Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 9(10): 27607–27626.
DOI:10.3934/math.20241341
Received: 05 July 2024
Revised: 10 September 2024
Accepted: 18 September 2024
Published: 24 September 2024

Research article

Generalized Jaccard feature screening for ultra-high dimensional survival data

Renqing Liu[1], **Guangming Deng**[1,2,*] **and Hanji He**[3,*]

[1] School of Mathematics and Statistics, Guilin University of Technology, Guilin, 541004, China
[2] Key Laboratory of Applied Statistics, Guangxi Colleges and Universities,Guilin, 541004, China
[3] School of Economics and Finance, South China University of Technology, Guangzhou, 510006, China

* **Correspondence:** Email: dgm@glut.edu.cn, 202110189651@mail.scut.edu.cn.

Abstract: To identify critical genomes that influence a cancer patient's survival time, feature screening methods play a vital role in this biomedical field. Most of the current research relies on a fixed survival function model, which limits its universality in practical applications. In this paper, we propose the Generalized Jaccard coefficient (GJAC), which extends the traditional Jaccard coefficient from comparing binary vectors' similarity to calculating the correlation between the general vectors. The larger the GJAC value, the higher the sample similarity. Using the GJAC, we introduce a novel model-free screening method to select the active set of covariates in ultra-high dimensional survival data. Through Monte Carlo simulations, GJAC-Sure Independence Screening (GJAC-SIS) shows a higher accuracy, lower errors, and an excellent applicability in different types of survival data compared with other existing model-free feature screening methods in survival data. Additionally, in the real cancer datasets (DLBCL), GJAC-SIS can screen out two additional important genomes, which are certified in the real biomedical experiment, while the other five methods can't. As a result, GJAC-SIS achieves a high screening precision, delivers a more effective screening outcome, and has a better utility and universality.

Keywords: generalized Jaccard coefficient; ultra-high dimensional survival data; model-free
Mathematics Subject Classification: 62F07, 62N01

1. Introduction

With the exponential growth of data, current research has focused on finding truly relevant covariates in ultra-high dimensional datasets, which can make calculations easier in the future. In the current medical industry, using feature screening to identify cancer genomes is essential due to the quick advancements of science and technology. This calls for the quick identification of critical genomes, which have significant impacts on patients among thousands of potential genomes. In medical, some patients may occasionally not reach the final time of observation due to some unforeseen situations, thus leading to survival time censoring, commonly known as survival data.

For high dimensional data, based on penalty functions, a series of traditional variable screening methods were proposed, such as Least Absolute Shrinkage and Selection Operator (LASSO) [1], Smoothly Clipped Absolute Deviation (SCAD) [2], and Marginal Propensity to Consume (MPC) [3]. However, when the dimension reaches tens of thousands, it is upgraded to ultra-high dimensional data, which means the p-dimension of covariates $X_{n\times p}$ is larger than the sample size n, thereby increasing exponentially with n (satisfies the formula $\log(p) = O(n\alpha)$, $\alpha > 0$). At this time, the traditional regularization methods may lead to increased prediction errors of the model and a decreased accuracy. Ultra-high dimensional data is widely used in many domains such as biomedical image recognition, and natural language processing. Although ultra-high dimensional data can give us a better understanding of the subject, through the continuous growth of data dimensions, the distribution will become sparse, thus resulting in dimensional disasters and overfitting. A method to effectively screen out the really important covariates has become the key point in processing ultra-high dimensional data and conducting subsequent analyses. However, the extremely high complexity of ultra-high dimensional variables leads to a greater need for large amounts of storage, long processing times, and significant computing expenses. Therefore, to address this exigency, researchers have proposed many screening methods for ultra-high dimensional data to select a true and accurate set of active predictors.

Fan and Lv [4] initiated the Sure Independence Screening (SIS) method, thereby employing a linear regression model to calculate the marginal Pearson's correlation coefficients between the covariates and the response variables. The ultra-high dimensionality of the covariates was reduced to a modest dimension, thus effectively screening out the significant covariates that had important effects on the response variables. Based on the SIS, under the assumption of fixed models, many scholars had proposed different methods for screening ultra-high dimensional covariates. Bühlmann et al. [5] proposed a screening method based on the Partial Correlation coefficient (PC-simple algorithm) to control the effects of other variables, which was a more accurate correlation measurement between two variables. Hall and Miller [6] proposed a generalized correlation screening method; this was an extension based on Pearson's correlation coefficient in a more general context and was more universal. By transforming the form of the response variables, Li et al. [7] creatively proposed a Robust Rank Correlation (RRC) screening method, which enhanced the robustness when the data was in either a non-normal distribution or an abnormal state. In the framework of generalized linear models, Fan and Song [8] proposed the Maximum Marginal Likelihood Estimation screening method (MMLE-SIS). Additionally, under this assumption, the method of conditional screening (CSIS) was first proposed by Barut [9]. Under the premise that some covariates had significant influences on the response variables, the remaining important covariates can be accurately selected by calculating the given covariates and the maximum marginal likelihood estimation of the remaining single covariates. However, in practical problems, since the process of generating real data is often more complicated than theoretical models, the model is often uncertain. The inaccuracy of model assumptions may bring large biases to the screening results, which affects the reliability of the entire data analysis and prediction. Therefore, the model-free feature screening method has recently raised a lot of attention and is the main research direction for scholars. Zhu et al. [10] first proposed a model-free screening method (SIRS), which only imposed a very general model framework that no longer relied on a specific function model and established a consistency in rank(CIR) property for the method. They utilized statistical correlations among variables for the feature screening. Later, Li et al. [11] developed a screening procedure based on the distance correlation coefficient (DC-SIS), which was also applicable to grouped predictors and multivariate response variables. For categorical variables, Huang et al. [12] proposed a feature screening process based on Pearson's chi-square statistic and predicted the categorical response variables. Zhu et al. [13] introduced an interval quantile index to measure and test the independence for feature screenings. He et al. [14] proposed the quantile adaptive (Qa-SIS) screening method, which solved the heterogeneity problem of ultra-high dimensional data and applied it to ultra-high dimensional survival data, further broadening the application range of the model-free assumption screening method in practice.

As a unique type of dataset, survival data is characterized by the presence of either censored or incomplete response variables. The incompleteness of this data stems from a variety of reality factors, including study duration constraints, the loss of patients, etc., which results in parts of the data not being adequately documented and generating missing values. Thus, developing a method to deal with the censored data poses a significant challenge in real data analyses. Due to the insufficient consideration of censored variables, typical ultra-high dimensional feature screening approaches may not be appropriate or precise enough when dealing with survival data. It is necessary to propose new methods to screen out the set of important covariates from the ultra-high dimensional survival data.

For the studies of ultra-high dimensional survival data, Fan et al. [15] initially addressed this issue by proposing the Iterative Sure Independence Screening (ISIS) method, thereby integrating a loop iteration procedure with the SIS method under a fixed Cox proportional hazards model. This approach successively eliminated insignificant covariates to obtain the set of important covariates. Building upon this, Zhao et al. [16] suggested the Principled Sure Independence Screening (P-SIS) method, which was based on the marginal Cox proportional hazards model and incorporated a novel threshold selection method by controlling the false-positive rate to determine the number of retained covariates. Gorst [17] introduced the 'Feature Abbreviation at Survival Times' (FAST) index, which was based on the single-index hazard rate model by calculating the abbreviation of each covariate from the average of the time-varying covariate in terms of the time-varying covariate's survival time. Due to the problem that fixed models may lack robustness in some circumstances, scholars have tended to develop model-free feature screening methods. Song et al. [18] proposed the Censored Rank Independence Screening method (CRIS) based on an inverse probability-of-censoring weighted Kendall's τ index. This method was robust enough against outliers and enhanced the accuracy of the screening results. Zhang et al. [19] proposed the CR-SIS screening method based on the correlation rank by using the Kaplan-Meier estimator. Zhou et al. [20] proposed the cSIRS screening procedure, a sure independent ranking and screening procedure for censored regression. Zhong et al. [21] grounded in the censored mean-variance index (cMV) proposed the the censored mean-variance screening method (cMV-SIS).

