Turkish Journal of Science & Technology
15(1), 1-12, 2020

An Alternative Approach to Variable Selection Using Hybrid Regression Modeling in Undersized Sample Data

Esra PAMUKÇU*
Department of Statistics, Faculty of Science, Fırat University, Elazığ, Turkey
epamukcu@firat.edu.tr

(Geliş/Received: 11./12/2019; Kabul/Accepted: 09/02/2020)

Abstract: The problems encountered in the analysis of data sets with undersized sample mainly arise from the singular covariance structure. As a solution to this problem, non-singular Hybrid Covariance Estimators (HCEs) have been proposed in the literature. Several multivariate statistical techniques where HCEs are used continue to be developed and introduced. One of these is the Hybrid Regression Model (HRM). Thanks to HCEs, since there is no longer the rank problem in covariance matrix, in HRM analysis the regression coefficients can be estimated as many as the number of variables. However, determining the best predictors in regression model is one of the biggest problems for researchers since the number of variables increases and there is insufficient knowledge about the model. Therefore, some numerical optimization techniques and strategies are required to explain such a wide solution space where the number of alternative subsets of candidate models of predictors can reach millions. In this paper, we introduced a new and alternative approach to variable selection for undersized sample data by using the Genetic Algorithm (GA) and Information Complexity Criteria (ICOMP) as a fitness function in the HRM analysis. To demonstrate the ability of proposed method, we carried out the Monte Carlo simulation study with correlated and undersized data sets. We compared our method with Elastic Net (EN) modeling. According to results, the proposed method can be recommended as an alternative approach to select variable in undersized sample data.

Key words: Genetic algorithm, hybrid regression model, information complexity criteria, variable selection, undersized sample problem

Cılız Örneklem Problemine Sahip Veri Setlerinde Hibrit Regresyon Modellemesi Kullanarak Değişken Seçimine Alternatif Bir Yaklaşım

Öz: Cılız örneklem problemine sahip veri setlerinin analizinde karşılaşılan problemler temel olarak singüler kovaryans matrisinden kaynaklanmaktadır. Bu probleme bir çözüm olarak literatürde Hibrit Kovaryans Tahmin Edicileri (HCE) önerilmiştir. HCE'lerin kullanıldığı bazı çok değişkenli istatistiksel yöntemler geliştirilmeye ve tanıtılmaya devam etmektedir. Bunlardan biri Hibrit Regresyon Modeli'dir (HRM). HCE sayesinde kovaryans matrisinde artık singülerlik problemi olmadığı için, HRM ile değişken sayısı kadar regresyon katsayısı tahmin edilebilir. Bununla birlikte, değişken sayısı çok fazla olduğu ve model hakkında yetersiz bilgiye sahip olunduğu için, regresyon modelindeki en iyi tahmin edicileri belirlemek araştırmacılar için en büyük problemlerden biridir. Bu nedenle tahmin edicilerin alternatif modellerinin alt küme sayısının milyonları bulabildiği böyle geniş bir çözüm uzayını açıklamak için bazı nümerik optimizasyon tekniklerine ve stratejilerine ihtiyaç vardır. Bu çalışmada, ICOMP'ın uygunluk fonksiyonu olarak kullanıldığı bir Genetik Algoritma yapısı ile HRM analizi yapılarak cılız örneklemli veri setleri için değişken seçimine alternatif bir yaklaşım önerilmiştir. Önerilen yaklaşımın kullanılabilirliğini göstermek için korelasyonlu ve cılız örneklemli veri setlerinin kullanıldığı bir Monte Carlo simulasyon çalışması yapılmıştır. Karşılaştırma amacıyla Elastik Net modellemesi kullanılmıştır. Elde edilen sonuçlara göre, önerilen yaklaşımın cılız örneklemli veri setlerinde değişken seçimi için alternatif bir yaklaşım olarak kullanılabileceği söylenebilir.

Anahtar kelimeler: Genetik algoritma, hibrit regresyon modeli, bilgi karmaşıklığı kriterleri, değişken seçimi, cılız örneklem problemi

1. Introduction

According to statistical viewpoint, if inferences about data are required, it is expected that the number of samples should increase exponentially against the number of variables. Nowadays, even if there are many observations, the number of variables may increase radically. In this case, a single observation can have thousands or even millions of dimensions whereas the number of observations that can be reached for the study is expressed as ten or hundred. The traditional techniques in statistics are not capable for analyzing such data sets [1].

Statisticians sometimes say "Big p, Small n" for this problem. Another definition is "the undersized sample problem" or "extremely small sample problem" [2,3].

* Corresponding author: epamukcu@firat.edu.tr ORCID: 0000-0002-5778-9626

The problems encountered in the analysis of data sets with undersized sample mainly arise from the singular covariance structure. As a solution of this problem, for the first time by [4], non-singular Hybrid Covariance Estimators (HCEs) is proposed by hybridizing the Maximum Likelihood Estimator (MLE) with smoothed covariance structures after the stabilization stage of eigenvalues.

