Multiclass Feature Selection and Classification with Support Vector Machine in Genomic Study

A. W. Banjoko[1]; W. B. Yahya[2]; M. K. Garba; O. R. Olaniran; L. B. Amusa; N. F. Gatta; K. A. Dauda[†]; K. O. Olorede[†]

Department of Statistics,
University of Ilorin,
Ilorin Nigeria.
email: [1]banjokoalabi@gmail.com; [2]dr.yah2009@gmail.com

[†]Department of Statistics and Mathematical Sciences,
Kwara State University, Malete,
Malete, Nigeria.

***Abstract*—This study proposes an efficient Support Vector Machine (SVM) algorithm for feature selection and classification of multiclass response group in high dimensional (microarray) data. The Feature selection stage of the algorithm employed the F-statistic of the ANOVA–like testing scheme at some chosen *family-wise-error-rate* (FWER) to control for the detection of some false positive features. In a 10-fold cross validation, the hyper-parameters of the SVM were tuned to determine the appropriate kernel using *one-versus-all* approach. The entire simulated dataset was randomly partitioned into 95% training and 5% test sets with the SVM classifier built on the training sets while its prediction accuracy on the response class was assessed on the test sets over 1000 Monte-Carlo cross-validation (MCCV) runs. The classification results of the proposed classifier were assessed using the Misclassification Error Rates (MERs) and other performance indices. Results from the Monte-Carlo study showed that the proposed SVM classifier was quite efficient by yielding high prediction accuracy of the response groups with fewer differentially expressed features than when all the features were employed for classification. The performance of this new method on some published cancer data sets shall be examined vis-à-vis other state-of-the-earth machine learning methods in future works.**

Keywords-*Support Vector Machines; Monte-Carlo Cross-Validation; F-Statistic; Family wise error rate; Misclassification Error Rate.*

I. INTRODUCTION

Clinical diagnosis, identification and classification of cancer types often require thorough examination of the tumour cells using microscope and some other clinical parameters. However, this clinical procedure has been reported to often take considerable longer time before the types of tumour presence could be detected [17]. A number of studies have reported that cancer types might be discovered earlier with the use of microarray analysis than with the clinical methods [2,19]. Non-clinical classification of the various cancer types using gene expression profiling has been the most recently efficient alternative technique to clinical methods due to its numerous advantages [2, 14, 15].

Non-clinical diagnosis of cancer problems with binary endpoints, being the most common, has been given prominent attentions in the literature (see [2,14,15] among others) while few discussions only exist for multiclass cancer problems [16].

Gene expression profiling has been utilized for tumor finding (grouping) in numerous omics studies and this often resulted to the selection of gene subsets that have meaningful biological relationships with the tumour classes of the mRNA samples [1-3,13,14]. Thus, the selection of useful genes requires the selection of those gene subsets that are factually (statistically) significant and are naturally (biologically) relevant to the response class. The main objective in feature selection exercise therefore is to arrive at a classification model that uses fewer most relevant gene subsets to maximize its predictive accuracy of the tumour class of the mRNA samples [8,19].

The Support Vector Machines (SVM) is one of the state–of–the–art tools in the field of statistical learning and pattern recognition. Its theoretical development and applications have been presented in many works (see [3,4, 12,14] among others). A thorough review of the SVM methodology for cancer tumour classification in a binary response microarray data problem was presented in [2,9,14, etc.].

In this work, the SVM technique was employed to classify tumor types in multiclass response microarray cancer cases. The selection of core most relevant genes for classification was performed using the F (Welch) – statistic of the ANOVA–like testing scheme at some chosen *family-wise-error-rate* (FWER) which were set purposely to control for the selection of some false positive genes. The *one-versus-all* approach of the multiclass response category was adopted based on the submission in [10]. The goodness of the proposed method was fully examined on simulated microarray cancer dataset.

II. Materials and Methods

A. Simulation scheme

A multiclass response microarray dataset was simulated by adapting the approach reported in [14] and [16] was extended to multiclass response data situation. In this study, a three-response class microarray cancer data case was conjectured.

A total of 150 observations were simulated in all with the first 50 samples ($n_1 = 50$) came from the first group (group 1), the second 50 samples ($n_2 = 50$) came from the second group (group 2) and the third 50 samples ($n_3 = 50$) came from the third group (group 3) such that $n_1 + n_2 + n_3 = n$.