In this paper, we propose an innovative screening method for ultra-high dimensional survival data, named the Generalized Jaccard coefficient screening (GJAC-SIS). The GJAC is a deformation of the Jaccard coefficient, which is designed to calculate the correlation between two vectors, and is a variant of the Pearson correlation coefficient. In terms of the screening accuracy and screening practicability, GJAC-SIS outperforms other model-free methods and does not require any model assumptions.

The main contributions of this paper are as follows:

(1) For ultra-high dimensional survival data, we propose a model-free feature screening procedure, which can be applied in all kinds of survival data with fewer restrictions. It demonstrates that GJAC-SIS offers a superior utility and universality.

(2) In simulation studies, GJAC-SIS exhibits a superior accuracy and reduced error rates in comparison to five existing model-free methods. This new method has advantages in ultra-high-dimensional survival data.

(3) In the real Diffuse large B-lymphocytoma cancer (DLBCL) dataset experiment, GJAC-SIS is capable of identifying genomes that have significant impacts on the survival time of the cancer patients. It can screen out two additional important genomes validated by biomedical experiments, which are not selected by the other five methods. It proves the practical value of this method in the actual survival data analysis. Moreover, it shows a better effectiveness, practicability, and predictability in the ultra-high dimensional feature screening procedure.

This paper is organized as follows: Section 2 describes the proposed GJAC-SIS methodology and the associated screening procedure; Section 3 proves the screening property of GJAC-SIS under specific conditions; Section 4 conducts Monte Carlo simulations to compare the screening accuracy in comparison to five screening methods; Section 5 provides a real cancer data (DLBCL) analysis; and Section 6 summarizes the advantages of this method and provides conclusions.

2. Screening method

2.1. Jaccard coefficient

The Jaccard coefficient is a widely recognized metric in computer science, ecology, genomics, and other fields. It is used to quantify the similarity and diversity between binary variables, which are characterized by two possible values: 0 and 1. Commonly known as a matching measure coefficient, it evaluates the degree of similarity or dissimilarity between two sets by examining their shared features. A higher Jaccard coefficient indicates a greater level of similarity between the two sets, whereas a lower coefficient reflects a higher degree of dissimilarity.

Mathematically, the Jaccard coefficient is formally defined as follows: Given two sets A and B, it is computed as the ratio of the size of their intersection (the elements common to both sets) to the size of their union (all unique elements across both sets).

It can be formally expressed as follows:

$$Jaccard = \frac{|A \cap B|}{|A \cup B|}, \tag{2.1}$$

where $|A \cap B|$ denotes the number of elements in the intersection of A and B, and $|A \cup B|$ represents the total number of unique elements in either A or B.

Suppose x and y are binary variables, $x = (x_1, x_1, \ldots\ldots, x_n)'$, $y = (y_1, y_2, \ldots\ldots, y_n)'$. Then, the number of two vector matches can be defined as follows:

$$a = \sum_{i=1}^{n} x_i y_i, \tag{2.2}$$

$$b = \sum_{i=1}^{n} y_i(1 - x_i), \tag{2.3}$$

$$c = \sum_{i=1}^{n} x_i(1 - y_i), \tag{2.4}$$

$$d = \sum_{i=1}^{n} (1 - x_i)(1 - y_i). \tag{2.5}$$

The Jaccard coefficient does not consider the condition that x and y are both 0, which is often used to deal with asymmetric binary vectors. In the formula, a is the number of matches when the two vectors x and y are both 1, b is the number of matches when x is 1 and y is 0, c is the number of matches when x is 0 and y is 1, and d is the number of matches when x and y are both 0.

Therefore the Jaccard coefficient can also be written as follows:

$$Jaccard = \frac{|A \cap B|}{|A \cup B|} = \frac{a}{a + b + c}. \tag{2.6}$$

Since the Jaccard coefficient is mainly used to calculate the similarity between individuals of symbolic or Boolean measures, its application is limited to determining whether two variables are identical or not. Consequently, there is a necessity to extend this measure to accommodate the similarity assessment of various types of vectors. To address this limitation, the GJAC is introduced.

2.2. Generalized Jaccad coefficient

When measuring the similarity between two vectors, we usually calculate the similarity coefficients between them, and the most common similarity coefficients are the Cosine Coefficient, the Pearson correlation coefficient, the Generalized Dice coefficient, and so on. The Generalized Jaccard coefficient, according to Zhang et al. [22], represents an extension of the Jaccard Coefficient applicable to general vectors and serves as a variant of the Pearson Correlation Coefficient. This coefficient is capable of dealing with diverse data types that the traditional Jaccard Coefficient cannot effectively manage. Since it is suitable to calculate the similarity of sets comprised of real-valued vectors, it can be applied to multiple domains to calculate the similarities, such as the text similarity, image similarity, audio similarity, and so on.

The GJAC is defined as follows:

For a pair of vector sets A and B,

$$GJaccard(A, B) = \frac{A \cdot B}{||A||^2 + ||B||^2 - A \cdot B}. \tag{2.7}$$

In the formula, and $A \cdot B$ denotes the dot product between the vectors, and $||A||^2$ denotes the square of the magnitude of vector A. The closer the GJAC is to 1, the higher the similarity between the two vectors. Based on the above definition, for continuous covariates $X = \left\{x_1, x_2, \ldots\ldots, x_p\right\}'$, and continuous response variables $Y = \{y_1, y_2, \ldots\ldots, y_n\}$, the estimated GJAC is defined as follows:

$$G\widehat{Jaccard}(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^{2} + \sum_{i=1}^{n} y_i^{2} - \sum_{i=1}^{n} x_i y_i}. \tag{2.8}$$

In this paper, we incorporate the GJAC into the procedure of feature screening for ultra-high dimensional survival data, referred to as the GJAC Sure Independence Screening method (GJAC-SIS).

Since the GJAC can quantify the similarity between any two sets, this calculation method does not rely on any specific model assumption. Therefore, the screening procedure can be classified as a model-free feature screening method.

2.3. Generalized Jaccard coefficient screening method

Let $\mathbf{X} = \left\{X_1, X_1, \ldots\ldots, X_p\right\}'$ be the covariates with p-dimension, T be the survival time, and C stand for the censoring time. Instead of observing the survival time, we define the observed time as $Y = min\{T, C\}$, which means that by taking the minimum of the survival time T and the censoring time C, $\delta = I(T \leq C)$ denotes the censoring index. $I(\cdot)$ represent the indicator function when $T \leq C$, $\delta = 1$. Suppose a condition that by knowing the value of $\mathbf{X}$, both T and C are independent, and the observing data sets $\{(x_i, y_i, \delta_i) : x_i \in \mathbb{R}^p, y_i \in \mathbb{R}^+, \delta_i \in 0, 1, i = 0, 1, 2, \ldots, n\}$ are identically independent distributed samples of $(\mathbf{X}, Y, \Delta)$. Let $S(t|\mathbf{X}) = P(T > t|\mathbf{X})$ be the conditional function of survival time T, when given the value of $\mathbf{X}$. Through the screening procedure, we need to screen out the set of active predictors that will impact the survival time T.

We can define the set of active predictors by the following:

$$\mathcal{A} = \left\{j : S(t|X_j)\ functionally\ depends\ on\ X_j,\ for\ some\ j = 1, 2, \ldots\ldots, p\right\}.$$

Therefore, the inactive set is defined as follows:

$$\mathcal{I} = \left\{j : S(t|X_j)\ does\ not\ depend\ on\ X_j,\ for\ j = 1, 2, \cdot, p\right\} \setminus \mathcal{A}.$$

If j belongs to $\mathcal{A}$, then X_j is an active predictor, which influences the survival time; otherwise, it is an inactive predictor.

