CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva and Carolina Marchant

Volume 11 Number 1
April 2020
ISSN: 0718-7912 (print)
ISSN: 0718-7920 (online)

Published by the
Chilean Statistical Society
SOCHE
SOCIEDAD CHILENA DE ESTADÍSTICA

Aims

The Chilean Journal of Statistics (ChJS) is an official publication of the Chilean Statistical Society (`www.soche.cl`). The ChJS takes the place of *Revista de la Sociedad Chilena de Estadística*, which was published from 1984 to 2000.

The ChJS is an international scientific forum strongly committed to gender equality, open access of publications and data, and the new era of information. The ChJS covers a broad range of topics in statistics, data science, data mining, artificial intelligence, and big data, including research, survey and teaching articles, reviews, and material for statistical discussion. In particular, the ChJS considers timely articles organized into the following sections: Theory and methods, computation, simulation, applications and case studies, education and teaching, development, evaluation, review, and validation of statistical software and algorithms, review articles, letters to the editors.

The ChJS editorial board plans to publish one volume per year, with two issues in each volume. On some occasions, certain events or topics may be published in one or more special issues prepared by a guest editor.

Contents

Carolina Marchant and Víctor Leiva
Starting a new decade of the Chilean Journal of Statistics in COVID-19 pandemic times with new editors-in-chief 1

Luz Milena Zea Fernandez and Thiago A.N. de Andrade
The erf-G family: new unconditioned and log-linear regression models 3

Thodur Parthasarathy Sripriya, Mamandur Rangaswamy Srinivasan, and Meenakshisundaram Subbiah
Detecting outliers in $I \times J$ tables through the level of susceptibility 25

Adolphus Wagala
A likelihood ratio test for correlated paired multivariate samples 41

Josmar Mazucheli, Sudeep R. Bapat, and André Felipe B. Menezes
A new one-parameter unit-Lindley distribution 53

Chilean Journal of Statistics
Vol. 11, No. 1, April 2020, 41–51

Multivariate statistical inference
Research Paper

A likelihood ratio test for correlated paired multivariate samples

Adolphus Wagala*

[1]Departamento de Probabilidad y Estadística, Centro de Investigación en Matemáticas, AC, Guanajuato, Mexico

(Received: 02 August 2019 · Accepted in final form: 20 April 2020)

Abstract

Many laboratory experiments in the fields of biological sciences usually involve two main groups say the healthy and infected subjects. In one of these kind of experiments, each specimen from each group can be divided in two portions; one portion is stimulated while the other remains unstimulated. Consequently resulting into two main groups with paired measurements that are correlated. For all the groups, p genes are measured for expression. The stimulation in this case can be done by introducing a known infection causing micro-organism like the group A streptococcus which is usually associated with the acute rheumatic fever. An important question in such experiment would be to statistically test for the differences in the differences in means for the healthy and the infected groups. That is, the difference in the means of the healthy group (stimulated and unstimulated) is tested against the difference in the means of the infected (stimulated and unstimulated) group. In this paper, a likelihood ratio test statistic is developed for such kind of problems. The developed statistics and the Hotelling T^2 statistic are both applied to the data are simulated from real biological situations and their performances are compared. The simulated data exhibit the correlation structure similar to that of real biological data obtained from experiments involving the milliplex analyst biomarker data sets. The results indicate that the proposed test statistic give the same conclusions for the hypotheses tested as those of the Hotelling T^2 test. However, the proposed test is intuitively more appealing since it takes care of the correlations between the pairs in the data. The simulation study confirms that the test statistics follow a chi-square distribution. This research contributes a theoretical analysis of paired correlated samples motivated by a practical problem for which the existing statistical methods in use have seldomly taken into account the correlation structure of the data.

Keywords: Correlated pairs· Likelihood ratio test · Multivariate samples

Mathematics Subject Classification: Primary62H15 · Secondary 62J15.

1. Introduction

Consider an experiment involving two groups of subjects namely the healthy (H) and the infected (I) donors. Each group is further divided into two sub-groups whereby one subgroup is stimulated using some infection causing organism for example group A streptococcus (GAS) which causes the acute rheumatic fever (ARF). The other subgroup remains unstimulated. As a result, we end up with paired samples for the H and also another paired

*Corresponding author. Email:adolphus.wagala@cimat.mx

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

samples for the I, resulting into two groups with paired measurements that are correlated. The samples from all these groups are then sequenced to measure the expression levels for the p genes under consideration. The genes whose expression levels are measured are the same for all the paired groups. It is expected that the GAS stimulation of H and I subjects can help in understanding how the GAS affects the H and I subjects thereby possibly able to identify the biomarkers associated with the ARF. The effect of GAS stimulation/unstimulation can lead to changes in the genes with regards to up or down regulations or no change. Assuming that the sample sizes for H and I subjects are m and k respectively and that p genes are considered in the experiment. It is easy to see that the m paired measurements for the H are correlated and at the same time the k paired measurements for the I are correlated while H and I groups are independent. Furthermore, since the genes usually act in a group, the p genes are expected to be correlated.

The main goal therefore is to develop a statistical framework for testing the changes in expression levels in the different sets of genes between the two main groups which have the properties of independence between them but paired correlation within the subjects. The observations are independent and identically distribute (IID). We use the well known likelihood ratio theory to formally derive a new test for formally testing for the difference in the differences of the mean expression levels for the healthy and infected subjects.

