Measuring Reproducibility of High-Throughput Deep-Sequencing Experiments Based on Self-adaptive Mixture Copula

Qian Zhang[1], Junping Zhang[1,⋆], and Chenghai Xue[2,⋆]

[1] Shanghai Key Lab of Intelligent Information Processing
School of Computer Science, Fudan University, China
[2] Cold Spring Harbor Laboratory, NY
qianzhang.fdu@gmail.com, jpzhang@fudan.edu.cn, xuec@cshl.edu

Abstract. Measurement of the statistical reproducibility between biological experiment replicates is vital first step of the entire series of bioinformatics analysis for mining meaningful biological discovery from mega-data. To distinguish the real biological relevant signals from artificial signals, irreproducible discovery rate (IDR) employing Copula, which can separate dependence structure and marginal distribution from data, has been put forth. However, IDR employed a Gaussian Copula which may cause underestimation of risk and limit the robustness of the method. To address the issue, we propose a Self-adaptive Mixture Copula (SaMiC) to measure the reproducibility of experiment replicates from high-throughput deep-sequencing data. Simple and easy to implement, the proposed SaMiC method can self-adaptively tune its coefficients so that the measurement of reproducibility is more effective for general distributions. Experiments in simulated and real data indicate that compared with IDR, the SaMiC method can better estimate reproducibility between replicate samples.

1 Introduction

During the past years, the biological technology revolution, next-generation high-throughput deep-sequencing, has produced mountains of data of DNA-Seq, RNA-Seq and ChIP-Seq. This mega-data allows biologists to observe the signals from tens of thousands of genes or related genomic elements in a single experiment, a way that was not possible before. One question arises here: how many of these signals are real biologically relevant? Generally, to avoid experiment noise or error, two experiment replicates of one biological sample should be produced, each of which has a collection of individual elements or signals such as a list of genes or transcripts. Then we need to verify the reproducibility of each individual signal. Only the individual signals with high reproducibility are considered as reliable results for further analysis such as differential gene expression identification or GO analysis. Here reproducibility of a signal's two observations in

[⋆] Corresponding authors.

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 301–313, 2013.

two replicates is a measure of the confidence that the two observations are consistent with each other. Hereinafter it is shorted as "*reproducibility of signal*". We propose a posterior probability to characterize the confidence and also define irreproducibility = 1 - reproducibility. By choosing a specific critical value, each signal can be determined whether or not to be confident. It is worth noting that we only have two replicates, two lists of observations of individual signals but need to detect the reproducibility of each individual signal. Such extremely small samples also lead to a big and difficult challenging task to traditional data mining where data are generally of remarkably larger size and density estimation can be effectively calculated based on these data.

IDR (Irreproducible Discovery Rate) which measures the reproducibility in high-throughput experiments has been put forth by Li [1]. They proposed to use copula, which can separate the dependency structure of random variables, and a measurement based on copula to detect the high reproducible signals. A remarkable advantage of copula is that it provides an effective way to infer the dependency structure between biological signals without knowing their respective marginal distribution, which is difficult to obtain from real biological signals. Reproducibility measure has been adopted as the standard of ENCODE (The Encyclopedia of DNA Elements) project and has been carried on each signal of all samples before these data are submitted to public database. The strategy of IDR has been generalized to use on other data types such as RNA-Seq.

Although IDR has shown its ability of distinguishing the bona fide signals from artificial signals, it employed the Gaussian Copula, which assumes that the dependence structure of random variables follows multivariate Gaussian distribution. This assumption causes that the Gaussian Copula is sensitive to extreme events and can not capture the asymmetric dependence structure [2]. In fact, despite its simplicity, Gaussian Copula often leads to an underestimation of the risk of the occurrence of joint extreme events [3, 4].Therefore, it is necessary to develop a novel and efficient approach to measure the reproducibility of replicate data without stronger assumption to data distribution.