We define the screening procedure of GJAC-SIS as follows: we compute the GJAC of every row between $X_j = (x_{1j}, x_{2j}, \ldots\ldots, x_{nj})$ and Y, that is

$$w_j = GJaccard(X_j, Y), j = 1, 2, \ldots\ldots, p. \tag{2.9}$$

The larger the w_j, the more similarities of X_j and Y. As for the jth observation of the sample, we construct a screening estimator $\widehat{w_j}$ to calculate the similarity between the observed covariate set $\left\{X_j, j = 1, 2, \ldots\ldots, p\right\}$ and actual observation time Y:

$$\widehat{w_j} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_{ij}{}^2 + \sum_{i=1}^n y_i{}^2 - \sum_{i=1}^n x_i y_i}. \tag{2.10}$$

In practice, we rank the covariates with large $\widehat{w_j}$‘s. Therefore, for GJAC-SIS, we define the estimated set of the active predictors set $\mathcal{A}$ as follows:

$$\widehat{\mathcal{A}} = \left\{1 \leq j \leq n : \widehat{w_j} \geq d_0, d_0 < n\right\}.$$

d_0 stands for the threshold of screening, which is the maximum number of important covariates that need to be selected and will be given in advance. As suggested by Fan and Lv [4], d_0 is always taken as $[n/\log(n)]$ or $n - 1$. In this paper, we only chose $d_0 = [n/\log(n)]$ as the screening threshold, which is the size of $\widehat{\mathcal{A}}$, i.e., $\left|\widehat{\mathcal{A}}\right| = d_0 = [n/\log(n)]$, where $[a]$ denotes the integer part of a.

3. Screening property

According to the definitions and properties proved by Fan and Lv [4], GJAC-SIS, also has a Sure Screening Property Property.

As the sample size n increases towards infinity, the probability that the true active predictors in set $\mathcal{A}$ will be included in the estimated set $\widehat{\mathcal{A}_*}$ converges to 1. This implies that, given a large enough sample size of n, the GJAC-SIS screening procedure will eventually identify and select all the active predictors.

This is defined as follows:

$$\widehat{\mathcal{A}_*} = \left\{ j : \widehat{w_j} \geq cn^{-\kappa}, for\ 1 \leq j \leq p \right\}.$$

Condition 1. *There exist positive constants c and κ, for $c > 0$, $0 \leq \kappa < 1/2$, such that* $\min_{j \in \mathcal{A}} w_j \geq 2cn^{-\kappa}$.

Proposition 1. *When X is a continuous variable, we have $0 < GJ(X_j, Y) \leq 1$, and X_j and Y are identical if and only if $GJ(X_j, Y) = 1$.*

Theorem 1. *For a continuous response Y and continuous covariates X, under Condition 1, for any* $0 < \epsilon < 1$, *we have*

$$P(\left|\widehat{w_j} - w_j\right| > 2\epsilon) \leq \exp\left\{ -\frac{\epsilon^2}{2n(\sum_{v=1}^{3} c_v {\sigma_v}^2 + c_4 \epsilon)} \right\}, \tag{3.1}$$

where c_1, c_2, c_3, c_4 is constant.

Theorem 2. *For a continuous response Y and continuous covariates X, under Condition 1, we have*

$$P(\mathcal{A} \subseteq \widehat{\mathcal{A}_*}) \geq 1 - O\left(\exp\left\{ -\frac{c^2 n^{-\kappa}}{c_6} \right\} \right), \tag{3.2}$$

where c_6 is a constant.

According to Fan and Lv [4], and Zhu et al. [10], Condition 1 assumes the minimum true signal to disappear to zero in the order of $n^{-\kappa}$ as the sample size goes to infinity.

GJAC-SIS has a Sure Screening Property. The detailed proof of the theoretical properties is in the Appendix.

4. Simulation

In this section, we generate different types of survival data by setting the appropriate parameter conditions. We conduct Monte Carlo simulations to illustrate the finite sample performance. In order to compare the accuracy and effectiveness of the new screening method GJAC-SIS, we consider five existing screening methods: (1) Two methods that are based on fixed survival models: P-SIS proposed by Zhao et al. [16], which was based on a fixed Cox's proportion hazards model, and FAST proposed by Gorst [17], which was based on a single-index hazard rate model; and (2) Three model-free screening methods: CRIS proposed by Song et al. [18], CR-SIS proposed by Zhang et al. [19], and CSIRS proposed by Zhou et al. [20]. The above six methods are applied concurrently to screen out the important predictors.

First, to scientifically assess the screening effect of the new method, we introduce three core evaluation criteria, suggested by Li et al. [11] as follows. The accuracy and effectiveness of the six model-free screening methods are compared by evaluating the performance of each criterion.

$\mathbf{P_j}$: The proportion of times a single active predictor X_j is included, thus representing the probability that an active predictor is correctly selected.

$\mathbf{P_{all}}$: The proportion of times all the active predictors $X_j s$ are included, thus signifying the probability that all active predictors are correctly selected.

MMS: The minimum model size to include all the active predictors. This indicator can measure the effectiveness of each screening procedure.

We set the number of active predictors through the formula $d = [n/\log(n)]$, where n stands for the sample size, and $[x]$ stands for the integer part of x.

In the simulation with 1000 replicates, the criterion P_1 indicates that when given a model size d, the average proportion of the first important covariate is screened out. Under an idealistic condition, after 1000 repetitions, the closer the variance is to 0, the smaller the screening error. The criterion P_{all} reveals that when given a model size d, the average proportion of all significant covariates are screened out. The closer the variances of 1000 repetitions are to 0, the better. It indicates that our screening method consistently maintains a high identification rate. The significance of the MMS criterion is the number of minimum significant covariates included in the corresponding quantiles at 5%, 25%, 50%, 75%, and 95%. At each quantile, a closer match between the number of screened-out active predictors and the true number of important covariates shows a better screening effect. It is used to measure the model complexity and represent the better screening effect.

In all the simulations, we set the covariate matrix $X_{n\times p}$ with a sample size n=100, coupled with a dimensionality p=1000; then, the number of active predictors is calculated by $d = [n/\log(n)] = 21$. According to the settings, we need to use the above six aforementioned screening methods to screen out the active covariates set, with 1000 repetitions. The screening performances of three criteria are compared to test whether GJAC-SIS is superior to other screening methods.

Simulation 1 (Cox Proportional Hazards Model)**.** The survival time T is generated by the conditional proportional hazards model:

$$h(t|X) = h_0(t)\exp\left(X_i^T\beta\right). \tag{4.1}$$

In this model, $h_0(t)$ denotes the baseline of the hazard function, which is set to $h_0(t) = -\log(u_i)$, where $u_i \sim U[0,1]$. β is the regression coefficient of p-dimension covariates. In this simulation, we set $\beta = (0.8, 0.8, 0.8, 0, \ldots, 0)^T \in \mathbb{R}^p$, thereby designating the first three covariates $\{X_1, X_2, X_3\}$ as the active predictors. The matrix $X_{n\times p}$ is generated from a multivariate normal distribution with a mean of $\mathbf{0}$, and a correlation matrix $\sum = \left(0.8^{|i-j|},\ i \neq j\right)$, $(i, j = 1, 2, 3)$. The censoring time C_i is independently generated by the uniform distribution $U \sim [0, c]$. c is selected to implement different censoring rates, which is used to control the amount of censoring time. Here, we consider the censoring rates(CR) of 30% and 60% to compare the impacts of different screening methods on the screening outcomes.

Table 1 presents the comparative screening performances of GJAC-SIS with the P-SIS, CR-SIS, CRIS, CSIRS, and FAST-SIS methods.

Under the fixed Cox proportional hazards model with a 30% censoring rate, GJAC-SIS and FAST-SIS achieve the same results in terms of the MMS criterion, thus encompassing all active predictors. However, following 1000 repetitions, the variance associated with the GJAC-SIS screening method was observed to be lower than that of FAST-SIS, thus indicating that GJAC-SIS exhibits reduced error rates. When the censoring rate increases to 60%, the three screening methods GJAC-SIS, P-SIS, and CSIRS reach the same results, which means they have equivalent screening capabilities.