Table 1: An overview of the simulated multiclass microarray data structure.

S/N	y	$g_1, g_2, \dots, g_{10}$	$X_{11}, X_{12}, \dots, X_{1000}$
1 2 . . 50	group 1 . . . group 1	$\mu_1 = 2$	$\mu_4 = 3$
51 . . . 100	group 2 . . . group 2	$\mu_2 = 4$	
101 . . . 150	group 3 . . . group 3	$\mu_3 = 3.5$	

On each observation, 1000 covariates, representing the observed gene expression profiles were simulated. Of 1000 genes simulated on the three sample groups, 10 of them: $g_1, g_2, \dots, g_3$were simulated from the mixture of three multivariate normal densities with respective mean vectors μ_1, μ_2 and μ_3, and variance-covariance matrices $\sum_1$, $\sum_2$ and $\sum_3$, $\mu_1, \mu_2, \mu_3 > \mu$. That is, $((g_1, g_2, \dots, g_{10})|Y) \sim [\pi_1 * N(\mu_1, \sum_1) + \pi_2 * N(\mu_2, \sum_2) + \pi_3 * N(\mu_3, \sum_3)]$ with the values of the mixing parameter π_i taken to be 0.3 for all i. These ten genes are those whose expression levels are strongly related to the tumour groups hence, they are called Biomarkers. The remaining 990 genes $(X_{11}, X_{12}, \dots, X_{1000})$ constitute the genes with relatively low expression levels and were simulated from multivariate normal densities with mean vectors μ and variance-covariance matrix Σ. The whole data set simulated is of dimension $n \times q \; i.e. (150 \times 1000)$ obviously with $n < q$, as usually the case with microarray data. An overview of the simulated data employed in this study is presented by Table 1. In all cases, the covariance matrix Σ defined as $\Sigma = \{\sigma_{ij}\}$, has a block structure such that:

$$\sigma_{ij} = \begin{cases} 0.2, & if|j-i| \le 5 \\ 0, & otherwise \end{cases}$$

B. Methodology

Firstly, the feature selection algorithm was applied on the datasets using the F – statistic in [6] defined as;

$$F = \frac{\frac{1}{k-1}\sum_{j=1}^{k} w_j(\bar{x}_j - \bar{x}')^2}{1 + \frac{2(k-2)}{k^2-1}\sum_{j=1}^{k}\left(\frac{1}{n_j - 1}\right)\left(1 - \frac{w_j}{w}\right)^2} \sim F(k - 1, df) \quad (1)$$

where:

$$w_j = \frac{n_j}{S_j^2} \;, \; w = \sum_{j=1}^{k} w_j \,, \qquad \bar{x}' = \frac{\sum_{j=1}^{k} w_j \bar{x}_j}{w}$$

$df = \frac{k^2-1}{3\sum_{j=1}^{k}\left(\frac{1}{n_j-1}\right)\left(1-\frac{w_j}{w}\right)^2}$ is the degree of freedom and k = the number of the response class in the data.

In order to control the number of false positive genes in the feature selection process of the proposed algorithm, the method of Sidak [20] was adopted at six different FWERs (1%, 5%, 10%, 15%, 20% and 100%) as follows;

$$\alpha_s = 1 - (1 - \alpha_F)^{1/G}$$

where; α_s is the actual level of significance for the test as proposed by Sidak[20] for controlling the number of false positive in a multiple hypothesis testing for comparing a number of genes, α_F if the FWER and G is the number of features in the microarray data. A particular feature say X_j will be termed significant biomarkers if its $P - value(P_j)$ from the F – statistic discuss above is less than α_s $(P_j < \alpha_s)$ [1].

The entire dataset was randomly partitioned into 95% training and 5% test sets as proposed by [1, 14]. The

traditional SVM proposed by [13] was employed for feature selection and tumour classification using the algorithm in [2] based on *one-versus-all* approach [16] for multiple class responses with $k > 2$. For the response group $y = \{1, 2, \ldots, k\}$, the idea is to fit k SVMs classifiers to distinguish one of the k classes in y from the remaining $k-1$ classes at each fit. The k^{th} reference class in y is coded $+1$ and the remaining other classes are coded -1. By this, all other complementary classes are put into one group and subgroups of binary responses are formed with the selected set of gene predictors on which the SVM algorithm is employed over 1000 Monte-Carlo Cross validation (MCCV) runs to stabilize the classification model. The final decision is taken based on majority vote after classification results are obtained [16].

