**tscvh R Package: Computational of the two samples test on microarray-sequencing data**

**Rohmatul Fajriyah¹, Dedi Rosadi²**

¹Department of Statistics, Universitas Islam Indonesia, Yogyakarta, Indonesia  
²Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia

E-mail: rfajriyah@uii.ac.id

**Abstract.** We present a new R package, a tscvh (two samples cross-variance homogeneity), as we called it. This package is a software of the cross-variance statistical test which has been proposed and introduced by Fajriyah ([3] and [4]), based on the cross-variance concept. The test can be used as an alternative test for the significance difference between two means when sample size is small, the situation which is usually appeared in the bioinformatics research. Based on its statistical distribution, the p-value can be also provided. The package is built under a homogeneity of variance between samples.

1. Introduction

The two sample tests problem in bioinformatics study actually based on the sample size in microarray experiment which is usually small. Therefore, to decide that there is a significant difference between two means by implementing the parametric t test will create a problem. As L"onnsted and Speed [7] pointed out the large mean and large t are caused by the outlier and small denominator SE respectively.

Tusher [13], Efron [2], Smyth [12] have modified the two samples t-statistic. If the samples come from normal distribution, in case $\sigma_x^2 = \sigma_y^2$, the test statistic has $t_{m+n-2}$ distribution under $H_0$ (see e.g. [10]). In case $\sigma_x^2 \neq \sigma_y^2$ then we use the approximation distribution as in [15] and [14].

Other the `t.test` function in R, those concepts have been implemented in the software packages, for instance `limma`, `SpatialPack`, `geoR`, `SScore`, and `Samr`.

Fajriyah [4] introduces the cross-variance concept and present a new approach to test the mean difference between two samples based on it. It is stated that the cross-variance is different from the cross-covariance, where cross-covariance is a function that gives the covariance of one process with the other at pairs of time points [16] and cross-variance is defined as using the second sample average to compute the first sample variance vice versa.

The purpose of the paper is to provide a general overview of the tscvh package (version 1.0), which is based on the homogeneity of the cross-variance test. The paper is organized as follows. In Section 2 the cross-covariance concept and test are introduced. Section 3 introduces the package main function and the implementation by providing some examples on how to use...
the package and the data come from genomics study. We end by comparing \texttt{tscvh} to other R packages for two samples test and describe future plans for extending the package for the heterogeneity cross-variances test. Finally, section 4 is the concluding remarks.

2. Cross variance concept and the proposed test

2.1. The Cross-Variance Concept

**Definition 2.1** Suppose we have two independent samples, $X_i$ and $Y_j; i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. Their sample mean and variance are denoted by $\bar{X}, \bar{Y}$ and $V_x, V_y$. Let

$$V_x^* = \frac{m}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})^2$$
$$V_y^* = \frac{n}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

be the cross-variance for each sample $X$ and $Y$ respectively. The cross-variance between samples $X$ and $Y$ is defined as

$$T = \frac{V_x^* + V_y^*}{2}, \quad (1)$$

where $V_x^* = \frac{V_x}{V_x^*}$, $V_y^* = \frac{V_y}{V_y^*}$.

The $T$ can also be written as

$$T = \frac{V_x}{2V_x + 2m(\bar{Y} - \bar{X})^2} + \frac{V_y}{2V_y + 2n(\bar{Y} - \bar{X})^2} = Z_1 + Z_2 \quad (2)$$

In what follows, we assume that

(i) the sample sizes are equal
(ii) $X_i$ and $Y_i$ are i.i.d. normally distributed with unknown means and known variances $\sigma_x^2, \sigma_y^2$.

Based on Rao [10] the Equation (2) can be written as

$$T = \frac{U}{2U + 2abV} + \frac{S}{2S + 2bcV} \quad (3)$$

with

$$U = \frac{(n-1)V_x}{\sigma_x^2}, \quad S = \frac{(n-1)V_y}{\sigma_y^2}, \quad V = \frac{n(Y - \bar{X})^2}{(\sigma_x^2 + \sigma_y^2)}$$

and

$$a = \frac{1}{\sigma_x^2}, \quad b = (\sigma_x^2 + \sigma_y^2)^2, \quad c = \frac{1}{\sigma_y^2}.$$

To compute the distribution of $T$ in Equation (3), it can be done by considering the fact that (see e.g. [10]) $U, V$ and $S$ are independent.