Table 1. When the censoring rate =30% and censoring rate=60%. The performance of six methods in Simulation 1.

	P_s		MMS				
Method	P_1	P_{all}	5%	25%	50%	75%	100%
			CR=30%				
GJAC-SIS	1(0)	1(0)	1.1	1.5	2	2.5	2.9
P-SIS	0.67(0.03)	0(0.03)	1.4	3	5	93.5	164.3
CR-SIS	1(0.01)	1(0)	1.1	1.5	2	3.5	4.7
CRIS	0(0.04)	0(0.05)	614.5	661.5	720.0	853.5	960.3
CSIRS	0.67(0.02)	0(0.05)	1.1	1.5	2.0	16.0	27.2
FAST-SIS	1(0.008)	1(0.003)	1.1	1.5	2	2.5	2.9
			CR=60%				
GJAC-SIS	1(0)	1(0)	1.1	1.5	2	2.5	2.9
P-SIS	1(0)	1(0)	1.1	1.5	2	2.5	2.9
CR-SIS	0.03(0.03)	0(0.1)	231.5	337.25	415.5	572	785.9
CRIS	0.13(0.1)	0.1(0)	538	592.75	654	766.75	856.95
CSIRS	1(0)	1(0)	1.1	1.5	2	2.5	2.9
FAST-SIS	1(0)	1(0)	97.35	140.75	205.5	264.5	302.3

Simulation 2 (Linear Transformation Model)**.** Generating the survival time T from the following transformation model:

$$H(T_i) = -\beta' X_i + \epsilon_i. \tag{4.2}$$

According to simulation experiments' parameter settings by Song et al. [18] and Zhang et al. [19], we set $\beta = (-1, -0.9, -0.5, -0.8, 0, ..., 0)^T$, $H(t) = \log\left\{0.5(e^{2t} - 1)\right\}$. The covariate X is generated by a multivariate normal distribution with a mean of $\mathbf{0}$ and a correlation matrix $\sum = \left(0.5^{|i-j|}, i \neq j\right)$, $(i, j = 1, 2, 3, 4)$, which means that only the first four covariates $\{X_1, X_2, X_3, X_4\}$ were chosen as the given active predictors. However, the other covariates are identically independent and generated by a multivariate standard normal distribution. The random error ϵ_i is generated from a standard normal distribution, and the censoring time C_i is independently generated from the uniform distribution $U \sim [0, c]$, where c is selected to implement different censoring rates. In this simulation,we chose either a 30% or 60% censoring rate to compare the effects of different screening methods on the screening results.

The performances of the screening comparison of GJAC-SIS with the P-SIS, CR-SIS, CRIS, CSIRS, and FAST-SIS methods are shown in Table 2.

At a censoring rate of 30%, the GJAC-SIS, P-SIS, and CSIRS methods produce identical outcomes across all evaluated criteria. When the censoring rate reaches 60%, GJAC-SIS, P-SIS, and FAST-SIS yield the same results on the MMS criterion, which encompasses all significant covariates. Furthermore, these three screening methods are capable of identifying four significant covariates that closely approximate the true model size at the 95% quantile. However, according to GJAC-SIS screening results, the variance is 0, which means a higher screening accuracy in comparison with others.

Table 2. When the censoring rate =30% and censoring rate=60%. The performance of six methods in Simulation 2.

	P_s		MMS				
Method	P_1	P_{all}	5%	25%	50%	75%	100%
			CR=30%				
GJAC-SIS	1(0)	1(0)	1.15	1.75	2.5	3.25	3.85
P-SIS	1(0)	1(0)	1.15	1.75	2.5	3.25	3.85
CR-SIS	0(0)	0(0)	171	363	516	675	771
CRIS	0.95(0.05)	0.8(0.2)	1.15	1.75	2.5	3.25	3.85
CSIRS	1(0)	1(0)	1.15	1.75	2.5	3.25	3.85
FAST-SIS	0.8(0.245)	0.4(1.12)	1.15	1.75	2.5	3.75	4.75
			CR=60%				
GJAC-SIS	1(0)	1(0)	1.15	1.75	2.5	3.25	3.85
P-SIS	0.95(0.05)	0.8(0.2)	1.15	1.75	2.5	3.25	3.85
CR-SIS	0(0)	0(0)	268	318	390.5	496.5	791.3
CRIS	0.65(0.187)	0.4(0.245)	2.15	2.75	12.5	37.25	53.1
CSIRS	0.9(0.061)	0.6(0.245)	1.15	1.75	2.5	4.5	8.1
FAST-SIS	1(0.2)	1(1.12)	1.15	1.75	2.5	3.25	3.85

Simulation 3 (Accelerated Failure Time Model)**.** The survival time T_i is generated from an AFT model.

$$\log(T_i) = c_0([X_i]^T\beta) + \epsilon_i. \tag{4.3}$$

Here, we set $\beta = (0.8, 0.8, 0.8, 0, \ldots, 0)^T \in \mathbb{R}^p$, which denotes the p-dimension regression coefficients and only chose the first three covarites $\{X_1, X_2, X_3\}$ as the given active predictors, which are generated by a multivariate normal distribution with a mean of $\mathbf{0}$, and a correlation matrix $\sum = \left([0.8]^{|i-j|}, i \neq j\right)$, $(i, j = 1, 2, 3)$. Additionally, we set the constant $c_0 = 1$ and the random error $\epsilon_i \sim N(0, 1)$. The censoring time C_i is independently generated by the uniform distribution $U \sim [0, c]$ and c is selected to implement different censoring rates. In this simulation,we chose either a 30% or 60% censoring rate to compare the effects of different screening methods on the screening results.

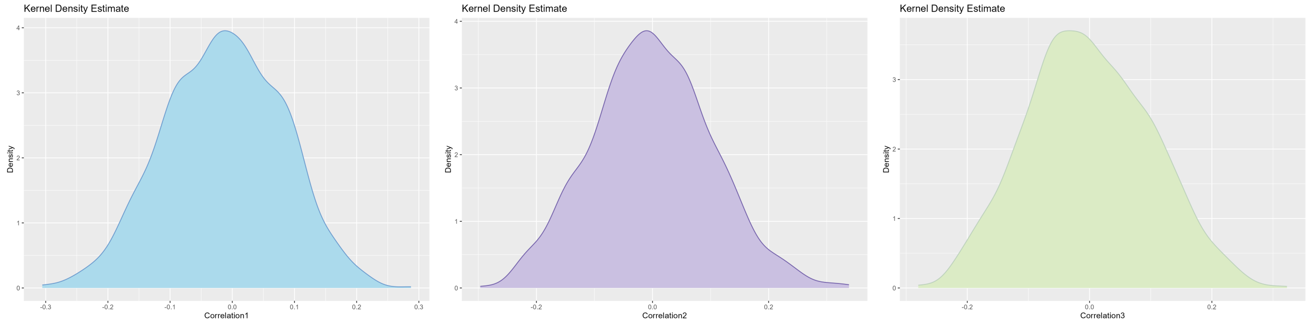
The screening comparison results of GJAC-SIS with the P-SIS, CR-SIS, CRIS, CSIRS, and FAST-SIS methods are presented in Table 3.

At the censoring rate of 30%, all four methods consistently indentify the first and all significant covariates, across 1000 replicated experiments with a zero variance. However, in relation to the MMS criterion, all methods overestimate the true number of active predictors in the 95% quantile, with the exception of the GJAC-SIS screening method, which closely approximates the actual number of the three predictors. When the censoring rate is 60%, the CR-SIS screening method achieves the same screening effect as the GJAC-SIS screening method.

Table 3. When the censoring rate =30% and censoring rate=60%. The performance of six methods in Simulation 3.