The remainder of this paper is organized as follows, Section 2 gives a brief review of the likelihood ratio testing. The proposed likelihood ratio test statistic for multivariate paired, correlated samples is presented in Section 3 while the simulation study is given in 4. Finally, the summary and conclusions are given in Section 5.

2. The likelihood ratio test

The theory of the likelihood ratio test (LRT) is well understood and has been utilized extensively in the field of statistical inference. Most standard multivariate statistics books like for example Anderson (2003), Seber (2004), Mardia et al. (1980), Johnson and Wichern (2007) to mention but a few, contain comprehensive treatment of this subject matter.

To review, the LRT, we start by letting $\boldsymbol{\theta}$ be the parameter vector for the likelihood function $L(\boldsymbol{\theta})$ with observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with a density function given by $f(\boldsymbol{x}; \boldsymbol{\theta})$. If the parameter space is given by $\boldsymbol{\Theta}$ and suppose that we want to test the null hypothesis $H_o : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ where $\boldsymbol{\Theta}_0$ is a subset of $\boldsymbol{\Theta}$. The parameter space $\boldsymbol{\theta}$ is unconstrained while $\boldsymbol{\theta}_0$ is constrained. The LRT statistic is given by

$$\Lambda = \frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_o} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta})}.$$

The null hypothesis H_o is rejected when $\Lambda < C$, where C is a critical value depending on the type-I error. The LRT has good power properties asymptotically and usually is as good or better than many other test statistics Seber (2004). The LRT statistic under general conditions and with large samples are approximately $\chi^2_{(d)}$ distributed where d is the degree of freedom which in general is given by the total number of variables under consideration. The LRT is given by

$$-2\text{Log}\Lambda = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \{-2\text{Log L}(\boldsymbol{\theta})\} - \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \{-2\text{Log L}(\boldsymbol{\theta})\}.$$

Some common problems that have been tackled in the said standard multivariate statistics analysis setting with regards to the LRT include the following.

- Suppose we have N observations on $\boldsymbol{X}$ that is multivariate normally distributed accord-

ing to $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, a test statistic is derived to test for the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ when Σ is unknown. The obvious MLE for Σ in this case is the sample covariance. The resultant test statistics is the T^2 statistics which follows the T^2 Hotelling distribution. This test can be used for testing the hypothesis about the mean vector $\boldsymbol{\mu}$ of the population and obtaining the confidence region for the unknown vector $\boldsymbol{\mu}$ see (Anderson, 2003; Seber, 2004; Mardia et al., 1980; Johnson and Wichern, 2007).

- The two sample problem with unequal covariance matrices has also been addressed. In this case, let $\left\{\boldsymbol{y_j^{(i)}}\right\}, j = 1, \ldots, N$ be samples from $\mathcal{N}(\boldsymbol{\mu}^{(i)}, \Sigma_i), i = 1, 2$ a test statistic for testing $H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ is developed. The distribution for the respective sample mean vectors is given by $\mathrm{E}(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(2)}) = \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$ while the covariance for the difference $\mathrm{Cov}(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(2)}) = \Sigma_1/N_1 + \Sigma_2/N_2$. It is shown that when $N_1 \neq N_2$ and assuming that $N_1 < N_2$ then a suitable test would be a T^2 test with $[N_1 - 1)$ degrees of freedom; see (Anderson, 2003).
- When Σ_1 and Σ_2 are assumed to be equal and unknown, then a pooled sample covariance is used as an estimate. The test statistic is found to be the usual T^2 which follows the T^2 distribution; see (Anderson, 2003; Seber, 2004).
- The topic of paired comparisons is also treated especially in Johnson and Wichern (2007) in which for the paired samples, the difference between them is calculated. The T^2 test is then applied to the differences.
- Most of the likelihood problems tackled only compare two mean vectors and the resultant statistic is the T^2 with a certain degree of freedom depending on the problem set-up.

In other related type of studies, Varuzza and Pereira (2010) developed an exact significance test for comparing digital expression profiles which took in to the asymptotic properties unlike the χ^2 test. Furthermore Lim et al. (2010) developed LRT to compare multiple multivariate normally correlated samples.

3. Proposed LRT statistic

Following the illustration in Section 1, for the healthy subjects, suppose that each gene has m paired measurements $[(h_{u1}, h_{s1}), (h_{u2}, h_{s2}), \ldots, (h_{um}, h_{sm})]$ where h symbolizes one of the groups, say healthy while the subscripts u and s stand for unstimulated and stimulated respectively. Therefore first measurement is for the expression level for the unstimulated specimen, while the second one is for a stimulated one for the same subject. In a similar manner let a represent the second group, say the infected subjects. Assume that each of the p genes has k paired measurements $[(a_{u1}, a_{s1}), (a_{u2}, a_{s2}), \ldots, (a_{uk}, a_{sk})]$ for the unstimulated and stimulated specimens in each pair respectively.