In this paper, we propose a **S**elf-**a**daptive **Mi**xture **C**opula, called SaMiC, to measure the reproducibility of the high-throughput deep-sequencing experiments. Unlike IDR, SaMiC doesn't assume the dependence structure of random variables to follow Gaussian distribution. SaMiC mixes several copulas and automatically determines the mixture coefficients based on the fitness of the data and the copulas. We prove theoretically that the new mixture copula is still a copula so that it can be used to measure the reproducibilities. Simple and easy to implement, SaMiC is effective and suitable for general distributions. Experiments in both simulated data and real biological data of RNA transcripts expression from human cells show that compared with IDR, SaMiC attains better performance in distinguishing bona fide signals from artificial signals.

The remainder of this paper is organized as following. In Section 2, we introduce the development and preliminary of copulas. In Section 3 we detail our proposed self-adaptive mixture copula and a novel measurement to detect the reproducibility of experiments. In Section 4, we perform experiments in simulated data and real data. In Section 5, we conclude the paper.

2 Related Work and Preliminary of Copula

As a tool of extracting the dependence structure from joint distributions of random variables, copula was first proposed by Sklar [5]. In his work, copula is obtained by a two-stage procedure, *i.e.*, estimating the marginal distribution of each random variable followed by measuring the dependence structure between different random variables. Deheuvels proposed several empirical functions, *i.e.*, the empirical copula of samples, to estimate the copula of population and constructed different non-parametric dependence tests from samples [6]. However, a well-recognized definition of copula hasn't been given until Nelsen's work [7].

There are two major categories in studying copulas: parameter estimation and test of goodness of fit. In the former category, Oakes and Genest proposed a common strategy to estimate the parameters of bivariate copula [8,9]. Later, Joe investigated the maximum likelihood estimation of parametric marginal distribution and parametric copula [10]. Furthermore, Chen studied two stage semi-parametric maximum likelihood estimation [11], and Abegaz derived asymptotic properties of the marginal and copula parameter estimators [12].

In the aspect of test of goodness of fit, an important goal is to measure how well a copula describes the dependence structure among random variables since it is closely related to the correctness of the proposed copula. According to the copula model, test of goodness of fit can be transformed into test of univariate distribution. Then Kolmogorov-Smirnov test can be used to test the goodness of fit of copula. In this manner, Klugman used Q-Q plot to measure the rationality of copula model [13]. Hu introduced M-statistics, which follows the chi-square distribution, to measure goodness of fit of copula model [14]. Engle proposed a test method named “Hit” [15], and Patton expanded the test method “Hit” to the nonlinear density model for checking the goodness of fit [16]. Theses methods can evaluate both the copula and the marginal distribution.

Since the function is useful to obtain the dependence structure of multivariate random variables with few assumptions, it has been applied in financial field. Embrechts employed copula for financial risk management [17]. Hu proposed to use mixed-copula to analyze the financial data [14]. However, such a mixture is not automatic and depends on experts' experience. Recently, copula was also employed in bioinformatics field. For example, Kim discussed the application in genetic data [18], Zhang used copula model to analyze ChIP-Seq data [19], and Li proposed a new method based on copula model to measure reproducibility of bioinformatics data [1]. A main reason of using copula in financial and bioinformatics fields is that it can obtain the dependence structure without knowing marginal distribution of random variables in advance. It is worth noting that the above-mentioned copulas still require more or less assumptions to data distribution, which may limit its effectiveness and extension. It is also noticeable that in bioinformatics field, copula is still a new tool of data analysis. For better understanding, we introduce some preliminaries of copula as follows.

Theorem 1 (Sklar's Theorem). *[5] Let H be a joint distribution function with margins F_1 and F_2. Then there exists a copula C such that*

$$H(x,y) = C(F_1(x), F_2(y)), \quad \forall x, y \in \boldsymbol{R} \tag{1}$$

If F_1 and F_2 are continuous, then C is unique; otherwise, C is uniquely determined on $RanF_1 \times RanF_2$. Here $RanF$ refers to the range of F. Conversely, if C is a copula and F_1 and F_2 are distribution functions, then the function H is a joint distribution function with margins F_1 and F_2.