	P_s		MMS				
Method	P_1	P_{all}	5%	25%	50%	75%	100%
			CR=30%				
GJAC-SIS	1(0)	1(0)	1.1	1.5	2	2.5	2.9
P-SIS	0.97(0.03)	0.9(0.02)	1.1	1.5	2	4.75	6.95
CR-SIS	1(0)	1(0)	1.1	1.5	2	3	3.8
CRIS	0.53(0.11)	0.2(0.17)	3.75	6.75	10.5	89	114.25
CSIRS	1(0)	1(0)	1.1	1.5	2	3.25	4.25
FAST-SIS	1(0)	1(0)	1.1	1.5	2	3.5	4.7
			CR=60%				
GJAC-SIS	1(0)	1(0)	1.1	1.5	2.0	2.5	2.9
P-SIS	1(0)	1(0)	1.15	1.75	2.5	3.5	3.9
CR-SIS	1(0)	1(0)	1.1	1.5	2.0	2.5	2.9
CRIS	0.57(0.07)	1(1.12)	2.2	5.25	9.5	115.25	168.15
CSIRS	1(1.12)	1(1.12)	1.1	1.5	2.0	3.0	3.8
FAST-SIS	0.97(1.31)	0.9(1.12)	1.1	1.5	2.0	3.0	3.8

Considering that there may be a strong correlation between the censoring time and the covariates, which may lead to inaccurate screening outcomes, we calculate the correlation coefficients between the censoring time and the covariates generated by each survival function model. Additionally, we also draw the Kernal Density plots to provide a more intuitive visualization of the distribution of these correlation coefficients. As illustrated in Figure 1, the majority of the correlation coefficients are concentrated within the interval of $[-0.2, 0.2]$, with the peaks are close to 0.



(a) Simulation 1.

(b) Simulation 2.

(c) Simulation 3.

Figure 1. The Kernel Density plots of correlation coefficients.

The figures suggest no obvious correlation between the variables. Consequently, we can draw the following experimental conclusions.

After the 1000 replicates for Monte Carlo simulation experiments, and by comparing the screening results of the GJAC-SIS method with the other five methods, combined with three evaluation criteria, we find that the GJAC-SIS method exhibits a superior overall performance. In terms of the screening accuracy, all the active predictors are accurately selected across all models in each simulation, with a variance equal to zero, thus signifying a minimal error. This outcomes reflects that GJAC-SIS has the highest accuracy and lower error rates. Regarding the screening efficiency, under the MMS criterion, GJAC-SIS successfully selects the corresponding number of covariates for each quantile, thereby effectively selecting all the active predictors specified at the 95% quantile.

5. Real data analysis

For this section, we chose the Diffuse Large B-cell Lymphoma (DLBCL) cancer patients dataset, which is accessible through the National Institutes of Health's affiliated webpage at `http://llmmpp.nih.gov/MCL/`. The DLBCL dataset was collected by Rosenwald's [23] team and had been instrumental in conducting extensive molecular diagnostics, exploring pathological mechanisms, and predicting patient survival, thus significantly contributing to the treatment of DLBCL cancers. In this part of the study, the GJAC-SIS is applied to the DLBCL dataset to accurately identify important genomes that have significant impacts on the survival time of the cancer patients. Statistically, the dataset contains gene expression profiling data from 92 patients diagnosed with cellular lymphoma, encomposing a total of 8810 unique genomic features. The survival times for these patients spanned a significant and wide range from 0.02 to 14.05 years, with a median of 1.96 years, an average of 2.76 years, and a variance of 2.78. Among the 92 patients, 28 patients were alive at the end of the study, meaning that their survival time was right censored, thus amounting to a censoring rate of approximately 30.4%. The remaining 64 patients unfortunately passed away during the study period, thus providing a complete survival time for this study. Let the response variable Y be the survival time in the data. We apply the six screening methods GJAC-SIS, P-SIS, CR-SIS, CRIS, CSIRS, and FAST-SIS to screen out important genomes that affect the patient's survival time. According to the selection threshold principle, it needs to screen out $d = [92/\log(92)] = 20$ genomes. Table 4 shows the summarized first 20 genomes' unique identifications (UNIQIDS).

Among them, six gene identifications "17326", "25234", "30157", "31420", "34771", and "34790" are selected as important genes that affect the patient's death from cellular lymphomas. In medicine, these identifications represent "Kinesin family member 23", "Antigen identified by monoclonal antibody Ki-67", "Centromere protein F", "Aurora kinase B", "Tubulin, alpha, ubiquitous", and "Thymidine kinase 1, soluble".

In Rosenwald's study, it was clearly confirmed that there were two gene identifications—"30142" and "16129"—that were very important in affecting the death of cancer patients. The newly proposed screening GJAC-SIS method successfully identifies these two genomes, ranking them at the top of the important variables set. In contrast, none of the other screening methods succeed in selecting these two genomes in the set of the top 20 important genomes.

These six genomes were simultaneously screened from the six methods as dependent variables for the regression prediction. We apply the Cox proportional hazards survival function model to predict survival time, and plot a survival probability curve, as shown in Figure 2. It is evident that the survival time stops around 14 years, which is in line with the patients' survival time in the real dataset.

Additionally, we apply the Cox survival regression model to the 20 genomes identified by the GJAC-SIS method to evaluate their predictive significance, with the results illustrated in Figure 3.

The C-index reaches 0.8, thus indicating that the top 20 genomes selected by the GJAC-SIS exhibit a strong predictive capability. Notably, the four gene identifications—"26950", "27678","30142", and "30157"—are the most critical genomes with a high significance. Moreover, this finding is also in line with the Rosenwald's experiment, further confirming the high accuracy and comprehensiveness of the screening method introduced in this paper.

Table 4. The top 20 gene identifications are screened by six methods.

GJAC-SIS	P-SIS	CR-SIS	CRIS	CSIRS	FAST-SIS
16312	27095	**30157**	28872	27095	28990
28990	**30157**	28346	**17326**	32187	**34790**
34790	**25234**	27762	28990	**34790**	**25234**
30142	32187	15936	17370	**30157**	16312
25234	**34790**	24723	**34790**	**25234**	**31420**
31420	28346	17198	**34771**	**31420**	27095
27762	24794	27116	**31420**	24794	27678
30157	**34771**	16312	27049	24723	24794
28872	**31420**	**34771**	**25234**	28346	**30157**
24794	16528	**34790**	16528	32699	28872
27095	**17326**	27095	32699	**34771**	26950
27678	28872	**31420**	**30157**	28990	32699
16129	28990	**25234**	30282	32049	17123
30917	32699	24610	27095	**17326**	**34771**
34201	17343	**17326**	32187	30282	27762
26950	27049	17434	33549	28872	17343
28726	34687	24656	24710	26962	**17326**
17326	26950	30917	24404	27019	17685
17123	24723	17174	17176	16528	24723
34771	24610	29330	24794	24610	28978

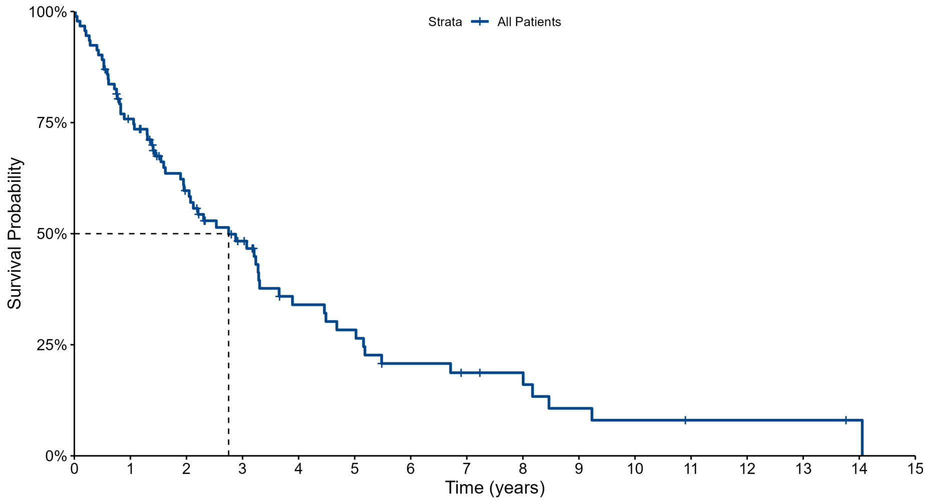


Figure 2. The survival probability curve.