2.2. The Cross-variance Test

Under normality assumption of $X$ and $Y$ then $U, V$ and $S$ are independent, where $V$ is $\chi^2_{(1)}$ distributed and $U, S$ are $\chi^2_{(n-1)}$ distributed.
In the case of $\sigma^2 = \sigma^2_n$, we estimate the $V_x$ and $V_y$ by the least square estimator of the pooled variance $S^2_p = \frac{V_x + V_y}{2}$, therefore the Equations (2) becomes

$$T^* = \frac{U^*}{U^* + 4V^*}$$

(4)

where $U^* = \frac{(n-1)(V_x + V_y)}{\sigma^2^2}$, $V^* = \frac{n(Y - \bar{X})^2}{2\sigma^2}$, and $U^*$ and $V^*$ are independents.

The pdf of $T^*$ is derived from the ratio of linear combination of chi-square random variables [9]. First, let $Y = 1 + 4V^*/U^*$, where $V^*$ is distributed $\chi^2_{(1)}$ and $U^*$ is distributed $\chi^2_{(2n-1)}$. Second, the pdf of $T^*$ is computed by taking $T^* = \frac{1}{V^*}$. In the following computation, the chi-square distribution will be represented as the Gamma distribution.

Suppose $G = \frac{1}{V^*}$ and if we take $U^* = H$, then we get $V^* = GH$. Further we got the Jacobian of this transformation variable random is $h$. Because $V^*$ and $U^*$ are independents, then the joint probability function of $G$ and $H$ is

$$f_{G,H}(g,h) = f_{V^*}(gh,h).h$$

(5)

where

$$f_{V^*}(gh,h) = f_{V^*}(gh).f_{U^*}(h), f_{V^*}(gh) = \frac{(gh)^{\alpha_1-1}e^{-\frac{gh}{\beta_1^2}}}{\beta_1^2 \Gamma(\alpha_1)}$$

and

$$f_{U^*}(h) = \frac{(h)^{\alpha_2-1}e^{-\frac{h}{\beta_2^2}}}{\beta_2^2 \Gamma(\alpha_2)}.$$ Therefore

$$f_{G,H}(g,h) = \frac{g^{\alpha_1-1}h^{\alpha_1+\alpha_2-1}e^{-\frac{h(\alpha_1+\alpha_2)}{\beta_1^2+\beta_2^2}}}{\beta_1^\alpha_1+\alpha_2 \Gamma(\alpha_1) \Gamma(\alpha_2)}$$

(6)

The pdf of $g$ then

$$f_G(g) = \int_0^\infty f_{G,H}(g,h)dh = \frac{g^{\frac{1}{2}-1}}{B\left(\frac{1}{2},n-1\right)(1 + g)^{n-\frac{1}{2}}}$$

(7)

$G$ is distributed beta of second kind.

The next step is determining the distribution of $Y$. We define $Y = 1 + 4G$ and by using the transformation random variable, where $G = \frac{Y-1}{4}$, the pdf of $Y$ is computed as follow

$$f_Y = \frac{4^{n-1}(y-1)^{\frac{1}{2}-1}}{B\left(\frac{1}{2},n-1\right)(y+3)^{n-\frac{1}{2}}}, \quad 1 \leq y \leq \infty$$

(8)

The pdf of $T^*$ is

$$f_{T^*}(t^*) = \frac{4^{n-1}(1 + \frac{1}{t^*})^{\frac{1}{2}-1}}{B\left(\frac{1}{2},n-1\right)(3 + \frac{1}{t^*})^{n-\frac{1}{2}} t^{t^*/2}}$$

$$= \frac{4^{n-1}t^*^{n-2}(1 - t^*)^{\frac{1}{2}-1}}{B\left(\frac{1}{2},n-1\right)(1 + 3t^*)^{n-\frac{1}{2}}}, \quad 0 \leq t^* \leq 1$$

(9)

Furthermore the cdf of $T^*$ analytically is computed as

$$F_{T^*}(t^*_0) = \int_0^{t^*_0} f_{T^*}(t^*)dt^* = \frac{4^{n-1}}{B\left(\frac{1}{2},n-1\right)} \times$$

$$\left[ \sum_{k=0}^{\infty} (-3)^k \frac{(n+k - \frac{3}{2})B\left(x,n+k-1,\frac{1}{2}\right)}{k!} \right]$$

(10)
where $B(x,\alpha,\beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt$ is the incomplete beta function see e.g. [5].