The m measurements from healthy subjects are assumed to be IID from a multivariate normal distribution $\begin{pmatrix} \boldsymbol{h}_u \\ \boldsymbol{h}_s \end{pmatrix} \sim \mathcal{N}_{2p}\left[\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_s \end{pmatrix}, \boldsymbol{\Sigma}\right]$ and the k measurements from the infected subjects are also assumed to be IID from a multivariate normal distribution $\begin{pmatrix} \boldsymbol{a}_u \\ \boldsymbol{a}_s \end{pmatrix} \sim \mathcal{N}_{2p}\left[\begin{pmatrix} \boldsymbol{\nu}_u \\ \boldsymbol{\nu}_s \end{pmatrix}, \boldsymbol{\Sigma}\right]$. Here, $\boldsymbol{\mu}_u$ and $\boldsymbol{\mu}_s$ represent the mean vectors for unstimulated and stimulated healthy subjects respectively. On the other hand, $\boldsymbol{\nu}_u$ and $\boldsymbol{\nu}_s$ denote mean vectors for unstimulated and stimulated infected subjects respectively while $\boldsymbol{\Sigma}$ is the covariance matrix which is assumed to be the same for the two groups of healthy and infected.

The hypotheses to be tested are:

$$H_0 : (\boldsymbol{\mu}_u - \boldsymbol{\mu}_s) = (\boldsymbol{\nu}_u - \boldsymbol{\nu}_s) \text{ versus } H_a : (\boldsymbol{\mu}_u - \boldsymbol{\mu}_s) \neq (\boldsymbol{\nu}_u - \boldsymbol{\nu}_s).$$

Case 1: Assuming the covariance matrix Σ is known For m healthy subjects denote a $2p \times 1$ vector of parameters $\boldsymbol{\mu} = \binom{\boldsymbol{\mu}_u}{\boldsymbol{\mu}_s}$ for the random vector $\boldsymbol{h} = \binom{\boldsymbol{h}_u}{\boldsymbol{h}_s}$ where the first p elements represent the elements of $\boldsymbol{h}_u$ while the remaining p represents the $\boldsymbol{h}_s$. Similarly for the k infected subjects we have the vector of parameters $\boldsymbol{\nu} = \binom{\boldsymbol{\nu}_u}{\boldsymbol{\nu}_s}$ and is associated with random variables $\boldsymbol{a} = \binom{\boldsymbol{a}_u}{\boldsymbol{a}_s}$ and $\boldsymbol{\nu}$ is of $2p \times 1$ dimension.

The joint probability density function is given as

$$f(\boldsymbol{h}, \boldsymbol{a}) = (2\pi)^{-p}|\boldsymbol{\Sigma}|^{-1}\exp\Big(-\frac{1}{2}\big[(\boldsymbol{h}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{h}-\boldsymbol{\mu}) + (\boldsymbol{a}-\boldsymbol{\nu})'\Sigma^{-1}(\boldsymbol{a}-\boldsymbol{\nu})\big]\Big).$$

A reduced -2log of the likelihood function in terms of sufficient statistics is given by

$$-2\text{Log L}(\boldsymbol{\mu}, \boldsymbol{\nu}) = B + m(\bar{\boldsymbol{h}}-\boldsymbol{\mu})'\Sigma^{-1}(\bar{\boldsymbol{h}}-\boldsymbol{\mu}) + k(\bar{\boldsymbol{a}}-\boldsymbol{\nu})'\Sigma^{-1}(\bar{\boldsymbol{a}}-\boldsymbol{\nu}), \tag{1}$$

where B is a constant that does not contain the parameters under consideration and vanishes during the optimization.

The MLEs under H_o are obtained by considering the parameter space given by $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{\nu} : -\infty < -\infty < \boldsymbol{\mu}, \boldsymbol{\nu} < \infty\}$ and then optimizing the constrained log-likelihood function using the Lagrangian $S(\boldsymbol{\Theta}, \boldsymbol{\lambda}) = -2\text{LogL}(\boldsymbol{\mu}, \boldsymbol{\nu}) + \boldsymbol{\lambda}'(\boldsymbol{\mu}_u - \boldsymbol{\mu}_s - \boldsymbol{\nu}_u + \boldsymbol{\nu}_s)$. The constraint $\boldsymbol{\lambda}'(\boldsymbol{\mu}_u - \boldsymbol{\mu}_s - \boldsymbol{\nu}_u + \boldsymbol{\nu}_s)$ is conveniently expressed in a matrix form as $A(\boldsymbol{\mu}-\boldsymbol{\nu}) = 0$ where $A = (\boldsymbol{I}, -\boldsymbol{I})$ and $\boldsymbol{I}$ is a $p \times p$ identity matrix. The constraint added to Equation (1) is of the form $2(\boldsymbol{\mu}-\boldsymbol{\nu})'A'\boldsymbol{\lambda} = 2[\boldsymbol{\lambda}'A(\boldsymbol{\mu}-\boldsymbol{\nu})]'$. The partial derivatives of the constrained function with respect to each unknown parameter are given as

$$\frac{\partial S(\boldsymbol{\Theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\mu}} = -2m\Sigma^{-1}(\bar{\boldsymbol{h}}-\boldsymbol{\mu}) + 2A'\boldsymbol{\lambda}, \tag{2}$$

$$\frac{\partial S(\boldsymbol{\Theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\nu}} = -2k\Sigma^{-1}(\bar{\boldsymbol{a}}-\boldsymbol{\nu}) - 2A'\boldsymbol{\lambda}, \tag{3}$$

$$\frac{\partial S(\boldsymbol{\Theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = 2A(\boldsymbol{\mu}-\boldsymbol{\nu}). \tag{4}$$

