix estimates. Let $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a random sample from an unknown continuous k -variate distribution. We thus wish to test the null hypothesis of multinormality, that is, the hypothesis that the unknown distribution where the data come from is a multivariate normal distribution $N_k(\boldsymbol{\mu}, \Sigma)$ with some (unknown) $\boldsymbol{\mu}$ and Σ .

Write $AX + \mathbf{b} = \{A\mathbf{x}_1 + \mathbf{b}, \dots, A\mathbf{x}_n + \mathbf{b}\}$ for a positive definite $k \times k$ matrix A and a k -vector \mathbf{b} . A vector valued statistic $\mathbf{T} = \mathbf{T}(X) \in \mathbb{R}^k$ is then called a **location statistic** if it is affine equivariant: $\mathbf{T}(AX + \mathbf{b}) = A\mathbf{T}(X) + \mathbf{b}$ for all choices A and \mathbf{b} . A matrix valued statistic $C = C(X)$ is said to be a **scatter statistic** if it positive definite symmetric $k \times k$ -matrix and affine invariant in the sense that $C(AX + \mathbf{b}) = AC(X)A^T$ for all positive definite A . In the following we assume that, under the hypothesis of multinormality $N_k(\boldsymbol{\mu}, \Sigma)$, all location and test statistics (\mathbf{T} and C) considered are weakly consistent estimates of $\boldsymbol{\mu}$ and Σ and $\sqrt{n}(\mathbf{T} - \boldsymbol{\mu})$ and $\sqrt{n} \text{vec}(C - \Sigma)$ have multivariate normal limiting distributions.

Let next \mathbf{T}_1 and \mathbf{T}_2 be two different location statistics, and C , C_1 and C_2 different scatter statistics. The test statistics considered are

then the Mahalanobis distance between the observed values of T_1 and T_2 ,

$$D_1(T_1, T_2) = \|T_1 - T_2\|_C^2,$$

and a distance between two scatter statistics C_1 and C_2 ,

$$D_2(C_1, C_2) = \|C_1^{-1}C_2 - I_k\|_F^2.$$

Here $\|\cdot\|_C$ is the Mahalanobis vector norm with a metric given by the scatter matrix C , and $\|\cdot\|_F$ is the Frobenius matrix norm ($\|A\|_F^2 = \text{Trace}(A^T A)$).

Asymptotic theory is developed to provide approximate null distributions of the test statistics as well as to consider their asymptotic efficiencies. Limiting efficiencies under certain contiguous sequence of contaminated normal distributions are found and the efficiencies of the new tests are compared to those of some classical tests. The efficiencies of the tests then naturally depend on choices of T_1 , T_2 , C , C_1 and C_2 ; good efficiencies are obtained if the distances are between robust and non-robust statistics. Simulations are used to compare finite sample efficiencies. The theory is illustrated by examples as well.