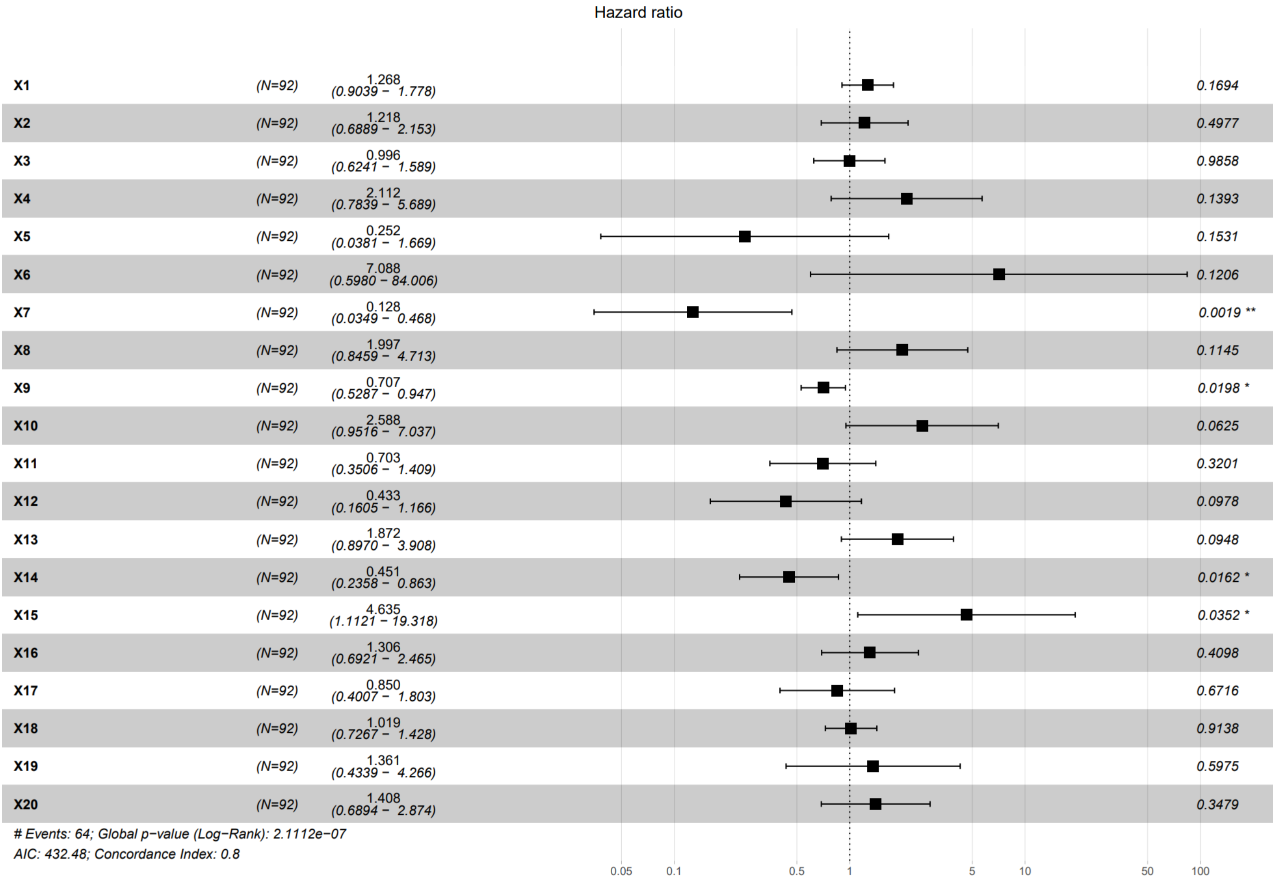


Figure 3. The significance of the top 20 genomes.

6. Conclusions

Based on the Jaccard coefficient, this paper proposed the GJAC to calculate the similarity between general vectors, introduced a new ultra-high dimensional model-free feature screening method, and applied this new method to the ultra-high dimensional survival data. Through repeated numerical simulations, this method proved to be both highly accurate and practical. Whether dealing with a fixed Cox Proportional Hazards Regression Model, a nonlinear model, or an Accelerated Failure Time model involving censored data, the GJAC-SIS screening method demonstrates a higher accuracy in screening out all active predictors and maintains relatively lower error rates compared to other methods.

In the real data experiment using the DLBCL cancer dataset, the GJAC-SIS method effectively pinpointed the top 20 gene identifications that have significant impacts on the cancer patients' survival time. To further verify the importance of these genes, we employ the six genes that are simultaneously by six different screening methods for survival analysis to predict the survival time. By plotting the survival curve, we confirmed that the GJAC-SIS screening method has a better predictive performance.

Drawing on the outcomes of this empirical study, the GJAC-SIS method is capable of identifying genomes that are closely associated with the survival time of cancer patients and can also screen out two additionally important genomes that were not selected by other methods. These screened genomes exhibited a high statistical significance, thus reinforcing the enhanced accuracy and applicability of this model-free screening method. As a result, the GJAC-SIS method holds a considerable practicle value in the application of biomedical research.

Author contributions

Renqing Liu: Conceptualization, Methodology, Software, Formal analysis, Writing-Original Draft; Guangming Deng: Supervision; Hanji He: Writing-Review & Editing.

Conflict of interest

The author certifies that the publication of this paper does not involve any conflicts of interest.

References

1. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Roy. Stat. Soc. B*, **58** (1996), 267–288. http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x
2. J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Stat. Assoc.*, **96** (2001), 1348–1360. http://dx.doi.org/10.1198/016214501753382273
3. C. Zhang, Nearly unbiased variable selection under minimax concave penalty, *Ann. Statist.*, **38** (2010), 894–942. http://dx.doi.org/10.1214/09-AOS729
4. J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, *J. Roy. Stat. Soc. B*, **70** (2008), 849–911. http://dx.doi.org/10.1111/j.1467-9868.2008.00674.x
5. P. Bühlmann, M. Kalisch, M. Maathuis, Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm, *Biometrika*, **97** (2010), 261–278. http://dx.doi.org/10.1093/biomet/asq008
6. P. Hall, H. Miller, Using generalized correlation to effect variable selection in very high dimensional problems, *J. Comput. Graph. Stat.*, **18** (2009), 533–550. http://dx.doi.org/10.1198/jcgs.2009.08041
7. G. Li, H. Peng, J. Zhang, L. Zhu, Robust rank correlation based screening, *Ann. Statist.*, **40** (2012), 1846–1877. http://dx.doi.org/10.1214/12-AOS1024
8. J. Fan, R. Song, Sure independence screening in generalized linear models with NP-dimensionality, *Ann. Statist.*, **38** (2010), 3567–3604. http://dx.doi.org/10.1214/10-AOS798
9. E. Barut, J. Fan, A. Verhasselt, Conditional sure independence screening, *J. Amer. Stat. Assoc.*, **111** (2016), 1266–1277. http://dx.doi.org/10.1080/01621459.2015.1092974
10. L. Zhu, L. Li, R. Li, L. Zhu, Model-free feature screening for ultrahigh-dimensional data, *J. Amer. Statist. Assoc.*, **106** (2011), 1464–1475. http://dx.doi.org/10.1198/jasa.2011.tm10563
11. R. Li, W. Zhu, L. Zhu, Feature screening via distance correlation learning, *J. Amer. Stat. Assoc.*, **107** (2012), 1129–1139. http://dx.doi.org/10.1080/01621459.2012.695654
12. D. Huang, R. Li, H. Wang, Feature screening for ultrahigh dimensional categorical data with applications, *J. Bus. Econ. Stat.*, **32** (2014), 237–244. http://dx.doi.org/10.1080/07350015.2013.863158