Table 2: The number of differentially expressed genes (Biomarkers) selected at different α_F level for simulated multiclass data.

α_F	No. of genes selected	Genes Selected
1%	9	$g_6, g_{10}, g_2, g_3, g_1, g_5, g_9, g_8, g_7$
5%	10	$g_6, g_{10}, g_2, g_3, g_1, g_5, g_9, g_8, g_7, g_4$
10%	10	$g_6, g_{10}, g_2, g_3, g_1, g_5, g_9, g_8, g_7, g_4$
15%	11	$g_6, g_{10}, g_2, g_3, g_1, g_5, g_9, g_8, g_7, g_4, X_{284}$
20%	11	$g_6, g_{10}, g_2, g_3, g_1, g_5, g_9, g_8, g_7, g_4, X_{284}$
100%	1000	**All the genes**

Table 3: The values of the tuning parameters determined for the Simulated Multiclass microarray data.

No of genes	Choice of Kernel	Kernel Parameter γ	SVM Parameter C	Minimum CV Error
9	RBF	0.01	1	0.3000
10	RBF	0.01	10	0.3000
10	RBF	0.01	10	0.3000
11	RBF	0.001	100	0.2533
11	RBF	0.001	100	0.2533
1000	LINEAR	N/A	0.01	0.5

The entire feature selection processes was performed in a sequential manner in which a set of gene subsets are formed at each gene selection stage and their contributions in term of the average Misclassification Error Rate (MER) or average Correct Classification Rate (CCR = 1-MER) of the classifier were determined. Generally, the gene subset that yielded the least average MER (highest CCR) among the rest is adjudged the best primary gene subset for the data.

In order to optimize the performance of the SVM classifier, all the primary gene subset selected through the sequential procedure above are ordered by their estimated p-values beginning from the gene with the least p-value to the one with the highest value. The p-values mark the strength of association of each gene with the response class. Thus, gene with the least p-value has the strongest association with the response class and it is considered the best followed by the second best, third best and so on as the p-values increase in the rank.

The optimization process begins by selecting the first 5, first 10, first 15, first 20 and first 25 genes and so on from the ranked (ordered) genes into the SVM for classification of the response class. This sequential-like gene selection process terminates at a point when addition of more genes failed to improve the classification accuracy of the current model. The crop of genes so selected at that point becomes the optimal gene biomarker for such microarray data.

The proposed algorithm was implemented in the R statistical environment using the **e1071** package.

III. Analysis And Results

The result of the proposed SVM algorithm for feature selection and classification of tumor samples with multiclass response for the simulated dataset is presented in this section.

In Table 2, the number and combination of gene subsets from the feature selection process with different FWER are presented. A grid search on each gene subsets to determine the optimal SVM parameter C and the kernel parameters were also obtained and the results are as presented in Table 3. The choice of the RBF and linear kernel was as reported in [14]. Possible range of values for the parameters was specified. Using a 10-fold cross validation on each of the gene subsets, the minimum misclassification error rate was then determined.

Having determined the optimal values for the parameters on each of the gene subsets, these values were use for the classification of tumor sample using the SVM classifier. The result of average misclassification error rates for the SVM algorithm with the appropriate kernel using the MCCV of 1000 runs on each subset is presented in Table 4.

The evaluation of the SVM algorithm on each subset in Table 3 suggest that the subset with 11 genes ($\alpha_F = 15\%$) provided the best correct classification rates among the subsets while the subset with 1000 genes ($\alpha_F = 100\%$) gives the worst result in terms of correct classification rate.

To optimize this set of genes selected, we ranked the first 25 best genes according to their p-values i.e. the gene with the least p-value was ranked first; the gene with the

second least was ranked second and so on. These genes were optimized by selecting a time the first 5, first 10, first 15, first 20 and first 25 genes in the ranked order into the SVM algorithm for classification of the response class based on RBF kernel. The classification results from this exercise are presented in Table 6.