Now, equating (2), (3) and (4) to zero and simplifying, we get

$$\Sigma^{-1}(\bar{\boldsymbol{h}}-\boldsymbol{\mu}) - \frac{1}{m}A'\boldsymbol{\lambda} = 0, \tag{5}$$

$$\Sigma^{-1}(\bar{\boldsymbol{a}}-\boldsymbol{\nu}) + \frac{1}{k}A'\boldsymbol{\lambda} = 0, \tag{6}$$

$$A(\boldsymbol{\mu}-\boldsymbol{\nu}) = 0.$$

Subtracting Equation (5) from (6) and with some algebraic manipulations results in

$$\begin{aligned}
\Sigma^{-1}(\bar{\boldsymbol{a}} - \boldsymbol{\nu} - \bar{\boldsymbol{h}} + \boldsymbol{\mu}) + \Big[\frac{1}{m} + \frac{1}{k}\Big] A'\boldsymbol{\lambda} &= 0 \\
(\bar{\boldsymbol{a}} - \boldsymbol{\nu} - \bar{\boldsymbol{h}} + \boldsymbol{\mu}) &= -\Big[\frac{1}{m} + \frac{1}{k}\Big]\Sigma A'\boldsymbol{\lambda} \\
A(\boldsymbol{\mu} - \boldsymbol{\nu}) + A(\bar{\boldsymbol{a}} - \bar{\boldsymbol{h}}) &= -\Big[\frac{m+k}{mk}\Big] A\Sigma A'\boldsymbol{\lambda} \\
A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) &= \Big[\frac{m+k}{mk}\Big] A\Sigma A'\boldsymbol{\lambda} \\
\boldsymbol{\lambda} &= \Big[\frac{mk}{m+k}\Big](A\Sigma A')^{-1} A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) \\
\boldsymbol{\lambda} &= \Big[\frac{mk}{m+k}\Big](A\Sigma A')^{-1}\Delta \qquad (7)
\end{aligned}$$

where $\Delta = A\bar{\boldsymbol{h}} - A\bar{\boldsymbol{a}}$. From Equations (5) and (6), we get

$$\hat{\boldsymbol{\mu}}_0 = \bar{\boldsymbol{h}} - \frac{1}{m}\Sigma A'\boldsymbol{\lambda} \qquad (8)$$

$$\hat{\boldsymbol{\nu}}_0 = \bar{\boldsymbol{a}} + \frac{1}{k}\Sigma A'\boldsymbol{\lambda} \qquad (9)$$

The MLEs under the alternative hypothesis H_a are obtained by maximizing the unconstrained likelihood function are given by; $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{h}}$ and $\hat{\boldsymbol{\nu}} = \bar{\boldsymbol{a}}$.

Now, let $\boldsymbol{\theta}$ be the parameter vector for the likelihood function $L(\boldsymbol{\theta})$ with observations from the paired samples of healthy and infected subjects. Consider the parameter space given by $\boldsymbol{\Theta}$; we wish to test the null hypothesis $H_o : \boldsymbol{\theta} \in \boldsymbol{\Theta}$ versus the alternative $H_a : \boldsymbol{\theta} \notin \boldsymbol{\Theta}$.

Substituting the MLEs under H_0 (Equations (8) and (9)) into the log likelihood function given by Equation (1) we get

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_o}\{-2\text{Log L}(\boldsymbol{\theta})\} \\
&= B + m\left(\frac{1}{m}\Sigma A'\boldsymbol{\lambda}\right)'\Sigma^{-1}\left(\frac{1}{m}\Sigma A'\boldsymbol{\lambda}\right) + k\left(\frac{1}{k}\Sigma A'\boldsymbol{\lambda}\right)'\Sigma^{-1}\left(\frac{1}{k}\Sigma A'\boldsymbol{\lambda}\right) \\
&= B + \frac{1}{m}\left(\boldsymbol{\lambda}'A\Sigma\right)\Sigma^{-1}\left(\Sigma A'\boldsymbol{\lambda}\right) + \frac{1}{k}\left(\boldsymbol{\lambda}'A\Sigma\right)\Sigma^{-1}\left(\Sigma A'\boldsymbol{\lambda}\right) \\
&= B + \frac{1}{m}\left(\boldsymbol{\lambda}'A\Sigma A'\boldsymbol{\lambda}\right) + \frac{1}{k}\left(\boldsymbol{\lambda}'A\Sigma A'\boldsymbol{\lambda}\right) \\
&= B + \frac{[k+m]}{mk}\left(\boldsymbol{\lambda}'A\Sigma A'\boldsymbol{\lambda}\right). \qquad (10)
\end{aligned}$$

We now substitute for the expression of $\boldsymbol{\lambda}$ from (7) into Equation (10) to get

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_o}\{-2\text{Log L}(\boldsymbol{\theta})\} = \\
&B + \frac{[k+m]}{mk}\left(\frac{mk}{[m+k]}(A\Sigma A')^{-1}\Delta\right)'(A\Sigma A')^{-1}\left(\frac{mk}{[m+k]}(A\Sigma A')^{-1}\Delta\right) \\
&= B + \frac{mk}{[m+k]}\Delta'(A\Sigma A')^{-1}\Delta.
\end{aligned}$$

Under the unconstrained hypothesis $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{-2\text{Log L}(\boldsymbol{\theta})\} = B$. The log LRT is therefore given as

$$\begin{aligned} 2\text{Log}\Lambda &= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_o}\{-2\text{Log L}(\boldsymbol{\theta})\} - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{-2\text{Log L}(\boldsymbol{\theta})\} \\ &= \frac{mk}{[m+k]}\Delta'(A\Sigma A')^{-1}\Delta \end{aligned} \tag{11}$$

