Journal of Mathematics and Statistics 7 (3): 227-229, 2011
ISSN 1549-3644

Inversion of Covariance Matrix for High Dimension Data

Samruam Chongcharoen
National Institute of Development Administration,
School of Applied Statistics, Bangkok, 10240,Thailand

Abstract: Problem statement: In the testing statistic problem for the mean vector of independent and identically distributed multivariate normal random vectors with unknown covariance matrix when the data has sample size less than the dimension n≤p, for example, the data came from DNA microarrays where a large number of gene expression levels are measured on relatively few subjects, the p×p sample covariance matrix S does not have an inverse.. Hence any statistic value involving inversion of S does not exist. **Approach:** In this study, we showed a version of some modification on S, S+cI and find a real smallest value c≠0 which makes $(S + cI)^{-1}$ exist. **Results:** The result from study provided when the dimension p tends to infinity and smallest change in S, the $(S + cI)^{-1}$ do exist when c = 1. **Conclusion:** In statistical analysis involving with high dimensional data that an inversion of sample covariance matrix do not exist, one way to modify a sample covariance matrix S to have an inverse is to consider a sample covariance matrix, S, as the form S + cI and we recommend to choose c = 1.

Key words: DNA micro arrays, eigenvalue, positive semidefinite, positive definite, gene expression, covariance matrix, statistic value, real vector, real number, determinant, symmetric matrix, definite matrix

INTRODUCTION

Now suppose $X_1, X_2,\ldots,X_n$ is a random sample from a p-dimensional multivariate normal distribution with unknown mean $\mu = (\mu_1, \mu_2,\ldots, \mu_p)'$ and unknown positive definite covariance matrix V with n≤p. The sample mean and p×p sample covariance matrix S are

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} \text{ and } \quad S = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(X_i - \bar{X})'}{n-1} \qquad (1)$$

Since covariance matrix is a real symmetric matrix and Harville (1997) showed that every symmetric matrix has an eigenvalues.

The Hotelling's T^2 statistic do not exist, so Dempster (1958; 1960), Bai and Saranadasa (1996) and Srivastava and Du (2008) developed test statistics using some other forms of *S* for their tests instead of inversion of S because it does not exist. By equipping with the knowledge of (Polymenis, 2011; Girgis *et al*., 2010; George and Kibria, 2010; Yahya *et al*., 2011; Nassiry *et al*., 2009) will help to transform ideas.

Searle (1982) also showed that the eigenvalues of every real symmetric matrix is real. For n≤p, Johnson and Wichern (2002) have shown that the determinant of sample covariance is zero for all samples, that is, S is singular. Now consider p×p an covariance matrix A with an eigenvalue λ, which for any real vector $v \neq 0$, then by the definition of eigenvalue, we have $Av = \lambda v$, then:

$$v'Av = \lambda v'v$$

and then:

$$\frac{v'Av}{v'v} = \lambda$$

Because v'v > 0, A is positive semidefinite (v'Av ≥ 0) if and only if $\lambda \geq 0$ and A is positive (v'Av > 0) definite if and only if λ>0. Thus the covariance matrix *S* is at least positive semidefinite. Searle (1982) also showed that for any p×p matrix A, the determinant of A is equal to the product of its eienvalues, that is, $|A| = \prod_{i=1}^{p} \lambda_i$. Hence, covariance matrix S must have at least one eigenvalue to be zero. Since every positive definite matrix is nonsingular and its determinant is positive, so the easiest way to makes covariance matrix S from high dimensional data to be nonsingular is to modify it to be positive definite matrix. We consider the form S + cI, $c \neq 0$ by looking for a smallest $c \neq 0$ which makes $(S + cI)^{-1}$ exist. Now suppose that S has r nonzero eigenvalues, that is, it has exactly r positive eigenvalues and p-r zero eigenvalues. We are interested in modifying S to be nonsingular with the smallest

change in S by considering S + cI, $c \neq 0$ for any real number.