Table 4: Result of the SVM method on each selected ranked gene subsets of the Simulated Data

Performance Measure	FWER (α_F)					
	1%	**5%**	**10%**	**15%**	**20%**	**100%**
No of genes →	9	10	10	11	11	1000
MER	0.2947	0.2949	0.2949	0.2774	0.2774	0.8145
CCR (%)	70.53	70.51	70.51	72.26	72.26	18.55
Sensitivity group 1	0.6218	0.6286	0.6286	0.6349	0.6349	0.3259
Sensitivity group 2	0.6550	0.6471	0.6471	0.6732	0.6732	0.3477
Sensitivity group 3	0.8965	0.8765	0.8765	0.8955	0.8955	0.3264
Specificity group 1	0.7890	0.7591	0.7591	0.7906	0.7906	0.6495
Specificity group 2	0.8593	0.8548	0.8548	0.8516	0.8516	0.6731
Specificity group 3	0.9272	0.9538	0.9538	0.9526	0.9526	0.6980
Positive Predicted Value group 1	0.6001	0.5702	0.5702	0.6102	0.6102	0.1766
Positive Predicted Value group 2	0.7196	0.7155	0.7155	0.7126	0.7126	0.2261
Positive Predicted Value group 3	0.8741	0.9176	0.9176	0.9178	0.9178	0.2425
Negative Predicted Value group 1	0.8036	0.8030	0.8030	0.8113	0.8113	0.6036
Negative Predicted Value group 2	0.8124	0.8122	0.8122	0.8240	0.8240	0.6289
Negative Predicted Value group 3	0.9630	0.9573	0.9573	0.9640	0.9640	0.6221

Table 5: Result of RBF and SVM tuning parameter for each of the ranked gene subsets for the simulated data

No of genes	γ	C	Minimum CV Error
5	0.05	1	0.3533
10	0.01	10	0.3000
15	0.01	1	0.2533
20	0.00001	10000	0.2000
25	0.1	1	0.1933

Table 6: Result of the SVM method on each selected ranked gene subsets of the Simulated Data

Performance Measure	Gene subset by p-value ranks				
No of genes →	5	10	15	**20**	25
MER	0.3783	0.2949	0.2311	**0.1887**	0.2280
CCR (%)	62.17	70.51	76.89	**81.13**	77.20
Sensitivity group 1	0.5319	0.6286	0.7614	**0.7701**	0.8068
Sensitivity group 2	0.5713	0.6471	0.7096	**0.8110**	0.6829
Sensitivity group 3	0.7980	0.8765	0.8585	**0.8605**	0.8625
Specificity group 1	0.7051	0.7591	0.7985	**0.8554**	0.7946
Specificity group 2	0.8144	0.8548	0.9312	**0.9135**	0.9329
Specificity group 3	0.9250	0.9538	0.9332	**0.9530**	0.9426
Positive Predicted Value group 1	0.4909	0.5702	0.6665	**0.7446**	0.6699
Positive Predicted Value group 2	0.6331	0.7155	0.8491	**0.8343**	0.8523
Positive Predicted Value group 3	0.8553	0.9176	0.8829	**0.9120**	0.8959
Negative Predicted Value group 1	0.7472	0.8030	0.8688	**0.8800**	0.8890
Negative Predicted Value group 2	0.7732	0.8122	0.8504	**0.9015**	0.8384
Negative Predicted Value group 3	0.9298	0.9573	0.9508	**0.9520**	0.9522

IV. Discussion Of Results

In this work, an efficient feature selection and classification method for multiclass response cancer tumor using Support Vector Machine has been proposed.

The propose feature selection procedure was able to detect 9 of the 10 simulated differentially expressed genes even at 1% FWER as presented in Table 1.

The optimal values of both the SVM and Kernel parameters with either the linear or RBF kernels (as the case may be) were equally determined using 10-fold cross validation for each of the gene subsets as presented in Table 3. The proposed SVM algorithm yielded good prediction accuracy (low misclassification error rates) with fewer biomarker genes than when all the genes were used for classification. This simply revealed the good parsimony property of the proposed method since it is capable to identify and select only the few core biomarkers that are present in microarray data.