The distribution of $\Delta = A\bar{\boldsymbol{h}} - A\bar{\boldsymbol{a}}$ is $\Delta \sim N\left((A(\boldsymbol{\mu}-\boldsymbol{\nu}), \frac{(k+m)}{mk}(A\Sigma A')^{-1}\right)$. If H_0 is true then $A(\boldsymbol{\mu}-\boldsymbol{\nu}) = 0$ so that $\Delta \sim N\left(0, \frac{(k+m)}{mk}(A\Sigma A')^{-1}\right)$. It is well known that given that $X \sim N_p(0,V)$ then $V^{-\frac{1}{2}} \sim N(0,I)$ implying that $(V^{-\frac{1}{2}}X)^T(V^{-\frac{1}{2}}X) \sim \chi^2_{(p)}$ and so $X^TV^{-1}X \sim \chi^2_{(p)}$, thus

$$-2\text{Log}\Lambda = \frac{mk}{(m+k)}\Delta'(A\Sigma A')^{-1}\Delta \sim \chi^2_{(p)}. \qquad \blacksquare$$

Case 2: Assuming the covariance matrix Σ is unknown We estimate the covariance matrix by first rewriting the -2log likelihood as

$$\begin{aligned} l = mp\log(2\pi) + m\log|\Sigma| + \text{tr}\Sigma^{-1}\boldsymbol{S}_h + \text{tr}\Sigma^{-1}(\bar{\boldsymbol{h}}-\boldsymbol{\mu})(\bar{\boldsymbol{h}}-\boldsymbol{\mu})' \\ +kp\log(2\pi) + k\log|\Sigma| + \text{tr}\Sigma^{-1}\boldsymbol{S}_a + \text{tr}\Sigma^{-1}(\bar{\boldsymbol{a}}-\boldsymbol{\nu})(\bar{\boldsymbol{a}}-\boldsymbol{\nu})', \end{aligned} \tag{12}$$

where $\boldsymbol{S}_h = \sum_{i=1}^m(\boldsymbol{h}_i-\bar{\boldsymbol{h}})(\boldsymbol{h}_i-\bar{\boldsymbol{h}})'$ and $\boldsymbol{S}_a = \sum_{j=1}^k(\boldsymbol{a}_j-\bar{\boldsymbol{a}})(\boldsymbol{a}_j-\bar{\boldsymbol{a}})'$. We obtain the partial derivative of l (12) with respect to Σ^{-1}, then equate the result to zero. The estimator for the variance-covariance matrix is then obtained as

$$\hat{\Sigma} = \frac{1}{[m+k]}\left[\boldsymbol{S}_h + \boldsymbol{S}_a + m(\bar{\boldsymbol{h}}-\hat{\boldsymbol{\mu}})(\bar{\boldsymbol{h}}-\hat{\boldsymbol{\mu}})' + k(\bar{\boldsymbol{a}}-\hat{\boldsymbol{\nu}})(\bar{\boldsymbol{a}}-\hat{\boldsymbol{\nu}})'\right].$$

By substituting the plug-in estimators for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\nu}}$ which are $\bar{\boldsymbol{h}}$ and $\bar{\boldsymbol{a}}$ respectively, we get the plug-in estimator for the covariance matrix as

$$\hat{\Sigma} = \frac{1}{[m+k]}\left[\boldsymbol{S}_h + \boldsymbol{S}_a\right].$$

The estimator $\hat{\Sigma}$ is then plugged-in into the LRT statistic given in Equation (11) which has $\chi^2_{(p)}$ distribution to get

$$-2\text{Log}\Lambda = \frac{mk}{[m+k]}\Delta'(A\hat{\Sigma}A')^{-1}\Delta. \tag{13}$$

Proposition Denote Equation (11) by $\Lambda_1 = \frac{mk}{[m+k]}\Delta'(A\Sigma A')^{-1}\Delta$ and (13) by $\Lambda_2 = \frac{mk}{[m+k]}\Delta'(A\hat{\Sigma}A')^{-1}\Delta$ and noting that $\hat{\Sigma}$ is a consistent estimator of Σ. Since $\Lambda_1 \overset{d}{\sim} \chi^2_{(p)}$ then $\Lambda_2 \overset{a}{\sim} \chi^2_{(p)}$, where $\overset{d}{\sim}$ means exactly distributed while $\overset{a}{\sim}$ stands for asymptotically distributed.

PROOF Since $\hat{\Sigma} \xrightarrow{p} \Sigma$ as $n \to \infty$ where $n = m + k$ and the fact that $(A\Sigma A')$ is positive definite, we had shown in Case 1 that $A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) \overset{d}{\sim} \mathrm{N}\left(0, \frac{m+k}{mk} A\Sigma A'\right)$ under H_0 then it follows that in a similar manner $A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) \overset{a}{\sim} \mathrm{N}\left(0, \frac{m+k}{mk} A\hat{\Sigma} A'\right)$ under H_0. Consequently the LRT statistic $\frac{mk}{[m+k]} \Delta' (A\hat{\Sigma} A')^{-1} \Delta \overset{a}{\sim} \chi^2_{(p)}$. ■

REMARK We note that the world applications, p is usually less than n, that is, $p < n$. In such a case, the derived statistic in Equation (13) becomes untenable because the matrix $(A\hat{\Sigma} A')$ is singular. In order to overcome this problem, the usage of the general inverse as in Ben-Israel and Greville (2003) of the covariance matrix is instead used.