Since HCEs overcome the singularity in covariance matrix, the analysis of undersized or high-dimensional data sets with multivariate statistical methods for $\boldsymbol{n << p}$ problem has become possible. Several multivariate statistical techniques in where HCEs are used continue to be developed and introduced [4,5,6]. One of these is the Hybrid Regression Model (HRM) in which HCEs is used as covariance input [7]. When the HRM is performed in case $\boldsymbol{n << p}$, the regression coefficients can be obtained as many as the number of variables even if they are hundreds or thousands. The next stage is to determine the best predictors that have the most effect on the response variable, i.e. the variable selection stage. This is one of the biggest problems for researchers when the number of predictors increases and there is insufficient or no prior knowledge about the model.

Model selection is basically a process of finding the best model from the subset of competing or candidate models by revealing which variables are effective on the response variable. Since the early 1970s, it has been possible to come across many studies on the model selection algorithm and criteria. These include classical methods for the model selection and the methods based on information criteria. The classical methods are generally performed by hypothesis testing. An arbitrary level of confidence is selected in a hypothesis testing process by the researcher to decide whether the variables are included in final model or not. However, most statisticians and other scientists have emphasized that the level of confidence used in the selection of classical models is baseless [8]. Some scientists have stated that the hypothesis testing approach does not have a theoretical accuracy and it is generally insufficiently valid [9]. Although almost all popular statistical packages have many classical model selection procedures based on hypothesis testing, such as Forward Selection, Backward Elimination, and Stepwise techniques, these selection methods do not deal with the dependency structure between variables. Also, since they contain the limitations of the hypothesis testing, these methods are criticized by Boyce et al. (1974, p:16) with the following words "These approaches do not guarantee optimal results". Therefore, the selection of the best predictors involves randomness. He mentioned that also "An exhaustive search would examine $\boldsymbol{2^p}$ possible equations" [10].

The shortcomings in the classical procedures for model selection impose limitations on the selection of the best or near-best model subsets. Although some statisticians and researchers propose to choose from all possible subsets, in many cases this method is not producible in terms of calculation and is also not possible in terms of time and cost [11]. For example, if we have p=20, the number of the subsets of the candidate models are $\boldsymbol{2^{20} = 1048576}$. Therefore, some numerical optimization techniques and strategies are required to explain such a wide solution space.

In general, two components are needed to use numerical techniques in a subset selection problem.

- An algorithm to effectively scan wide solution space
- A measure for comparison of candidate models

In this study, for the HRM, it is constructed a Genetic Algorithm (GA) structure where the variables in the model assigned as '1' and the others variables assigned as '0'. By using ICOMP criterion as the fitness function, it is compared the candidate regression models in population for transferring to the next generation.

The sketch of paper is given by following orders. In Section 2.1, HRM and its background will be introduced. In Section 2.2, the GA structure will be presented in order to explain how to determine the variables exist in the model by using the GA in the HRM. In Section 2.3, we briefly introduce the Elastic Net (EN) modeling. In Section 3, we provide our Monte Carlo simulation study. The last part is divided to conclusion and brief discussion.

2. Material and Methods

2.1. Hybrid regression model (HRM) with information complexity and hybrid covariance estimators

Let us consider the multiple linear regression model in matrix form given by

$$y = X\beta + \varepsilon \tag{1}$$

where y_{nx1} is vector of observaitons on a response variable, X_{nxp} is full rank matrix of variable, β_{px1} is regression coefficient vector and $\varepsilon_{nx1} \sim N(0, \sigma^2)$ is random error vector. The maximum likelihood estimates of β_{px1} are given by

$$\hat{\beta}_{MLE} = (X'X)^{-1}X'y \tag{2}$$

As is seen in equation (2), the Gram matrix $(X'X)$ must be non-singular and invertible in order to obtain the estimate of $\hat{\beta}_{MLE}$. In case of undersized sample, i.e. $n \ll p$, it is clear that the X_{nxp} is not full rank matrix and the Gram matrix is not invertible. As a solution to this problem, in the framework of the regression analysis it has been proposed Hybrid Covariance Estimators (HCEs) as a well-conditioned and non singular covariance estimate instead of using Gram matrix [4]. In the following section, HCEs will be introduced.

2.1.1. Smoothed covariance structures

For a covariance matrix, the Condition Number (CN) defined as the largest eigenvalue λ_{max} divided by the smallest eigenvalue λ_{min} is given by

$$CN = \lambda_{max}/\lambda_{min} \tag{3}$$

The inverse of CN can be used for the definition of singularity of covariance matrix [12] and it is defined as follow,

$$\kappa(\Sigma) = \frac{1}{CN} \tag{4}$$

If $\kappa(\Sigma)$ is near to zero, the covariance matrix is close to the singularity. As a solution to singularity, shrinkage estimators that will shrunk eigenvalues of Σ to a central value have been developed. The main idea of these estimators is to take the convex combination of the maximum likelihood estimation of the sample covariance, i.e. $\hat{\Sigma}_{MLE}$, and a target diagonal matrix $\widehat{D}$ (i.e., taking the weighted average). Then, the shrinkage or smoothed covariance estimator is given by

$$\hat{\Sigma}_S = (1 - \hat{\rho})\hat{\Sigma}_{MLE} + \hat{\rho}\widehat{D} \tag{5}$$

where, $\hat{\rho} \in [0,1]$ is estimate of the optimal shrinkage coefficient ρ, The $\widehat{D}$ matrix is called as shrinkage target matrix and its naive form is as follow,

$$\widehat{D} = \frac{tr(\hat{\Sigma}_{MLE})}{p} I_p = \left(\frac{1}{p}\sum_{j=1}^{p}\lambda_j\right) I_p = \bar{\lambda} I_p \tag{6}$$

where $tr(.)$ represents the trace of matrix, $\lambda_j, j = 1,2,\dots,p$ are eigenvalues of sample covariance matrix and $\bar{\lambda} = \sum_{i=1}^{p}\lambda_i/p$ is arithmetic mean of eigenvalues. By using the weighted average in equation (5), it is placed to a lower weight on extremely large or small eigenvalues. Thus, the effect of these eigenvalues is reduced and a smoothed estimator is obtained.

