**Executive summary of the research work done by SAJESH T A for the period of July 2014 to July 2016.**

UGC Reference No. : MRP(S) – 0604/13-14/KLCA019/UGC-SWRO

Title of research project: ***Robust Multivariate Analysis of Variance to Test the Equality of Mean Vectors of Multiple Groups***

One-way multivariate analysis of variance (MANOVA) deals with testing the null hypothesis of equal mean vectors across the $g $considered groups. The setup is similar to that of the one-way univariate analysis of variance (ANOVA) but the inter-correlations of the independent variables are taken into account, i.e. the variables are considered multivariate. Under the classical assumptions that all groups arise from multivariate normal distributions, many test statistics are discussed in the literature, one of the most widely used being the likelihood ratio test. This test statistic is better known as Wilk’s Lambda in MANOVA. The Wilks' Lambda is reported as part of the test output in almost all statistical packages. However, this measure which uses the classical normal theory as well as the inference based on it can be adversely affected by outliers present in the data.

Let$ x\_{k1}, x\_{k2}, …, x\_{kn\_{k}} $be$ n\_{k} $independent and identically distributed $p$-dimensional observations from a continuous $p$-variate distribution with distribution function$ F\_{k}\left(u\right) $where$ k=1, 2, …, g$ and the number of groups$ g\geq 2$. If all$ g $distributions are exactly the same but only their locations differ we have

$$F\_{k}\left(u\right)=F\left(u-μ\_{k}\right)$$

Then the hypothesis we want to test is that all $ F\_{k} $are identical, i.e.

$$H\_{0}:μ\_{1}=μ\_{2}=…=μ\_{g}$$

against the alternative hypothesis

 $H\_{a}:μ\_{i}\ne μ\_{j} $for at least one $i\ne j$

Under the classical assumptions that all groups arise from multivariate normal distributions, the most widely used test statistic is the Wilk’s Lambda (the likelihood ratio test). The Wilks' Lambda statistic is the ratio of the within generalized dispersion to the total generalized dispersion. The within generalized dispersion is the determinant of the within-group sums of squares and cross-products matrix $W $and the total generalized dispersion is the determinant of the total sums of squares and cross-product matrix $T$ . The statistic

$$ Λ\_{Wilks}=\frac{det⁡(W)}{det⁡(T)} (1)$$

takes values between zero and one ( here $det⁡(A) $means the determinant of$A$). The non-robustness of the Wilk’s Lambda statistic in the context of variable selection in linear discriminant analysis was demonstrated in Todorov (2007).

Therefore we propose to use robust estimators instead of the classical ones for computing Wilk’s Lambda statistic. The non-robustness of the normal theory based test statistic has led many other authors also to search for alternatives. For this purpose we will use the Kurtosis estimator of Pena and Prieto (2001) which is a highly robust estimator of location and scatter. Since the distribution of the robust Wilk’s Lambda statistic based on kurtosis differs from the classical one it is necessary to find a good approximation for this distribution.

In order to obtain a robust procedure with high breakdown point for inference about the means in the one-way MANOVA model we construct a robust version of the Wilk’s Lambda statistic by replacing the classical estimators by the kurtosis estimators. The Kurtosis estimator is introduced by Pena and Prieto (2001) looks for a subset of $h$observations which optimize (maximize or minimize) the kurtosis coefficient. The method is affine equivariant, and it shows a very satisfactory practical performance, especially for large sample space dimensions and concentrated contamination. The method also produces good robust estimates for the covariance matrix, with low bias. We construct an approximate distribution based on a Monte Carlo study and examined its accuracy.

We start by finding estimates of the group means $m\_{k}^{0} $and the common covariance matrix $C\_{kur} $based on the kurtosis estimate. The proposed estimators become

$$ \begin{matrix}m\_{k}^{0}=\frac{1}{\left|U\right|} \sum\_{iϵU}^{}x\_{ik } ,\\C\_{kur}=\frac{1}{(\left|U\right|-1)}\sum\_{iϵU}^{}\left(x\_{ik}-m\_{k}^{0}\right)(x\_{ik}-m\_{k}^{0})' ,\end{matrix} (2)$$

where $U $is the set of all observations is not labeled as outliers, $\left|U\right|$ denotes the number of observations in this set. Using this obtained estimates$ m\_{k} $and$ C\_{kur}$ in $\left(2\right) $we can calculate the robust distances (Pena and Prieto, 2001) as

$$ RD\_{ik}=\sqrt{\left(x\_{ik}-m\_{k}^{0}\right)^{'}C\_{kur}\left(x\_{ik}-m\_{k}^{0}\right)} (3)$$

with these robust distances we can define a weight for each observation
$x\_{ik} , i=1,…, n\_{k} $and $k=1, …, g $by setting the weight to 1 if the corresponding robust distance is less or equal to $\sqrt{χ\_{p, 0.99}^{2}} $and to 0 otherwise, i.e.,