4. Simulation study

In this section, a synthetic data are generated and then analyzed using the proposed LRT method and the well known Hotelling T^2 statistic. All the simulations and data analysis were done using the `R` software (R Core Team, 2020). The data are simulated with the following different set-ups.

- The mean vector for the "healthy unstimulated" is obtained by first simulating p uniform random variables in the range of $(0, 0.5)$ to the vector $\boldsymbol{\mu}_u$.
- Similarly we generate p uniform random variables in the interval $(0.6, 0.75)$ to create $\boldsymbol{\mu}_s$ which is the "healthy unstimulated".
- For the "infected unstimulated", the values for simulation of $\boldsymbol{\nu}_u$ used to generate uniform random variables of dimension p is $(0, 0.55)$.
- The $\boldsymbol{\nu}_s$ are obtained by generating a p uniform random variables of the interval $(0.001, 0.2)$ to obtain the mean vector for the "infected stimulated".

For each of the category, we assume that all the two paired measurements we generate at randomly a $2p \times 2p$ positive definite covariance matrix V. The number of subjects for the healthy group is arbitrarily set at 20 while the infected group is set at 19.

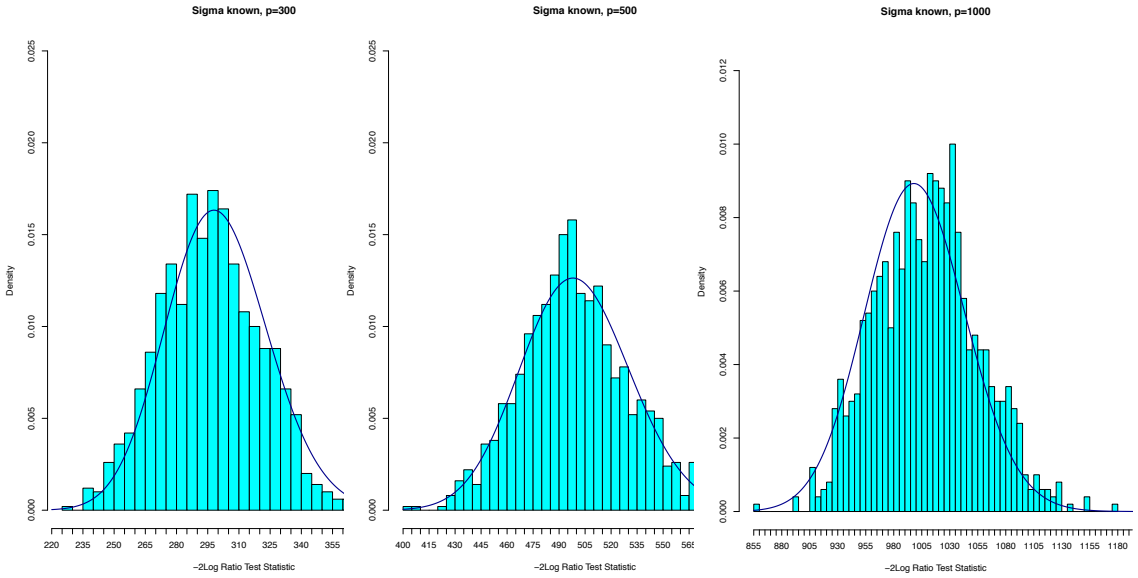
The data are simulated for four different values of p namely $p = \{300, 500, 1000\}$ while the sample sizes were fixed at $m = 20$ and $k = 19$. A LRT statistic and the corresponding p-value are calculated when Σ is assumed to be unknown and when it is known. A resampling distribution is then obtained from which an approximate p-value is then computed. The results are shown in Table 1 in addition to the plots in Figure 1.

The proposed statistic is applied to the simulated data. The results presented in Table 1 reveal that both the calculated p-value and the one obtained from resampling lead to the same conclusions regarding the hypothesis testing. In this case, for all the cases, the difference in the means was statistically significant at 5% level.

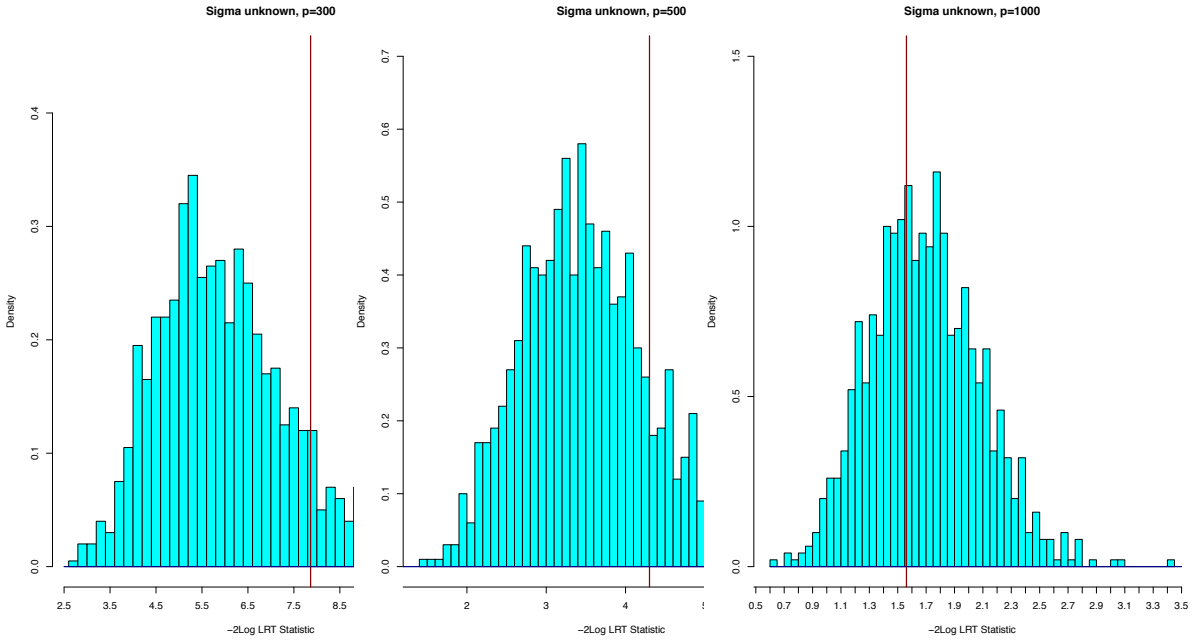
Table 1.: Calculated LRT statistic and the p-values for the simulation experiment 1.