The smoothed or regularized covariance estimators under linear or quadratic loss functions have been introduced in the literature. Some of them: Empirical Bayes Estimator (EB)[13], Stipulated Ridge Estimator (SRE)[14], Stipulated Diagonal Estimator (SDE)[14], Convex Sum Estimator (CSE)[15,16], Bozdogan's Convex Sum Estimator (BCSE)[17], Oracle Approximation (OAS)[18], Ledoit-Wolf Estimator (LW)[19] are given in Table 1.

Table 1. Smoothed or regularized covariance estimators

$\hat{\Sigma}_S$	**Form***	$\hat{\boldsymbol{\rho}}$
$\hat{\Sigma}_{EB}$	$\hat{\Sigma}_{MLE} + \hat{\rho}_{EB} D_p$	$\frac{p-1}{n(\hat{\Sigma}_{MLE})}$
$\hat{\Sigma}_{SRE}$	$\hat{\Sigma}_{MLE} + \hat{\rho}_{SRE} D_p$	$p(p-1)\left[2ntr(\hat{\Sigma}_{MLE})\right]^{-1}$
$\hat{\Sigma}_{SDE}$	$(1-\hat{\rho}_{SDE})\hat{\Sigma}_{MLE} + \hat{\rho}_{SDE} D_p$	$p(p-1)\left[2ntr\left(\hat{\Sigma}_{MLE}{}^{-1}\right) - p\right]^{-1}$
$\hat{\Sigma}_{CSE}$	$\hat{\rho}_{CSE}\hat{\Sigma}_{MLE} + (1-\hat{\rho}_{CSE}) D_p$	$\frac{n}{n+m}$ where $0 < m < \frac{2[p(1+\beta)-2]}{p-\beta}$, $\beta = \frac{tr(\hat{\Sigma})^2}{tr(\hat{\Sigma}^2)}$ for $p \geq 2$ dimensions.
$\hat{\Sigma}_{BCSE}$	$\hat{\rho}_{BCSE}\hat{\Sigma}_{MLE} + (1-\hat{\rho}_{BCSE}) D_p$	$\frac{1}{\alpha}$, where $\alpha = \frac{1}{n-1}\sum_{j=1}^{p} \sigma_{jj}$
$\hat{\Sigma}_{OAS}$	$(1-\hat{\rho}_{OAS})\hat{\Sigma}_{MLE} + \hat{\rho}_{OAS} D_p$	$min\left(\frac{\left(1-\frac{2}{p}\right) tr(\hat{\Sigma}^2) + (tr\hat{\Sigma})^2}{(n+1-\frac{2}{p})\left[tr(\hat{\Sigma}^2) - \frac{(tr\hat{\Sigma})^2}{p}\right]}, 1\right)$
$\hat{\Sigma}_{LW}$	$(1-\hat{\rho}_{LW})\hat{\Sigma}_{MLE} + \hat{\rho}_{LW} D_p$	$min\left(\frac{\sum_{i=1}^{n} \|x_i' x_i - S\|^2_F}{n^2\left[tr(\hat{\Sigma}^2) - \frac{(tr\hat{\Sigma})^2}{p}\right]}, 1\right)$

*: n is number of observations, p is number of variables in data, D_p is target matrix

2.1.2. Hybrid covariance estimators (HCEs)

Since the problem of interest of this study is able to perform regression analysis in data sets with undersized sample, it is clear that the maximum likelihood estimates of covariance matrices of these data sets is singular and/or ill conditioned. It is important to obtain a well-conditioned and non-singular covariance structure. From this point, Pamukcu et al. (2015) [4] developed Hybrid Covariance Estimators-HCEs by using following eigenvalue stabilization algorithm defined by Thomaz [20]

- Step-1: Calculate the eigenvalues λ_j and their eigenvectors v_j of $\hat{\Sigma}_{MLE}$, where $j = 1,2,\dots,p$ and p is the number of variables of the data. $\hat{\Sigma}_{MLE}$ represents the maximum likelihood covariance estimator.
- Step-2: Calculate the arithmetic mean of eigenvalues by using: $\bar{\lambda} = \frac{1}{p}\sum_{j=1}^{p} \lambda_j$
- Step-3: Produce the following matrix of eigenvalues based on largest dispersion values:

$$\Lambda^* = \begin{bmatrix} max(\lambda_1, \bar{\lambda}) & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & max(\lambda_p, \bar{\lambda}) \end{bmatrix} \quad (7)$$

- Step-4: The new stabilized covariance matrix is given by:

$$\hat{\Sigma}_{MLE\_STA} = V \Lambda^* V \quad (8)$$

where V is matrix of eigenvectors v_j of $\hat{\Sigma}_{MLE}$. As is seen, the algorithm stabilizes eigenvalues by expanding only the smaller and consequently less reliable eigenvalues of the covariance matrix and by keeping its larger eigenvalues unchanged. Then, one can obtain HCEs by following two stage process:

- Stage-1: Perform the stabilization algorithm above
- Stage-2: Produce the stabilized-smoothed covariance estimators in Table 1.