We reject the null hypothesis of the equality of mean of two independent samples if $T^* < t^*_{0,\alpha}$ or $P(t^* < T_0^*) = p$-value $\leq P(t^* < t^*_{0,\alpha}) = \alpha$.

**Theorem 2.2** Let $\sigma^2_x = \sigma^2_y$ and the $V_x$ and $V_y$ is estimated by the pooled variance $S_p^2 = \frac{V_x + V_y}{2}$, then $T^*$ in equation (4) has a pdf

$$f_{T^*}(t^*) = \frac{4^{n-1}n^{n-2}(1-t^*)^{\frac{1}{2}-1}}{B\left(\frac{1}{2}, n-1\right) (1 + 3t^*)^{n-\frac{3}{2}}}, \quad 0 \leq t^* \leq 1 \quad (11)$$

Equation (11) is the probability function for the cross-variance test in the **tscvh** package.

3. The package

3.1. Main functions of the package

There are 4 main functions in the **tscvh** package, namely

(i) cumr: This function is a cumulative probability density function for the Tstar random variable. It provides a cumulative probability density function value of the Tstar

(ii) fr: This function is a density function for the Tstar random variable. It provides a probability value of the Tstar

(iii) quantr: This function provides a value for the qth quantile of the Tstar

(iv) Tstf: This function is a main function for doing the first cross variance test. A test for testing the difference between two samples

3.2. The implementation

3.2.1. Installing the package

- **tscvh_1.0.tar.gz** is the tarball package
- `install.packages('tscvh')`
- **R CMD install ~/Downloads/tscvh_1.0.tar.gz**

3.2.2. The data

The data set we use in this paper is the data used in Corrada Bravo et al. [1]. We take the second probeset, which is the 238493_at as a microarray data example for the package. The data comes from the **antiProfilesData** package in Bioconductor. In this data set, there are 5339 probesets (genes) and 68 samples. It contains a colon normal tissue and cancer samples, from different tissues: colon, rectum mucosa and sigmoid colon mucosa. This data set is from the Affymetrix platform.

3.2.3. How to use the package

- **Step 1.** After installing the package then its library has to be called before using it. `library(tscvh)`
- **Step 2.** Write or load the data

**Data set 1**

```
 u=(0.962552767, 0.849068175, -0.006143108, -1.797205349, 1.885255606, 
 0.914242952, -1.171969005, -0.700293945, -0.611782849, -1.042330679, 
 -0.461310875, 4.722690836, 1.597091975, 0.640279134, 0.684615816 )
 v=(1.86832946, 0.14578081, 0.49505374, 1.75326037, 0.91047225, 
 1.94170092, 3.8472690836, 1.597091975, 0.640279134, 0.684615816, 
 -0.06117995, -0.72727366, 1.07775479, 1.42560956, 2.21901185, 2.17089092)
```
Data set 2

\[ x = (2.8538601, 0.5906237, 0.6176468, 0.8219640, 2.0970868, 0.1187993, 1.5110154, -0.1979436) \]

\[ y = (1.86832946, 0.14578081, 0.49505374, 1.75326037, 0.91047225, 1.94170092, 3.84792213, 0.32982953, -0.06117995, -0.72727366, 1.07775479, 1.42560956, 2.21901185, 0.96661355, 2.17089092) \]

- **Step 3.** Called the \( Tstf \) function

\[
Tstf(d1=x,d2=y,alpha=0.05) \\
Tstf(d1=u,d2=v,alpha=0.05)
\]

The results can be seen at Table 1 below.