	$p = 300$		$p = 500$		p =1000	
	Σ known	Σ Unknown	Σ known	Σ unknown	Σ known	Σ unknown
Log LRT	373.61	7.86	617.43	4.3	1202.05	1.56
calculated p-value*	0.00025	1.00	0.002	1.00	0.00	1.00
p-value from resampling	0.001	0.396	0.001	0.166	0.00	0.601

* p-values calculated from the exact $\chi^2_{(p)}$ distribution.



(a) p =300, Σ-known (b) p =500, Σ-known (c) p =1000, Σ-known



(d) p =300, Σ-unknown (e) p =500, Σ-unknown (f) p =1000, Σ-unknown

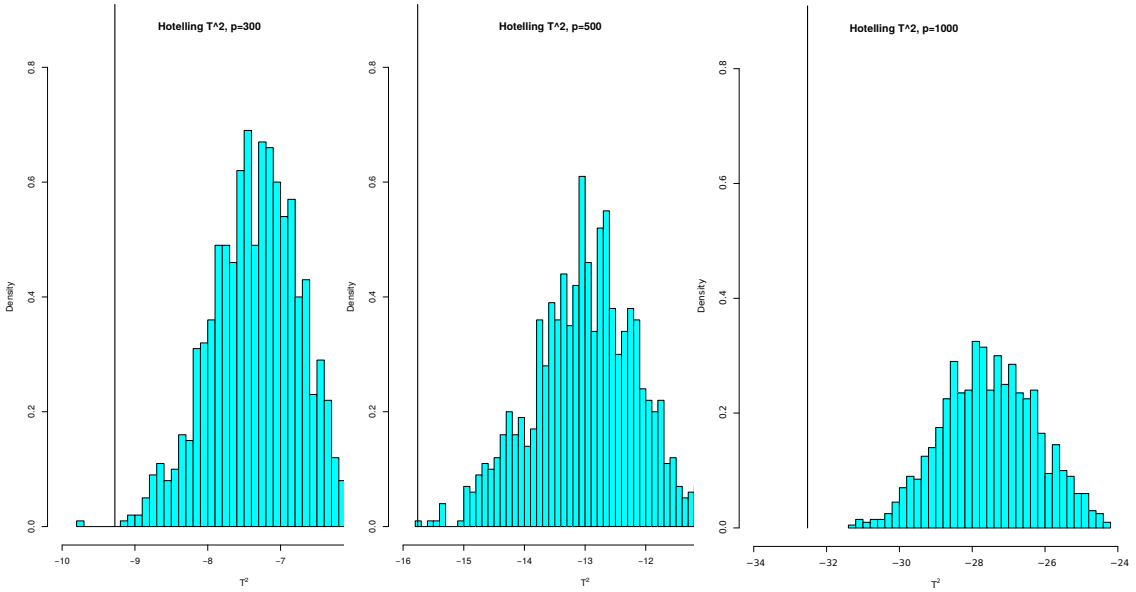
Figure 1.: Histograms for the proposed LRT from a the permutation of the statistic for p =300, 500 and 1000, Σ known and unknown.

In Figure 1, the curves are the chi-squared densities for the corresponding degrees of freedom p. The plots show that the distributions for the -2 log likelihood test statistic follow a chi-square distribution and are also positive skewed. However, as the number of p increases, the distributions look like normal distribution and the skewness is less when the degree of freedom is higher. The normal looking distribution are still a chi-squared, for they approach $\mathrm{N}(p, 2p)$ distribution as the degree of freedom gets large. The red vertical lines (when shown) indicates the position of the computed statistic for the un-resampled data. The plots without the vertical lines are the ones whose computed statistic is far too small beyond the scale used in plotting.

The simulated data are analyzed using the Hotelling T^2 statistic (Hotelling, 1931) in order to compare the performance of our proposed method with it. During the computation, when number of variables p is much greater than the number of samples n, then the covariance matrix is estimated using the shrinkage approach of Schäfer and Strimmer (2005). The results are presented in Table 2 . The results indicate that there is a significant difference in the means at 5% significance level. The permutations for Hotelling T^2 statistic is done and the different values plotted on an histogram shown in Figure 2 which reveals that the statistic is chi-square distributed for all the different values of p. The results are consistent with the one obtained by the proposed algorithm.