For example, $\hat{\Sigma}_{HCE} = \hat{\Sigma}_{MLE\_STA\_BCSE}$ is defined as follow

$$\hat{\Sigma}_{MLE\_STA\_BCSE} = (1 - \hat{\rho}_{BCSE})\hat{\Sigma}_{MLE\_STA} + \hat{\rho}_{BCSE} D_p \tag{9}$$

For more on this we refer the readers to [4,21].

The logical and mathematical theme of using Stabilization + Regularization is to achieve positive definition with respect to shrunk and to expand the less appropriate and smaller eigenvalues by using stabilizing. Considering Table 1, there are several forms of HCEs which can be used. It is an important problem which one is optimal covariance for a researcher who wants to use them. For this reason, it is investigated for the choosing optimal covariance structure by [4,5,6,21] especially in the framework of regression analysis and observed that $\hat{\Sigma}_{MLE\_STA\_CSE}$ and $\hat{\Sigma}_{MLE\_STA\_BCSE}$ have superior performance on the others. Therefore, in our simulation study, we use $\hat{\Sigma}_{MLE\_STA\_CSE}$ and $\hat{\Sigma}_{MLE\_STA\_BCSE}$ covariance structures and their model are represented by HRM1 and HRM2, respectively.

2.1.3. Information complexity criteria (ICOMP) for model selection

In the literature, Akaike [22,23] Information Criterion (AIC) is widely used for statistical model selection. This is given by

$$AIC = -2logL(\theta) + 2k \tag{10}$$

where $logL(\theta)$ is log likelihood function of θ parameter in a probability density function, k is the number of parameters. ICOMP (I; Information-COMP; Complexity) are criteria developed by Bozdogan (1988) for the model selection in multivariate linear and nonlinear models [24]. Wheras the AIC is only intended to strike a balance between the lack of fit and the penalty terms, ICOMP aims to establish this balance by taking into account a complexity measure that measures how the parameters in the model relate to each other. Therefore, instead of directly penalizing the number of parameters, it penalizes the covariance complexity of the model introduced by [24]. ICOMP is given by

$$ICOMP = -2logL(\theta) + 2C_1(\Sigma) \tag{11}$$

The second part of equation 11 is called the measure of complexity of model. It is given as follow:

$$C_1(\Sigma) = \frac{p}{2} log\left[\frac{tr(\Sigma)}{p}\right] - \frac{1}{2} log|\Sigma| \tag{12}$$

where $|\Sigma|$ represents determinant of Σ and p is dimension of Σ. As seen, $C_1(\Sigma)$ include two simplest scales of multivariate scattering called determinant and trace in a single function. There are several forms ICOMP criteria, for more about this we refer the reader to [11,25,26,27]. For these criteria, the model which has minimum value of criteria is called as the best model.

2.1.4 Hybrid regression model

In the case of undersized sample, i.e. $n << p$, where the sample variance covariance matrix is singular, we can analyze the data with HRM by using following steps:

- Step-1: $\hat{\Sigma}_{HCEs} = \hat{\Sigma}_{MLE\_STA\_CSE}$ and $\hat{\Sigma}_{HCEs} = \hat{\Sigma}_{MLE\_STA\_BCSE}$ are estimated for the data set
- Step-2: $\hat{\Sigma}_{HCEs}$ are used instead of Gram matrix in multiple regression analysis
- Step-3: The model which has minimum value of information criteria is determined as the best model among the candidate models created with different $\hat{\Sigma}_{HCEs}$.

Indeed, AIC, BIC [28] and Consistent Akaike Information Criterion (CAIC) [29] may be used as information criteria for model selection tool. Specifically, we prefer to use ICOMP in our computations since it is demonstrated

that ICOMP has superior performance on the other criteria. Also, it has been demonstrated that they tend to select more variables that may be related to each other since their penalty terms are not deal with correlation structure between variables [11,17,21,26,27]. The derived form of ICOMP in HRM is defined as follows:

$$ICOMP_{Miss}(HRM) = nlog(2\pi) + nlog(\hat{\sigma}^2) + n + 2C_1\left(Cov(\hat{\beta}_{HRM})_{Missspec}\right) \tag{13}$$

where $Cov(\hat{\beta}_{HRM})_{Missspec} = \hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\hat{\mathcal{F}}^{-1}$is called the "sandwiched covariance" estimator. $\hat{\mathcal{F}}^{-1}$ and $\hat{\mathcal{R}}$ represent the inverse of Fisher information matrix and outer-product form of Fisher information matrix, respectively. These are given by

$$Cov(\hat{\beta}_{HRM})_{Missspec} = \begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}\begin{bmatrix} \frac{n}{\hat{\sigma}^2} & \frac{nS_k}{2\hat{\sigma}^3} \\ \frac{nS_k}{2\hat{\sigma}^3} & \frac{n(K_t-1)}{4\hat{\sigma}^4} \end{bmatrix}\begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix} \tag{14}$$

is called Sandwich covariance matrix and where

$$S_k = Skewness\ coefficient = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^3\right)}{\hat{\sigma}^3} \tag{15}$$

$$K_t = Kurtosis\ coefficient = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^4\right)}{\hat{\sigma}^4} \tag{16}$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$and n is the number of observations. When the model is correct we get $\hat{\mathcal{F}}^{-1} = \hat{\mathcal{R}}$ and the formula reduces to the usual inverse Fisher information matrix