13. L. Zhu, Y. Zhang, K. Xu, Measuring and testing for interval quantile dependence, *Ann. Statist.*, **46** (2018), 2683–2710. http://dx.doi.org/10.1214/17-AOS1635

14. X. He, L. Wang, H. Hong, Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, *Ann. Statist.*, **41** (2013), 342–369. http://dx.doi.org/10.1214/13-AOS1087

15. J. Fan, Y. Feng, Y. Wu, High-dimensional variable selection for Cox's proportional hazards model, In: *Borrowing strength: theory powering applications—a festschrift for Lawrence D. Brown*, Durham: Institute of Mathematical Statistics, 2010, 70–86. http://dx.doi.org/10.1214/10-IMSCOLL606

16. S. Zhao, Y. Li, Principled sure independence screening for Cox models with ultra-high-dimensional covariates, *J. Multivariate Anal.*, **105** (2012), 397–411. http://dx.doi.org/10.1016/j.jmva.2011.08.002

17. A. Gorst-Rasmussen, T. Scheike, Independent screening for single-index hazard rate models with ultrahigh dimensional features, *J. Roy. Stat. Soc. B*, **75** (2013), 217–245. http://dx.doi.org/10.1111/j.1467-9868.2012.01039.x

18. R. Song, W. Lu, S. Ma, X. Jessie Jeng, Censored rank independence screening for high-dimensional survival data, *Biometrika*, **101** (2014), 799–814. http://dx.doi.org/10.1093/biomet/asu047

19. J. Zhang, Y. Liu, Y. Wu, Correlation rank screening for ultrahigh-dimensional survival data, *Comput. Stat. Data Anal.*, **108** (2017), 121–132. http://dx.doi.org/10.1016/j.csda.2016.11.005

20. T. Zhou, L. Zhu, Model-free feature screening for ultrahigh dimensional censored regression, *Stat. Comput.*, **27** (2017), 947–961. http://dx.doi.org/10.1007/s11222-016-9664-z

21. W. Zhong, J. Wang, X. Chen, Censored mean variance sure independence screening for ultrahigh dimensional survival data, *Comput. Stat. Data Anal.*, **159** (2021), 107206. http://dx.doi.org/10.1016/j.csda.2021.107206

22. D. Zhang, X. You, S. Liu, K. Yang, Multi-colony ant colony optimization based on generalized Jaccard similarity recommendation strategy, *IEEE Access*, **7** (2019), 157303–157317. http://dx.doi.org/10.1109/ACCESS.2019.2949860

23. A. Rosenwald, G. Wright, A. Wiestner, W. Chan, J. Connors, E. Campo, et al., The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell Lymphoma, *Cancer Cell*, **3** (2003), 185–197. http://dx.doi.org/10.1016/S1535-6108(03)00028-X

Appendix. Proofs

Proof of Proposition 1. According to the definitions of $GJaccad(X_j, Y)$, suppose A stands for the set of X_j, B stands for the set of Y, we have

$$GJ(X_j, Y) = \frac{AB}{||A||^2 + ||B||^2 - AB} = \frac{AB}{||A||^2 + ||B||^2 - 2AB + AB} = \frac{AB}{(A-B)^2 + AB}.$$

Because of $||A|| > 0$ and $||B|| > 0$, then

$$(A-B)^2 \geq 0, AB > 0, GJ(X_j, Y) > 0.$$

If $A = B$, then $(A - B)^2 = 0$, we have

$$GJ(X_j, Y) = \frac{AB}{(A - B)^2 + AB} = \frac{AB}{AB} = 1.$$

If $A \neq B$, then $(A - B)^2 > 0$, we have

$$GJ(X_j, Y) = \frac{AB}{(A - B)^2 + AB} = \frac{1}{(A - B)^2/AB + 1} < 1,$$

and $GJ(X_j, Y) \to 0$, when $|A - B|$ approaches infinity.
If $GJ(X_j, Y) = 1$, we have

$$GJ(X_j, Y) = \frac{AB}{||A||^2 + ||B||^2 - AB} = \frac{AB}{(A - B)^2 + AB} = 1.$$

Because $(A - B)^2 \geq 0,\ AB > 0$,we have

$$AB = (A - B)^2 + AB,$$

then $A = B$.

Therefore, we can get the conclusion that, $0 < GJ(X_j, Y) \leq 1$, X_j and Y are the same if and only if $GJ(X_j, Y) = 1$. □

Lemma 1 (Bernstein Inequality). *If $Z_1, Z_2, ..., Z_n$ is an independent random variable with a mean value of 0 and bounded supporter is $[-M, M]$, then the inequality*

$$P\left(\left|\sum_{i=1}^{n} Z_i\right| > t\right) \leq 2\exp\left\{-\frac{t^2}{2(v + Mt/3)}\right\},$$

where $v \geq Var(\sum_{i=1}^{n} Z_i)$.

Lemma 2. *For continuous response Y and continuous covariates X, we have the following three inequalities:*
(a) $P\left(\left|\sum_{i=1}^{n} x_{ij}^2 - ||A||^2\right| > t\right) \leq 2\exp\left\{-\frac{t^2}{2n(\sigma_1^2 + t/3)}\right\}$,
(b) $P\left(\left|\sum_{i=1}^{n} y_i^2 - ||B||^2\right| > t\right) \leq 2\exp\left\{-\frac{t^2}{2n(\sigma_2^2 + t/3)}\right\}$,
(c) $P\left(\left|\sum_{i=1}^{n} x_{ij} y_i^2 - ||AB||^2\right| > t\right) \leq 2\exp\left\{-\frac{t^2}{2n(\sigma_3^2 + t/3)}\right\}$,
where $\sigma_1^2 = Var(x_{ij}^2), \sigma_2^2 = Var(y_i^2), \sigma_3^2 = Var(x_{ij} y_i)$.

Proof of Lemma 2. Because

$$\sum_{i=1}^{n} x_{ij}^2 - ||A||^2 = \sum_{i=1}^{n} (x_{ij}^2 - \frac{1}{n}||A||^2).$$

Let $Z_i = x_{ij}^2 - \frac{1}{n}||A||^2 = x_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n} A_i^2 = x_{ij}^2 - E(A^2)$, we have

$$E(Z_i) = E(x_{ij}^2 - E(A^2)) = E(x_{ij}^2) - E(A^2) = 0,$$

$$Var(\sum_{i=1}^{n} Z_i) = nVar(Z_i) = nVar(x_{ij}^2 - E(A^2)) = nVar(x_{ij}^2) = n\sigma_1^2,$$

then we have by Lemma 1

$$P\left(\left|\sum_{i=1}^{n} x_{ij}^2 - ||A||^2\right| > t\right) = P\left(\left|\sum_{i=1}^{n} Z_i\right| > t\right) \leq 2\exp\left\{-\frac{t^2}{2(n\sigma_1^2 + nt/3)}\right\} = 2\exp\left\{-\frac{t^2}{2n(\sigma_1^2 + t/3)}\right\}.$$

Similarly,

$$P\left(\left|\sum_{i=1}^{n} y_i^2 - ||B||^2\right| > t\right) \leq 2\exp\left\{-\frac{t^2}{2n(\sigma_2^2 + t/3)}\right\}, \sigma_2^2 = Var(y_i^2).$$

$$\sum_{i=1}^{n} x_{ij}y_i - AB = \sum_{i=1}^{n}(x_{ij}y_i - \frac{1}{n}AB).$$

Let $Z_i = x_{ij}y_i - \frac{1}{n}AB = x_{ij}y_i - E(AB)$, we have

$$E(Z_i) = E(x_{ij}y_i - E(AB)) = E(x_{ij}y_i) - E(AB) = 0,$$

$$Var(\sum_{i=1}^{n} Z_i) = nVar(Z_i) = nVar(x_{ij}y_i - E(AB)) = nVar(x_{ij}y_i) = n\sigma_3^2,$$

then we have by Lemma 1

$$P\left(\left|\sum_{i=1}^{n} x_{ij}y_i - AB\right| > t\right) = P\left(\left|\sum_{i=1}^{n} Z_i\right| > t\right) \leq 2exp\left\{-\frac{t^2}{2(n\sigma_3^2 + nt/3)}\right\}) = 2exp\left\{-\frac{t^2}{2n(\sigma_3^2 + t/3)}\right\}).$$