Table 2.: Hotelling T^2 values for the simulated data.

	p=300	p=500	p=1000
Hotelling T^2 value	-9.28	-15.76	-32.52
p-value from resampling	0.001	0.00	0.00



(a) p =300, Σ-known (b) p =500, Σ-unknown (c) p =1000, Σ-unknown

Figure 2.: Histograms of the distribution of the permuted test statistics for Hotelling T^2 when p =300, 500 and 1000, Σ unknown.

5. Conclusions

In this research, we have considered two main groups (say, healthy and infected specimens) with paired measurements that are correlated. We aim to provide a proper statistical framework for testing the difference in the difference in the means for the healthy and infected subjects. We have shown that this is not a trivial problem and so derived a likelihood ratio test for these differences. The derived test do follow a chi-square distribution with p degrees of freedom when the variance-covariance matrix is known. We have assumed that the observed measurements follow a multivariate normal distribution with a known variance-covariance matrix which can be deduced from the prior network that has been chosen. Finally, a likelihood ratio test statistic has been derived when the variance-covariance matrix is unknown. A simulation study has been done and demonstrated that the developed tests can be useful when applied to other cases which have similar problem set-ups. The study demonstrated that the proposed test statistic give the same conclusions for the hypotheses tested as those of the Hotelling T^2 test. However, the proposed test is intuitively more appealing since it takes care of the correlations between the pairs in the experiments. This research contributes a theoretical analysis of paired correlated samples motivated by a practical problem for which no formal statistical method is in use.

Acknowledgements

This research was done during my academic visit at the Speed Lab as a PhD candidate (partly supported by the Mexico's Consejo Nacional de Ciencias y Tecnología (CONACyT) scholarship number 384101), Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research (WEHI), Melbourne, Australia. I am very grateful to Prof. Terence Speed for his guidance through out this period. Much appreciation to Dr. Jo Keeble and Prof. Ian Wicks for their support and discussions on the Acute Rheumatic Fever (ARF) project that lead to the conception of ideas relating to this work. Much appreciation to Dr. Graciela González Farías of CIMAT, Gto, México for the useful comments that helped in improving this article. This research was partially supported by the Mexico's Consejo Nacional de Ciencias y Tecnología (CONACyT) project number 252996 through Dr. Graciela González Farías.

References

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Ben-Israel, A. and Greville, T., 2003. Generalized Inverses:Theory and Applications. Springer, New York.

Hotelling, H., 1931. The generalization of student's ratio. Annals of Mathematical Statistics, 2:360–378.

Johnson, R.A. and Wichern, D.W., 2007. Applied Multivariate Analysis. Prentice Hall, New Jersey.

Lim, J., Li, E., and Lee., S.J., 2010. Likelihood ratio tests of correlated multivariate samples. Journal of Multivariate Analysis, 101:541–554.

Mardia, K.V., Kent, J.T., and Bibby, J.M., 1980. Multivariate Analysis. Academic Press, New York.

Schäfer, J. and Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology, 4:32.

Seber, A. 2004. Multivariate Observations. Wiley, New York.

R Core Team 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Varuzza, L. and Pereira, C., 2010. Significance test for comparing digital gene expression profiles: Partial likelihood application. Chilean Journal of Statistics, 1:91–102.

Information for authors

The editorial board of the Chilean Journal of Statistics (ChJS) is seeking papers, which will be refereed. We encourage the authors to submit a PDF file of the manuscript in a free format to Editors of the ChJS (E-mail: `chilean.journal.of.statistics@gmail.com`). Submitted manuscripts must be written in English and contain the name and affiliation of each author followed by a leading abstract and keywords. The authors must include a"cover letter" presenting their manuscript and mentioning: "We confirm that this manuscript has been read and approved by all named authors. In addition, we declare that the manuscript is original and it is not being published or submitted for publication elsewhere".

Preparation of accepted manuscripts

Manuscripts accepted in the ChJS must be prepared in Latex using the ChJS format. The Latex template and ChJS class files for preparation of accepted manuscripts are available at `http://chjs.mat.utfsm.cl/files/ChJS.zip`. Such as its submitted version, manuscripts accepted in the ChJS must be written in English and contain the name and affiliation of each author, followed by a leading abstract and keywords, but now mathematics subject classification (primary and secondary) are required. AMS classification is available at `http://www.ams.org/mathscinet/msc/`. Sections must be numbered 1, 2, etc., where Section 1 is the introduction part. References must be collected at the end of the manuscript in alphabetical order as in the following examples:

Arellano-Valle, R., 1994. Elliptical Distributions: Properties, Inference and Applications in Regression Models. Unpublished Ph.D. Thesis. Department of Statistics, University of São Paulo, Brazil.

Cook, R.D., 1997. Local influence. In Kotz, S., Read, C.B., and Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 1., Wiley, New York, pp. 380-385.

Rukhin, A.L., 2009. Identities for negative moments of quadratic forms in normal variables. Statistics and Probability Letters, 79, 1004-1007.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tsay, R.S., Peña, D., and Pankratz, A.E., 2000. Outliers in multivariate time series. Biometrika, 87, 789-804.

References in the text must be given by the author's name and year of publication, e.g., Gelfand and Smith (1990). In the case of more than two authors, the citation must be written as Tsay et al. (2000).

Copyright