2.2. Genetic algorithm for hybrid regression model

After performing HRM to undersized sample data, since the regression coefficients are obtained as many as the number of variables, many of these could be irrelevant or redundant variables. So, the performance of the HRM can be evaluated after they are detected and sorted.

Let p be the number of variables in HRM. If the p increases, the number of subsets of candidate models can reach millions or even billions. Therefore, an effective algorithm is necessary to effectively scan the such a wide solution space. For this purpose, we prefer to use the principles of GA in order to select best predictors in HRM.

The GA is an evolutionary search algorithm that borrows concepts from biological evolution and is a stochastic optimization method inspired by the principles of evolution in nature. The search method is based on the principle of survival of the best. To this end, we begin to work with a community of potential solutions to the problem to be solved. It is called as initial population. Each individual in a population is a potential solution and is coded as chromosomes in accordance with the nature of the problem being studied.

For the variable selection problem in HRM analysis with GA, our implementation basically follows Goldberg's GA (1988) [30]. For a detailed information we refer the readers to [30,31].

Considering in the framework of the regression analysis, the length of the chromosomes is equal to the number of all the variables in data set. Let *k* be the number of the variables. Assuming that there are 10 variables in the data set, i.e. $k = 10$. In this case, a chromosome is a sequence of *k* units. Each unit on the chromosome is called a "gene" and each unit can have a value of '0' or '1'. '0' on the chromosome represents that the corresponding variables are not included in the regression model and '1' represents the included variables in the regression model. A chromosome sample is given below for the regression model in which the variables 1,2,4,5,6,8 and 10 are included [11,27].

Variables	1 2 3 4 5 6 7 8 9 10
Chromosome	1 1 0 1 1 1 0 1 0 1

The general procedure in the GA is straightforward and it is summarized as follow:

1. Generate the initial population of chromosomes
2. Calculate the fitness value for each chromosome in the population: The fitness value of the i[th] member is $f\ (i^{th})$ which is the value of the objective function f at that point [32]. For each chromosome, there is a

numerical fitness value that is proportional to the use or ability of the solution represented by the chromosome. This information guides the selection of more appropriate solutions for each generation. In this study, ICOMP is our fitness function. Firstly, a chromosome will be modeled by using HRM and the performance of the model will be measured by ICOMP. Since the model which has minimum value of ICOMP is defined as probably good model, the chromosomes will be transferred to the next generation.

3. Determine how current population is matched for the next generation: Firstly, it is sorted current population according to the values of fitness function. In our case, the most popular chromosomes which has minimum ICOMP values come to in the beginning of the list. Chromosomes in the list are mated by using a sequential pairs (pair (1,2), pair (3,4)...etc.).
4. Perform the GA operations: By changing the structure of chromosomes with the genetic processes such as mutation and crossover, it is provided to investigate the space of all possible solutions. There are three crossover techniques such as single, multiple and uniform crossover. However, these are related to the processing of the GA and no further details will be given.
5. Pass on offspring to new generation.
6. Loop back to step 2 until stopping criteria met: These steps given above are performed only for one generation. After we reach to a certain number of generations or to optimal value of fitness function, the process is terminated. In our case, there is no an optimal value for the fitness function, i.e. ICOMP, therefore the algorithm continue until the number of generations is reached.

Note that, each iteration in the GA is called as generation. The chromosome with the highest fitness in the final generation is proposed as a solution to the problem. It is expected that this proposed solution is optimal or near to the optimal solution.

2.3 Elastic net modeling

Let us recall the multiple regression model in matrix form given equation (1). The objective function of the vanilla matrix representation of EN is defined by

$$L(\beta, \lambda_1, \lambda_2) = \|y - X\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 \tag{17}$$

where $\|\beta\|_1 = \sum|\beta|$ is the L_1 norm penalty (the LASSO penalty) and $\|\beta\|_2^2 = \sum\beta^2$ is the L_2 norm penalty (the ridge penalty) To avoid the double shrinkage in equation (17), one can undo the penalty by scaling up the estimates from the vanilla matrix representation of EN. Then an improved estimator is

$$\hat{\beta} = \sqrt{1+\lambda_2}\tilde{\beta} \tag{18}$$

Further, one can show that the improved estimates in matrix form are given by [33]

$$\hat{\beta}_{EN} = arg\min_{\beta}\left[\beta'\left(\frac{X'X+\lambda_2 I}{1+\lambda_2}\right)\beta - 2y'X\beta + \lambda_1\|\beta\|_1\right] \tag{19}$$

It is clear that

$$\left(\frac{X'X+\lambda_2 I}{1+\lambda_2}\right) = \frac{1}{1+\lambda_2}X'X + \frac{\lambda_2}{1+\lambda_2}I \tag{20}$$

We can write the equation (20) equivalently as

$$\left(\frac{X'X+\lambda_2 I}{1+\lambda_2}\right) = (1-\hat{\rho}_{EN})X'X + \hat{\rho}_{EN}I \tag{21}$$

where $\hat{\rho}_{EN} = \frac{\lambda_2}{1+\lambda_2} \epsilon (0,1)$ is shrinkage coefficient and $X'X$ is the Gram matrix.