The performance of the SVM algorithm with the RBF kernel using the optimal parameter values obtained via the grid search on each of the ranked gene subsets shows an increment in the accuracy from each of the selected gene subset up to twenty after which the performance of the classifier dropped with the inclusion of additional genes (i.e. first 5 genes, CCR=62.17%, first 10 genes, CCR= 70.51%, first 15 genes, CCR= 76.89%, first 20 genes, CCR= 81.13% and first 25 – genes, CCR= 77.2%) as shown in Table 6. The addition of five genes (first 25) reduces the predictive accuracy of the SVM algorithm. This indicates the presence of noisy genes in the selected gene subset which has negatively affected the performance of the classifier as earlier mentioned. This performance behavior of the proposed method is clearly shown by Fig 2.

V. Conclusion

An efficient algorithm that is based on SVM technique which optimizes feature selection process and improves the classification of cancer tumor samples in multiclass response microarray cancer data has been presented in this study. Various results obtained showed the efficiency of the new proposed algorithm as it is capable at selecting the core relevant gene biomarkers that significantly enhance the predictive accuracy of the response class in any high-dimensional micro array data.

In other words, the propose algorithm has efficient filtering mechanism to filter out several irrelevant genes variables with very week expression levels that often characterize any typical microarray cancer data.

Results of the Monte-Carlo study reported in this work shall be validated with a number of published cancer data sets for general applicability.

Finally in future work, the performance of the proposed method here shall be compared some of the machine learning methods in the literature to determine its relative efficiency over others.

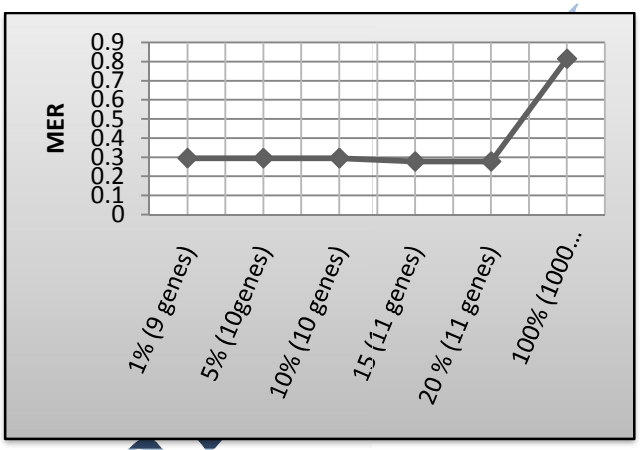

Figure1: Graph of MER for the simulated multiclass data at different FWER along with the number of gene selected in parenthesis.

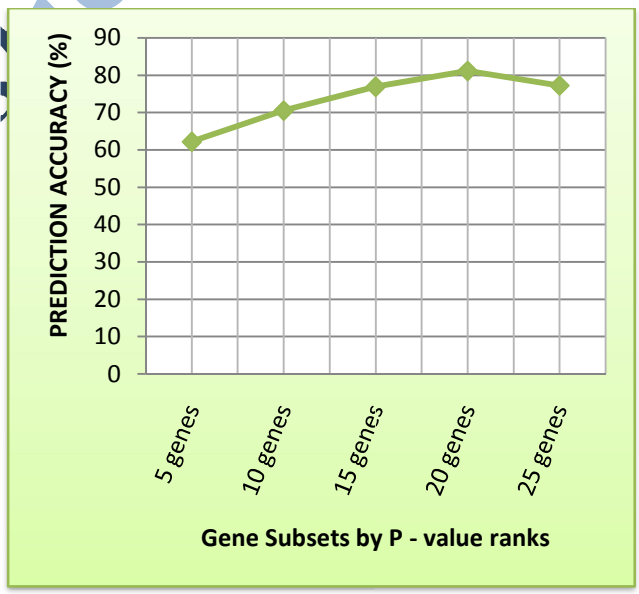

Figure 2: Graph of Prediction Accuracy (%) for the simulated multiclass data at different p-value rank gene subsets.

References

[1] A. Hapfelmeier, W. B. Yahya, R. Rosenberg, and K. Ulm, "Predictive Modeling of gene Expression data", In: *Handbook of Statistics in Clinical Oncology*, Chapman and Hall/CRC, New York, (2012), pp. 463-475.