 $w\_{ik}=\left\{\begin{matrix}1& RD\_{ik} \leq \sqrt{χ\_{p, 0.99}^{2}}\\0&otherwise\end{matrix}\right.$ (4)

with these weights we can calculate the final estimates, namely the group means$ m\_{k}$, the within-groups sum of squares and cross-products matrix$ W\_{R}$, the between-groups sum of squares and cross-products matrix $B\_{R} $and the total sum of squares and cross-products matrix $T\_{R}=W\_{R}+B\_{R} $which are necessary for constructing the robust Wilks' Lambda $Λ\_{R} $statistic as defined in equation $(1)$.

$$m\_{k}={(\sum\_{i=1}^{n\_{k}}w\_{ik}x\_{ik})}/{v\_{k}} , m={(\sum\_{k=1}^{g}v\_{k}m\_{k})}/{v}$$

$$W\_{R}=\sum\_{k=1}^{g}\sum\_{i=1}^{n\_{k}}w\_{ik}\left(x\_{ik}-m\_{k}\right)\left(x\_{ik}-m\_{k}\right)^{t}$$

$B\_{R}=\sum\_{k=1}^{g}v\_{k}\left(m\_{k}-m\right)\left(m\_{k}-m\right)^{t}$ (5)

$$T\_{R}=\sum\_{k=1}^{g}\sum\_{i=1}^{n\_{k}}w\_{ik}\left(x\_{ik}-m\right)\left(x\_{ik}-m\right)^{t}=W\_{R}+B\_{R}$$

where$ v\_{k} $are the sums of the weights group$ k $for$ k=1,…, g $and $v $is the total sum of weights:

$$v\_{k}=\sum\_{i=1}^{n\_{k}}w\_{ik} and v=\sum\_{i=1}^{n\_{k}}v\_{k}$$

Substituting these estimates of the matrices$ W $and$ T $into equation $\left(1\right) $we obtain a robust version of the test statistic$ Λ $given by

$$ Λ\_{R}=\frac{det⁡(W\_{R})}{det⁡(T\_{R})} \left(6\right)$$

For computing the kurtosis and related estimators the kurtosis algorithm of Pena and Prieto will be used as implemented in the Matlab software. The approximate distribution of the proposed robust statistic was derived using simulations and its fitting is examined using QQ – plots.

Further, Monte Carlo simulations were used to investigate the efficiencies of the proposed method and the level of significance and power of the proposed robust MANOVA are compared with that of classical MANOVA. The results of simulation study says that the estimated significance level in uncontaminated data using this robust method is approximately equal to the actual size, especially in high dimensional data sets. The size -power curve proposed by Davidson and McKinnon used here to compare the power of the proposed robust method with the classical one. Curve shows that the calculated power of the robust method is slightly below those obtained in classical. In most cases it is equal to the classical one. That means there is only an acceptable level of power loss is occurred for the proposed method in uncontaminated data sets, which shows the advantage of the robust test over classical test. In contaminated situation the robust method is more suitable than the classical one.

Additionally, the proposed robust test is applied and investigated its performance in a real life benchmark data (Oslo-transect data). The result shows that the proposed Robust MANOVA is much less affected by the presence of outliers compared to the classical MANOVA. That is, the proposed robust MANOVA technique performs well under contaminated as well as un-contaminated situations and is a better alternative to classical test for the data containing contaminated observations. Thus we can apply this proposed Robust MANOVA in real life data sets, as they should necessarily have outliers.

**Research papers presented/published/ Accepted for publication/ Communicated for publication:**

1. Presented in the ***International Conference on "New Horizons in Statistical Modeling and Applications (NHSMA – 2015)*** organised by the Department of Statistics, Presidency College, Chennai, Tamilnadu, India during Feb 27-28, 2015.
2. Robust Statistic for one way MANOVA (2016) *Journal of Engineering Technology and Applied Sciences,* **Communicated.**

**References**

1. T.W. Anderson, *An Introduction Multivariate Statistical Analysis*, John Wiley, 1958
2. R. Davidson and J. McKinnon, “Graphical methods for investigating the size and power of hypothesis tests”, *Technical report*, *The Manchester School*, 1998.
3. S. Gelper andC. Croux, “Multivariate out-of-sample tests for granger causality”, *Computational Statistics and Data Analysis*, Vol. 51, pp. 3319 -3329.
4. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall International, 2002
5. D. Pena and F.J. Prieto, “Multivariate Outlier Detection and Robust Covariance Matrix Estimation”, *Technometrics*, Vol. 43, pp. 286-300, 2001
6. C. Reimann, A. Arnoldussen, R. Boyd, T.E. Finne, F. Koller, O. Nordgullen and P. Englmair, “Element contents in leaves of four plant species (birch, mountain ash, fern and spruce) along anthropogenic and geogenic concentration gradients”, *Science of the Total Environment*, Vol. 377, pp. 416.433, 2007
7. C. Siani and C. de Peretti, “Analysing the performance of bootstrap neural tests for conditional heteroskedasticity in ARCH-M models”, Computatational Statistics and Data Analysis, Vol.51, no. 5, pp. 2442-2460, 2007
8. V. Todorov, “Robust selection of variables in linear discriminant analysis”, *Statistical Methods and Applications*, Vol. 15, pp. 395.407, 2007.