3. Results

3.1. Monte Carlo simulation study

In this part, we present a Monte Carlo simulation study on correlated data sets with undersized sample. According to explanations about HCEs performance in Section 2.1.2, we study two different HCEs in HRM to able to select the best predictors by using GA and ICOMP. For comparison, we study EN modeling with different three λ_2 tuning parameters as grid values. These are given in Table 2.

Table 2. Compared Models in Simulation Study

Models	With
HRM1	$\hat{\Sigma}_{MLE\_STA\_CSE}$
HRM2	$\hat{\Sigma}_{MLE\_STA\_BCSE}$
EN1	$\lambda_2 = 0.001$
EN2	$\lambda_2 = 0.01$
EN3	$\lambda_2 = 0.1$

The simulation protocol for generating correlated data sets with undersized sample is below:

- p : Number of variables=30, 40, 50, 60
- n : Number of observations=10, 20, 30
- $X_{(nxp)}$: The data set
- r : Correlation between the correlated variables = fixed as 0.5
- σ : Error variance = fixed as 0.01
- ε : Error

Note that the data sets are derived from the multivariate normal distribution with zero mean and Σ covariance matrix, i.e. $MVN(0, \Sigma)$.For $r = 0.5$ and $p = 5$, an example to calculate the variance-covariance matrix Σ is as follow:

$$\Sigma = \begin{bmatrix} 1 & r & \frac{r}{2} & \frac{r}{4} & \frac{r}{8} \\ r & 1 & r & \frac{r}{2} & \frac{r}{4} \\ \frac{r}{2} & r & 1 & r & \frac{r}{2} \\ \frac{r}{4} & \frac{r}{2} & r & 1 & r \\ \frac{r}{8} & \frac{r}{4} & \frac{r}{2} & r & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.500 & 0.250 & 0.125 & 0.062 \\ 0.500 & 1 & 0.500 & 0.250 & 0.125 \\ 0.250 & 0.500 & 1 & 0.500 & 0.250 \\ 0.125 & 0.250 & 0.500 & 1 & 0.500 \\ 0.062 & 0.125 & 0.250 & 0.500 & 1 \end{bmatrix} \quad (22)$$

As is seen, the covariance structure in equation (22) provides correlation between sequential variables. We generate the data sets with different samples and variables size. After having the data sets, we use following steps in order to show the performance of detection of true variables of proposed method and EN modeling, for comparison.

- We generate the response variable by using $y = 5 + 3X_1 - 2X_2 + 6X_3 + \sigma\varepsilon$. So, the variables $\{X_0, X_1, X_2, X_3\}$ are important variables which have effect on response variable.
- *'elasticnet'* module in MATLABR2015a © is used to perform EN modeling. It can be easily found in https://in.mathworks.com/matlabcentral/mlcdownloads/downloads/submissions/58182/versions/2/previews/elasticnet.m/index.html
- To perform GA with HRM and ICOMP, a graphical user interface GUI is written by the author following the steps in Section 2.2.
- In order to compare the performance of detection of true variables of models, we use the average value of True Positive Rate (TPR) [34] given by

$$TPR = \frac{1}{r}\sum_{q=1}^{r} TPR^q \tag{23}$$

where $TPR^q = \frac{\#truly\ selected\ variables}{\#total\ variables}$ for the models in q^{th} run and r is replication number.
We note that the TPR value is equal to 1, if the list of total selected variables is $\{X_0, X_1, X_2, X_3\}$, otherwise $TPR\epsilon[0,1)$. If it is added an unimportant variable to model, then the TPR value is equal to 4/5=0.8. Additionally, we use following GA parameters in Table 3 in our simulation study. This process for all data sets with different size is repeated 100 times. The simulation results are given in Table 4:

Table 3. GA parameters

Parameter	**Value**
Generation number	10
Population size	50
Generation seeding	Sorted
Probability of crossover	0.75
Type of crossover	Two-points
Probability of mutation	0.1
Objective functions	ICOMP

Table 4. The Simulation results according to TPR values

Data size	**HRM1**	**HRM2**	**EN1**	**EN2**	**EN3**	**CPU time (sec)**
10x30	0.1723	**0.2643**	0.1664	0.1697	0.1814	42.1
10x40	0.1116	**0.1582**	0.0994	0.1017	0.1373	44.5
10x50	0.0862	0.1283	0.1003	0.0933	**0.1323**	**47.9**
20x30	**0.2748**	0.2054	0.1067	0.1040	0.1575	42.9
20x40	**0.2423**	0.1776	0.1165	0.1151	0.1242	47.1
20x50	0.1333	**0.1478**	0.0895	0.0911	0.0946	48.2
30x40	**0.2078**	0.1712	0.0837	0.0848	0.0875	45.4
30x50	0.1396	**0.1475**	0.0894	0.0861	0.0921	52.5
30x60	0.0330	0.0541	0.0727	**0.0742**	0.0740	52.1

