On the Suitability of Multivariate Normal Models for Statistical Inference Based on Traffic Measurements

F. Mata[1] (student), J.L. García-Dorado (student), J. Aracil

Universidad Autónoma de Madrid, Spain

The assumption of multivariate normality (MVN) underlies many commonly used multivariate statistical procedures. However, Monte Carlo studies have shown that the performance of many of such procedures is adversely affected when they are used with non-MVN data [1-3]. We present practical evidence that such procedures are suitable for analyzing network traffic measurements when they are sufficiently aggregated. To do so, we aggregate daily network load MRTG measurements, coming for different links within the Spanish academic network RedIRIS [4], in $p = 16$ non-overlapping time intervals that we model with a MVN distribution, being each day a realization. Our dataset contains more than 300 days per direction (incoming/outgoing for traffic sent to/by RedIRIS institutions) and per link (we analyzed more than 20 different links).

We first check for univariate normality of all the marginal distributions, as it is commonly known that if non-normality is indicated for one or more of the variables, MVN can be rejected [5, p. 133]. We follow the procedure of [6], where the linear correlation coefficient γ between the order statistics of the sample and the percentiles of a normal distribution is used to measure the goodness-of-fit to univariate normality. Our results are similar to those presented in [6], having γ > 0.9 in more than 80% of the cases, suggesting fairly Gaussian marginal distributions.

To verify the suitability of the MVN, we applied to those subpopulations where the univariate normality cannot be rejected for all the variables the well-known Mardia's tests for multivariate skewness and kurtosis [6], denoted by $b_{1,p}$ and $b_{2,p}$, respectively. For convenience of applying existing statistical tables, in practice are used the standardized forms $sb_{1,p} = \frac{n b_{1,p}}{6}$ and $sb_{2,p} = \frac{b_{2,p} - p(p+2)(n-1)/(n+1)}{\sqrt{8p(p+2)/n}}$, which converge in distribution to χ^2_d and $N(0,1)$ distributions, respectively, under the null hypothesis, being $d = p(p+1)(p+2)/6$. To avoid our procedure of being affected by non-stationarity (i.e. parameters of the MVN distribution changing with time), we divide our dataset into subpopulations of size $n = 20$, and apply Mardia's tests to such subpopulations. However, with so small subpopulations size, the distributions of the standardized statistics are bad estimates. We therefore use Monte Carlo simulation to approximate the critical values of the statistics. We run 100000 Monte Carlo simulations using MVN samples of size n. The resulting critical values for α = 0.01 are 732.4614 for $sb_{1,p}$ (unilateral test) and -0.0054 and 0.01 for $sb_{2,p}$ (bilateral test). The results show that approximately 6% of the times in the worst case for $sb_{1,p}$ and 8% for $sb_{2,p}$ the MVN is rejected, which supports the assumption of being fairly multivariate Gaussian.

Mardia [2] has shown that the size of the normal theory tests of mean vectors is more sensitive to skewness than to kurtosis, i.e. large values of $sb_{1,p}$ adversely affects the size of the tests based on the T^2 statistic, whereas in [7] he shown that the size of the likelihood ratio test of equality of covariance matrices when the parent population is non-normal is seriously influenced by kurtosis, i.e. large values of $|sb_{2,p}|$ significantly reduce the size of equality of covariance matrices tests. Our results show that normal theory tests of mean vectors and covariance matrices can be applied to network traffic measurements if they are properly preprocessed, because in less than 8% of the situations, the values of the $sb_{1,p}$and $sb_{2,p}$ statistics are big enough to adversely affect the sizes of the normal theory tests being applied to such network measurements.

[1] J.W. Hopkins and P.P.F. Clay, "Some Empirical Distributions of Bivariate T^2 and Homocedasty Criterion M Under Unequal Variance and Leptokurtosis," Journal of the American Statistical Association, 58, pp. 1048-1053, 1963.

[2] K.V. Mardia, "Assessment of Multinormality and the Robustness of Hotelling's T^2," Applied Statistics, 24, pp. 163-171, 1975.

[3] W.J. Conover and R.L. Iman, "The Rank Transformation as a Method of Discrimination with Some Examples," Communications in Statistics-Theory and Methods, A9, pp. 456-487, 1980.

[4] Spanish National Resarch and Education Network RedIRIS, http://www.rediris.es/index.php.en.

[5] R.A. Johnson and D.W. Wichern, "Applied Multivariate Statistical Analysis," (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall.

[6] K.V. Mardia, "Measures of Multivariate Skewness and Kurtosis with Applications," Biometrika 57, pp. 519-530, 1970.

[7] K.V. Mardia, "Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies," Shankyā: The Indian Journal of Statistics, Series B, 36, pp. 115-128, 1974.

[1] Email contact address: felipe.mata@uam.es