MATERIAL AND METHODS

Suppose $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r > 0$ and $\lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_p = 0$ are all eigenvalues of S. From the definition of evigenvalue, for any eigenvalue λ; i = 1, 2,…, p of S and for any real vector $v \neq 0$, we have $Sv = \lambda v$, then $(S + cI)v = Sv + cv = \lambda v + cv = (\lambda + c)v$. So, $\lambda + c$ is an eigenvalue of S + cI. Thus all p eigenvalues of S + cI are $\lambda_1 + c \geq \lambda_2 + c \geq \lambda_3 + c \geq \ldots \geq \lambda_r + c > c = \ldots = c$. We can see that *c* cannot be negative because if it does, S + cI cannot be positive definite matrix. Now if $0 < c < 1$, the determinant of S + cI is:

$$|S + cI| = \prod_{i=1}^{r} (\lambda_i + c)(c)^{p-r}$$

which approaches to zero as p tends to infinity, that makes $(S + cI)^{-1}$ does not exist. Therefore the only one possible case is $c \geq 1$, but we are looking for a smallest value *c* that makes $(S + cI)^{-1}$ exist. So we pick c = 1. The proof is completed.

RESULTS

The result from this study provided a way to modify a sample covariance matrix, S, came from the data with the number of the dimension p larger than the number of observation n available, $n \leq p$, which its inversion of sample covariance matrix do not exist, to be S + cI with smallest change in S and then $(S + cI)^{-1}$ do exist with c = 1.

DISCUSSION

At present, there are a number of data with the number of the dimension p larger than the number of observation n available, $n \leq p$, in many diverse applied fields, e.g., medical, pharmaceutical, agricultural, psychological, educational, social, behavioral, political, criminal, industrial, meteorological, zoological and biological sciences but there are barely any statistical technique for analyzing this kind of data. The resulted technique we found may help the researchers to develop new statistical techniques for analyzing high dimensional data.

CONLUSION

In statistical analysis, when one involves with high dimensional data, the number of sample size less than the number of dimension(variables), any statistic values involving with inversion of sample covariance matrix will not exist because inversion of sample covariance matrix do not exist. One way to modify a sample covariance matrix, S, to have an inverse is to consider a sample covariance matrix, S, as the form S + cI. For this form of sample covariance matrix, we showed when $c \geq 1$, that $(S + cI)^{-1}$ do exist and for smallest change in S, we recommend to choose c = 1.

ACKNOWLEDGMENT

Professor Dr. A.K. Gupta, department of Mathematics and statistics, Bowling Green State University, Bowling Green ,USA, for his great suggestions.

REFERENCES

Bai, Z. and H. Saranadasa, 1996. Effect of high dimension: An example of a two sample problem. Statist. Sinica, 6: 311-329. http://www3.stat.sinica.edu.tw/statistica/j6n2/j6n21/j6n21.htm

Dempster, A.P., 1958. A high dimensional two sample significance test. Ann. Math. Stat., 29: 995-1010. http://projecteuclid.org/euclid.aoms/1177706437

Dempster, A.P., 1960. A significance test for the separation of two highly multivariate small samples. Biometrics, 16: 41-50. http://www.jstor.org/stable/2527954

George, F. and B.M.G. Kibria, 2010. Some test statistics for testing the binomial parameter: empirical power comparison. Am. J. Biostat., 1: 82-93. DOI: 10.3844/amjbsp.2010.82.93

Girgis,H., R. Hamed and M. Osman, 2010. Testing the equality of growth curves of independent populations with application on Egypt case. Am. J. Biostat., 1: 46-61. DOI: 10.3844/amjbsp.2010.46.61

Harville, D.A., 1997. Matrix Algebra From A Statistician's Perspective. Springer, ISBN: 0-387-94978-X, pp: 533. http://springer.com/statistics/statistical

Johnson, R.A. and D.W. Wichern, 2002. Applied Multivariate Statistical Analysis. 5th Edn., Prentice Hall, ISBN: 0-13-092553-5, pp: 135. http://www.prenhall.com/

Nassiry, M.R., A. Javanmard and R. Tohidi, 2009. Application of statistical procedures for analysis of genetic diversity in domestic animal populations. Am. J. Anim. Vet. Sci., 4: 136-141. DOI: 10.3844/ajavsp.2009.136.141

Polymenis, D.A., 2011. An application of univariate statistics to hotelling's T^2. J. Math. Stat., 7: 86-94. DOI: 10.3844/jmssp.2011.86.94

Srivastava, M.D. and M. Du, 2008. A test for mean vector with fewer observations than the dimension. J. Multivariate Anal. 99: 386-402. http://www.elsevier.com/locate/jmva

Searle, S.R., 1982. Matrix Algebra Useful for Statistics. John Wiley and Sons, Inc., ISBN: 0-471-86681-4, pp: 278.

Yahya, A.A., A. Osman, A.R. Ramli and A. Balola, 2011. Feature selection for high dimensional data: An evolutionary filter approach. J. Comput. Sci., 7: 800-820. DOI: 10.3844/jcssp.2011.800.820