We note that the highest TPR value the models are indicated as bold in the Table 4. According to results, we observe that our proposed method has superior performance against EN modeling. It is clear that the performance of EN modeling depend on the λ_2 tuning parameter. How to choose the λ_2 tuning is an important problem in EN. In practice, it is difficult to assign which value of the λ_2 tuning parameter is appropriate. Often the λ_2 tuning parameter is fixed arbitrarily at the beginning. In the literature cross-validation (CV) (5-fold, or 10-fold) has been used which is a time consuming operation. To avoid time consuming operation, we used some grid values for λ_2 tuning parameter. Otherhand, the computation of the HCEs covariance matrix and to perform HRM analysis with GA is fast in terms of the CPU time and it is not heavy. We should emphasize the values of CPU time in Table 4 are belong to whole simulation study. Compared to EN modeling, which is current and frequently used method, the HRM yields better results and it can be suggested as an alternative approach in variable selection for undersized sample data.

4. Conclusion and Brief Discussion

The variable selection is important as much as modeling. In the literature, GA has been proposed in different ways for variable selection in regression analysis [33-37]. But none of them is related to the solution of the

undersized sample problem, i.e. $n << p$. As mentioned before, if the number of the variable is equal to p, the number of subset of candidate model is $2^p - 1$. For example, if we have $p = 20$, the number of the subset of candidate models are equal to $2^{20} = 1048576$. Therefore it is difficult to determine which model is the best among these models without an effective search tool such as GA.

In this paper it was proposed a new andalternative approach to variable selection for undersized data by using HRM with HCEs and GA. We introduced the GA structure where $ICOMP_{Miss}$ used as a fitness function in order to select the best model from such a wide solution space. To demonsrate the effectiveness and utility of the proposed method, the Monte Carlo simulation study was performed on correlated and undersized data sets. The results were compared with Elastic Net (EN) modeling.

EN is a hybrid regression model between Ridge Regression (RR) and Least Absolute Shrinkage and Selection Operator (LASSO) [38]. It is claimed that it can be used for simultaneous modeling and variable selection in high dimensional or undersized data in [38]. However, in our simulation study ,we observed that EN modeling has poor performance to detect true variables when compared with our method. This has demonstrated the success of HCEs covariance matrices used in HRM analysis.

Nowadays, data sets can have radically increasing dimensions. In order to cope with these data sets, it is necessary dimension reduction or feature selection or to increase the number of samples. It may not always be possible to reach the number of samples required to be able to perform classical statistical methods. Therefore the methods that work well are required even if the sample size is too small. According to the results, it can be recommended that the proposed approach based on HRM and GA can be used in the analysis of these data sets.

Acknowledgments: I would like to thank to referees and the editor for their thoughtful and constructive suggestions that greatly improved the paper. I wish to thank to my supervisor Prof. Dr. Hamparsum Bozdogan from University of Tennessee since he introduced and guided to me in this interesting problem area and continued support.

Abbreviations: The following abbreviations are used throughout the text:

AIC	Akaike Information Criterion
BCSE	Bozdogan's Convex Sum Estimator
BIC	Bayesian Information Criterion
CAIC	Consistent Akaike Information Criterion
CN	Condition Number
CSE	Convex Sum Estimator
EB	Empirical Bayes Estimator
EN	Elastic Net
GA	Genetic Algorithm
HCEs	Hybrid Covariance Estimators
HRM	Hybrid Regression Model
HRM1	HRM model with MLE_STA_CSE covariance matrix
HRM2	HRM model with MLE_STA_BCSE covariance matrix
ICOMP	Information Complexity Criteria
ICOMPMiss	Information Complexity Criterion under misspecification
LASSO	Least Absolute Shrinkage and Selection Operator
LW	Ledoit-Wolf Estimator
MLE	Maximum Likelihood Estimator
MLE_STA_BCSE	Maximum Likelihood Stabilized Bozdogan's Convex Sum Estimator
MLE_STA_CSE	Maximum Likelihood Stabilized Convex Sum Estimator
OAS	Oracle Approximation Estimator
RR	Ridge Regression
SRE	Stipulated Ridge Estimator
SDE	Stipulated Diagonal Estimator
TPR	True Positive Rate

References

[1] Donoho, DL. Available online: statweb.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf. 2 (accessed on 20 April 2019)

[2] Cunningham P. Dimension Reduction. In: Machine Learning Techniques for Multimedia. Cord M, Cunningham P. Eds, Cognitive Technologies. Springer, Berlin, Heidelberg, 2008.

[3] Fiebig DG. On the maximum entropy approach to undersized samples. Appl Math Comput 1984; 14:301-312

[4] Pamukçu E, Bozdogan H, Çalık S. A novel hybrid dimension reduction technique for undersized high dimensional gene expression data sets using information complexity criterion for cancer classification. Comput Math Method M 2015; Article ID 370640: 1-14

[5] Bozdogan H, Pamukçu E. Novel Dimension Reduction Techniques for High-Dimensional Data Using Information Complexity. In Optimization Challenges in Complex, Networked and Risky Systems INFORMS 2016. p:140-170

[6] Mohebbi S, Pamukcu E, Bozdogan H. A new data adaptive elastic net predictive model using hybridized smoothed covariance estimators with information complexity. J Stat Comput Sim 2019: 89(6); 1060-1089.