□

Proof of Theorem 1. According to the definitions of $\widehat{w_j}$ and w_j, we have

$$\begin{aligned}
&\widehat{w_j} - w_j \\
=&\frac{\sum_{i=1}^{n} x_{ij}y_i}{\sum_{i=1}^{n} {x_{ij}}^2 + \sum_{i=1}^{n} {y_i}^2 - \sum_{i=1}^{n} x_{ij}y_i} - \frac{AB}{||A||^2 + ||B||^2 - AB} \\
=&\frac{1}{||A||^2 + ||B||^2 - AB}(\sum_{i=1}^{n} x_{ij}y_i - AB) + (\sum_{i=1}^{n} x_{ij}y_i)\left(\frac{1}{\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} x_{ij}y_i} - \frac{1}{||A||^2 + ||B||^2 - AB}\right) \\
=&: I_1 + I_2,
\end{aligned}$$

then,

$$P(\left|\widehat{w_j} - w_j\right| > 2\epsilon) \leq P(|I_1| > \epsilon) + P(|I_2| > \epsilon).$$

For I_1, then we have inequality

$$|I_1| = \left|\frac{1}{||A||^2 + ||B||^2 - AB}(\sum_{i=1}^{n} x_{ij}y_i - AB)\right| = \frac{1}{||A||^2 + ||B||^2 - AB}\left|\sum_{i=1}^{n} x_{ij}y_i - AB\right| \leq \left|\sum_{i=1}^{n} x_{ij}y_i - AB\right|.$$

$$P(|I_1| > \epsilon) \leq P(\left|\sum_{i=1}^{n} x_{ij}y_i - AB\right| > \epsilon) \leq 2\exp\left\{-\frac{\epsilon^2}{2n(\sigma_3^2 + \epsilon/3)}\right\}.$$

For I_2, then

$$\begin{aligned}
I_2 \leq&(\sum_{i=1}^{n} x_{ij}y_i)\left\{\frac{1}{\sum_{i=1}^n {x_{ij}}^2 + \sum_{i=1}^n {y_i}^2 - \sum_{i=1}^n x_{ij}y_i} - \frac{1}{||A||^2 + ||B||^2 - AB}\right\}\\
=&(\sum_{i=1}^{n} x_{ij}y_i)\left\{\frac{||A||^2 + ||B||^2 - AB - (\sum_{i=1}^n x_{ij}^2 + \sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_{ij}y_i)}{(\sum_{i=1}^n x_{ij}^2 + \sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_{ij}y_i)(||A||^2 + ||B||^2 - AB)}\right\}\\
=&(\sum_{i=1}^{n} x_{ij}y_i)\left\{\frac{||A||^2 - \sum_{i=1}^n x_{ij}^2 + ||B||^2 - \sum_{i=1}^n y_i^2 + \sum_{i=1}^n x_{ij}y_i - AB}{(\sum_{i=1}^n x_{ij}^2 + \sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_{ij}y_i)(||A||^2 + ||B||^2 - AB)}\right\}\\
=&(\sum_{i=1}^{n} x_{ij}y_i)\left\{\frac{||A||^2 - \sum_{i=1}^n x_{ij}^2 + ||B||^2 - \sum_{i=1}^n y_i^2 + \sum_{i=1}^n x_{ij}y_i - AB}{(\sum_{i=1}^n x_{ij}^2 + \sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_{ij}y_i)(||A||^2 + ||B||^2 - AB)}\right\}\\
=&\frac{\sum_{i=1}^n x_{ij}y_i}{(\sum_{i=1}^n x_{ij}^2 + \sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_{ij}y_i)(||A||^2 + ||B||^2 - AB)}\\
&\left\{(||A||^2 - \sum_{i=1}^{n} x_{ij}^2) + (||B||^2 - \sum_{i=1}^{n} y_i^2) + (\sum_{i=1}^{n} x_{ij}y_i - AB)\right\}\\
\leq&(||A||^2 - \sum_{i=1}^{n} x_{ij}^2) + (||B||^2 - \sum_{i=1}^{n} y_i^2) + (\sum_{i=1}^{n} x_{ij}y_i - AB)\\
=&:I_{21} + I_{22} + I_{23},
\end{aligned}$$

then,

$$P(|I_2| > \epsilon) \leq P(|I_{21}| > \frac{\epsilon}{3}) + P(|I_{22}| > \frac{\epsilon}{3}) + P(|I_{23}| > \frac{\epsilon}{3}).$$

For I_{21}, then we have inequality

$$P(|I_{21}| > \epsilon) \leq P\left(\left|||A||^2 - \sum_{i=1}^{n} x_{ij}^2\right| > \frac{\epsilon}{3}\right) = P\left(\left|\sum_{i=1}^{n} x_{ij}^2 - ||A||^2\right| > \frac{\epsilon}{3}\right) \leq 2\exp\left\{-\frac{\epsilon^2}{2n(9\sigma_1^2 + \epsilon)}\right\}.$$

For I_{22}, then we have inequality

$$P(|I_{22}| > \epsilon) \leq P\left(\left|||B||^2 - \sum_{i=1}^{n} y_i^2\right| > \frac{\epsilon}{3}\right) = P\left(\left|\sum_{i=1}^{n} y_i^2 - ||B||^2\right| > \frac{\epsilon}{3}\right) \leq 2\exp\left\{-\frac{\epsilon^2}{2n(9\sigma_2^2 + \epsilon)}\right\}.$$

For I_{23}, then we have inequality

$$P(|I_{23}| > \epsilon) \leq P\left(\left|\sum_{i=1}^{n} x_{ij}y_i - AB\right| > \frac{\epsilon}{3}\right) \leq 2\exp\left\{-\frac{\epsilon^2}{2n(9\sigma_3^2 + \epsilon)}\right\}.$$

In sum, we have inequality

$$P(\left|\widehat{w}_j - w_j\right| > 2\epsilon) \leq \exp\left\{-\frac{\epsilon^2}{2n(c_1\sigma_1^2 + c_2\sigma_2^2 + c_3\sigma_3^2 + c_4\epsilon)}\right\} = \exp\left\{-\frac{\epsilon^2}{2n(\sum_{v=1}^3 c_v{\sigma_v}^2 + c_4\epsilon)}\right\},$$

where c_1, c_2, c_3, c_4 is constant. □

Proof of Theorem 2. By Theorem 1 and Condition1, we have

$$\begin{aligned}
P(\mathcal{A} \subseteq \widehat{\mathcal{A}_*}) &\geq P(\left|\widehat{w_j} - w_j\right| \leq cn^{-\kappa}, \forall j \in D) \\
&\geq P(\max_{1\leq j\leq J} \left|\widehat{w_j} - w_j\right| \leq cn^{-\kappa}) \\
&\geq 1 - \sum_{j=1}^{J} P(\max_{1\leq j\leq J} \left|\widehat{w_j} - w_j\right| > cn^{-\kappa}) \\
&\geq 1 - O\left(\exp\left\{-\frac{c^2 n^{-2\kappa}}{4n(2\sum_{v=1}^{3} c_v \sigma_v^2 + c_4 cn^{-\kappa})}\right\}\right) \\
&\geq 1 - O\left(\exp\left\{-\frac{c^2 n^{-\kappa}}{c_6}\right\}\right),
\end{aligned}$$

where c_6 is a constant. □