[2] A. W. Banjoko, W. B. Yahya, M.K. Garba, O. R. Olaniran, K.O. Olorede, and K.A. Dauda, "Efficient Support Vector Machine Classification of Diffuse large B-Cell Lymphoma and Follicular Lymphoma mRNA Tissue Samples", *Annals. Computer Science Series*, Vol. 13 (2015a), pp. 69 – 79.

[3] A.W. Banjoko, W. B. Yahya, and M. K. Garba, "Efficient Support Vector Machine Classification of Diffuse Large B-Cell Lymphoma and Follicular Lymphoma mRNA Tissue Samples". *Proceedings of the14th Regional Scientific Conference of the International Biometric Society – group Nigeria, Nigeria*, (2015d).

[4] A.W. Banjoko, W. B. Yahya and M. K. Garba, "Support Vector Machine for Feature Selection and Classification of Small Node–Negative Breast Carcinomas". *Proceeding of the 3rd International Conference of the U6 Consortium, Nigeria*, (2015c).

[5] A.W. Banjoko, W. B. Yahya and M. K. Garba, "Efficient Support Vector Machine Method for Tissue Samples Classification in Colon Cancer Genomic data". *Proceedings of the 34th Annual Conference of The Nigeria Mathematical Society, Nigeria*, (2015b).

[6] B. L. Welch, "The generalization of "student's" problem when several different population variances are involved". *Biometrika*, Vol. 34, (1947), pp. 28-35.

[7] G. James, D. Witten, T. Hastie, R. Tibshirani, "An *Introduction to Statistical Learning with Applications in R*", Springer Science + Business Media, New York, (2013).

[8] G. T. Aremu, W. B. Yahya, "Competing Algorithms For Microarray Based Multiclass Sequential Feature Selection and Classification*", Proceedings of 4th International Science, Technology, Education, Arts, Management & Social Sciences (iSTEAMS) Research Nexus Conference, Nigeria*, (2015), pp. 675 – 682.

[9] N. Cristianini and J. Shawe-Taylor, "*An introduction to Support Vector Machines*", Cambridge University Press, United Kingdom, (2012).

[10] P. Cichosz, "*Data mining algorithms explained using R*", John Wiley & Sons,New York., (2015).

[11] R. R. Witold. Rudnicki, W. Mariusz and P. Wiesław,"All Relevant Feature Selection Methods and Applications", *Studies in Computational Intelligence*, 584, (2015), pp. 11 – 28.

[12] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisenh, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and Anne-LiseBørresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications", *Proceeding of the National Academy of Sciences of the United State of America* (PNAS), Vol. 98, (2001), pp. 10869 – 10874.

[13] V. N. Vapnik, "*The Nature of Statistical Learning Theory*", Springer-Verlag, New York, (1995).

[14] W. B. Yahya, "*Genes selection and Tumour Classification in Cancer Research*: *A new approach*", Säbruck, Germany: Lambert Academic Publishing (2012).

[15] W. B. Yahya, "*Sequential dimension reduction and prediction methods with high dimensional microarray data*", Universitätsbibliothek, Ludwig Maximilians-Universität, München, Germany, Ph.D. Thesis, (2009), URL:http://edoc.ub.uni - muenchen.de/10254/.

[16] W. B. Yahya, G. T. Aremu, M. K. Garba, "Multiclass Sequential Feature Selection and Classification Method for Gene Expression Data", *Journal of Applied Science and Technology*, 20(1&2), (2015), pp. 50 – 61.

[17] W. B. Yahya, K.Ulm, F. Ludwig, A. Hapflemeir, "*k*-SS: A sequential feature selection and prediction method in microarray study*", International Journal of Artificial Intelligence*, spring,6(S11), (2011), pp. 19- 47.

[18] W. B. Yahya, M. O. Oladiipo, E. T. Jolayemi, "A fast algorithm to construct neural networks classification models with high-dimensional genomic data". *Annals. Computer Science Series*, 10, (2012), pp. 39- 58.

[19] W. B. Yahya, R. Rosenberg, K. Ulm, "Microarray-based Classification of Histopathologic Responses of Locally Advanced Rectal Carcinomas To Neoadjuvant Radio-chemotherapy Treatment", TurkiyeKlinikleri Journal of Biostatistics, 6(1), (2014), pp. 8- 23.

[20] Z. K. Sidak, "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions". *Journal of the American Statistical Association*, 62, (1967), pp.626–633.