[7] Pamukcu E. A new hybrid regression model for undersized sample problem. Celal Bayar University Journal of Science 2017; 13(3): 803-813

[8] Linhart H, Zucchini W. Finite sample selection criteria for multinomial models. Statistiche Hefte 1986; 27: 173-178

[9] Burnham KP, Anderson DR. Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife Res. 2001; 28: 111-119

[10] Boyce DE, Faire A, Weischedel R. Optimal subset selection: multiple regression, interdependence, and optimal network algorithms. Springer-Verlag, 1974. p:16.

[11] Bozdogan H. Intelligent Statistical Data Mining with Information Complexity and Genetic Algorithm. In: Statistical Data Mining and Knowledge Discovery. H. Bozdogan (ed). Chapman and Hall/CRC. Florida, 2004.

[12] Bozdogan H. Information Complexity and Multivariate Learning in High Dimensions with Applications in Data Mining. Forthcoming book. 2020

[13] Haff LR. Emprical bayes estimation of the multivariate normal covariance matrix. The Annals of Statistics. 1980; 8(3):586-597

[14] Shurygin A. The linear combination of the simplest discriminator and Fisher's one. In Applied Statistics. Nauka (ed). Moscow. Rusia. 1983.

[15] Press S. Estimation of a normal covariance matrix. Technical Report. University of British Columbia. 1975.

[16] Chen MCF. Estimation of covariance matrices under a quadratic loss function. Research Report S-46, Department of Mathematics, SUNY at Albany (Island of Capri, Italy). 1976; p:1–33.

[17] Bozdogan H. A new class of information complexity (ICOMP) criteria with an application to customer profiling and segmentation. Invited paper. In Istanbul University Journal of the School Business Administration 2010; 39(2): 370-398

[18] Chen Y, Wiesel A, Eldar YC, Hero AO. Shrinkage algorithms for mmse covariance estimation. IEEE Trans. On Signal Processing 2010; 58 (10): 5016–5029.

[19] Ledoit O, Wolf M. A well conditioned estimator for large dimensional covariance matrices. J Multivariate Anal 2004; 88: 365-411

[20] Thomaz CE. Maximum Entropy Covariance Estimate for Statistical Pattern Recognization. PhD Dissertation, Department of Computing Imperial College. University of London. UK, 2004.

[21] Pamukçu E. Choosing the optimal hybrid covariance estimators in adaptive elastic net regression models using information complexity. J Stat Comput Sim 2019; 89(16): 2983-2996.

[22] Akaike H. Information theory and extension of the maximum likelihood principle. 2nd International Symposium on Information Theory. Budapest: Academiai Kiado 1973; p:267-281,

[23] Akaike H. A new look at the statistical model identification. IEEE Transaction and Automatic Control 1974; AC-19:719-723

[24] Bozdogan H. ICOMP: A new model-selection criterion. In: 1. Conference of the international federation of classification societies 1987; p. 599-608.

[25] Bozdogan H. On the information based measure of covariance complexity and its application to the evaluation of multivariate linear models. Commun Stat Theor M 1990; 1: 221-278

[26] Bozdogan H, Haughton DMA. Information complexity criteria for regression models. Comput Stat Data An 1998; 28: 51-76

[27] Bozdogan H, Howe JA. Misspecified multivariate regression models using the genetic algorithm and information complexity as the fitness function. European Journal of Pure and Applied Mathematics 2012; 5(2), 211-249

[28] Schwarz G. Estimating the dimension of model. Ann Stat 1978; 6: 461-464

[29] Bozdogan H. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. Psychometrika 1987; 52(3): 345-370

[30] Goldberg DE. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA. 1989

[31] Michalewicz Z. Genetic algorithms+ data structures= evolution programs. Springer Science & Business Media, 2013.

[32] Jang JSR. Derivative-Free Optimization. In Neuro-Fuzzy and Soft Computing: A Computational Approach To Learning and Machine Intelligence. Prentice-Hall, USA. 1997; p: 173-196
[33] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc B 2005; 67(2): 301-320.
[34] Shahriari S, Faria S, Gonçalves AM. Variable selection methods in high-dimensional regression—A simulation study. Commun Stat-Simul C 2015; 44(10): 2548-2561.
[35] Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. J Chemometr 1992; 6(5): 267-281
[36] Chatterjee S, Laudato M, Lynch LA. Genetic algorithms and their statistical applications: an introduction. Comput Stat Data An 1996; 22(6): 633-651
[37] Minerva T, Paterlini S. Evolutionary approaches for statistical modelling. Published in: Proceedings of the 2002 Congress on Evolutionary Computation CEC'02. Honolulu, HI, USA 2002; Cat. No.02TH8600
[38] Tolvi J. Genetic algorithms for outlier detection and variable selection in linear regression models. Soft Comput 2004; 8(8); 527-533
[39] Paterlini S, Minerva T. Regression Model Selection Using Genetic Algorithms. Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing. Proceedings of the 11th WSEAS. 2010.