| Sample | \( t \) test p-value | \( t \) test Decision | proposed test p-value | proposed test Decision |
|--------|---------------------|----------------------|----------------------|----------------------|
| Data set 1 (u,v) | 0.1299 | Accept | 0.129 | Accept |
| Data set 2 (x,y), max | 0.7173 | Accept | 0.715 | Accept |
| Data set 2 (x,y), ave | 0.7173 | Accept | 0.716 | Accept |

3.2.4. *Comparison to the \texttt{t.test}*. Based on Table 1, we can see that the decision from the \texttt{tscvh} package is equal to the \texttt{t.test}. The slightly different is in the p-value only. In the \texttt{tscvh} package we use 3 digits after comma, therefore for the Data 1, the package provides the p-value as 0.129. For the Data 2, where the sample size is different, then based on Fajriyah [4] the package gives you two best options: maximum and average and from both of them, the decisions are equals.

The two examples give the insight on how to use the package and show that the decision are equal to the \texttt{t.test}. Therefore, this package can be considered as an alternative package to compare means of two samples, in case the variance of the two samples is homogeneous. In the example above, we only show you the two-sided hypothesis testing. The one-sided right and left can be implemented in the package as well.

4. CONCLUDING REMARKS

We introduce the concept of cross-variance and derive a method to test the statistically significance of the means between two samples where the variance is homogeneous. The \texttt{tscvh} package is built with the R program language under the S3 class. It is intended for the bioinformatics data, where the replication is usually not many, but it can be used for data from other fields as well. In this paper we compare the result based on the available method, \texttt{t.test}.

In the future the package will be developed under the S4 class and included the heterogeneity case.

**Acknowledgement.** The authors would like to thanks to the Indonesian Government for its financial support of the Research Grant No. 012/DirDPPM/70/DPPM/ PascaDoktor-KEMENRISTEKDIKTI/IV/2017, under the Ministry of Research, Technology and Higher Education, Republic of Indonesia and Universitas Islam Indonesia, Jogjakarta, Indonesia.
References

[1] Bravo HC, McCall M and Irizarry RA 2017 antiProfilesData: Normal colon and cancer preprocessed affy data for antiProfile building R package version 1.14.0
[2] Efron B, Tibshirani R, Storey J D, and Tusher V 2001 "Empirical Bayes Analysis of a Microarray Experiment" Journal of the American Statistical Association, 96 1151-60
[3] Fajriyah R 2014 Microarray Data Analysis: Background Correction and Differentially Expressed Genes, Dissertation. Graz, Austria: Graz University of Technology
[4] Fajriyah R 2016 Introducing a cross variance concept and its application, 12th ICMSA Proceeding, IEEE CFP 16 H58 ART
[5] Gradshteyn I S and Ryzhik I M 2007 Table of integrals, series, and products. 7th edition (New York: Academic Press)
[6] Kennedy Richard E, Kerns Robnet T, Kong X, Archer, K J, and Miles, M F 2006 "SScore: an R package for detecting differential gene expression without gene expression summaries", Bioinformatics 22 127274
[7] Lönstedt I and Speed T P 2002 Replicated Microarray Data, Statistica Sinica, 12 31-46
[8] Phipson B, Lee S, Majewski I J, Alexander W S, and Smyth, G K 2016 Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression”, Annals of Applied Statistics 10 94663
[9] Provost S and Rudik E M 1994 The exact density function of the ratio of two dependent linear combinations of chi-square variables Annals of the Institute of Statistical Mathematics, 46 5571
[10] Rao C R 1971 Linear statistical inference and its applications. Second edition (New York: Wiley)
[11] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK 2015 “limma powers differential expression analyses for RNA-sequencing and microarray studies”, Nucleic Acids Research 43 e47
[12] Smyth G K 2004 "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments", Statistical Applications in Genetics and Molecular Biology 3 article 3
[13] Tusher V, Tibshirani R and Chu G 2001 "Significance analysis of microarrays applied to the ionizing radiation response", PNAS 98 5116-21
[14] Welch B L 1938 The significance of the difference between two means when the Population variances are unequal Biometrika 29 350-362
[15] Welch B L 1947 The generalization of students problem when several different population variances are involved Biometrika 34 2835
[16] https://en.wikipedia.org/wiki/Cross-covariance, last access November 8, 2017. At 12.55 PM.