Sklar's Theorem shows that a joint distribution function can be divided into each variable's marginal distribution function and a copula which present their statistical consistence. Therefore, it is possible to calculate copulas from joint distribution function and its marginal distribution functions. From Sklar's Theorem we can get the follow properties, which are important to measure the reproducibility of high-throughput deep-sequencing experiments:

Property 1. Let $G(X_1, X_2, \cdots, X_n)$ be a joint distribution function of n random variables $X_1, X_2, \cdots, X_n$, $F_1(x_1), F_2(x_2), \cdots, F_n(x_n)$ are marginal distribution functions of these random variables and $C(u_1, u_2, \cdots, u_n)$ is the corresponding copula, then for every $\mathbf{u} = (u_1, u_2, \cdots, u_n) \in [0,1]^n$ it satisfies that

$$C(u_1, u_2, \cdots, u_n) = G(F_1^{-1}(u_1), F_2^{-1}(u_2), \cdots, F_n^{-1}(u_n)), \tag{2}$$

where $F_i^{-1}(u_i)$ is the right-continuous inverse of F_i, defined as $F_i^{-1}(u_i) = \inf\{z : F_j(z) \geq u_i\}$.

Property 2. Let $G(X_1, X_2, \cdots, X_n)$ be a joint distribution function of n random variables $X_1, X_2, \cdots, X_n$, $C(u_1, u_2, \cdots, u_n)$ be a copula, $c(u_1, u_2, \cdots, u_n)$ be its density function and $F_1, F_2, \cdots, F_n$ are marginal distribution functions of the random variables, we can get that

$$g(X_1, X_2, \cdots, X_n) = c(F_1(X_1), F_2(X_2), \cdots, F_n(X_n)) \prod_{i=1}^{n} f_i(X_i), \tag{3}$$

where $c(u_1, u_2, \cdots, u_n) = \frac{\partial C(u_1, u_2, \cdots, u_n)}{\partial u_1 \partial u_2 \cdots \partial u_n}$, and $f_i(X_i)$ and $g(X_1, X_2, \cdots, X_n)$ are density functions of $F_i(X_i)$ and $G(X_1, X_2, \cdots, X_n)$, respectively.

Currently, there are two main types of commonly-used copulas: Elliptical copulas and Archimedean copulas. Elliptical copulas are a kind of copulas with contoured elliptical distributions, such as Gaussian copula and t-copula. Easy to construct, Elliptical copulas don't have a closed form of function expression, and all of them are radially symmetric and hard to extend to high-dimensional situation. As a result, it is difficult to use Elliptical copulas to describe unpredictable dependence structures or adapt to complex situations.

Different from Elliptical copulas, Archimedean copulas are an associative class of copulas which satisfy the following equations:

$$C(u_1, u_2, \cdots, u_n) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2) + \cdots + \varphi(u_n)) \tag{4}$$

where $\varphi(\cdot)$ is usually called the generator of Archimedean copula. Among a lot of Archimedean copulas, three frequently used Archimedean copulas are Frank

Copula, Clayton Copula and Gumbel Copula. Specifically, the reproducibility structures of Frank Copula and the variables drawn from Frank Copula are symmetric in both tails of their distributions. So any asymmetric consistence between random variables can't be captured using Frank Copula. Clayton Copula is sensitive to the low tail dependence of random variables, and can easily capture the changes around the low tail. Finally, Gumbel Copula is sensitive to the upper tail dependence of random variables.

3 The Proposed SaMiC Approach

It is obvious that each copula has its respective pros and cons in different distributions. To deal with more general distributions, we propose self-adaptive mixture copula, which is a linear combination of several copulas. For simplification, we discuss the proposed copula in two-dimensional situation.

Definition 1 (Mixture Copula). *A function is called mixture copula if it satisfies:* $C_M(u,v) = \sum_{i=1}^{m} \alpha_i C_i(u,v)$, *where* $0 \leq \alpha_i \leq 1$, $\sum_{i=1}^{m} \alpha_i = 1$. *Here* $C_1(u,v), C_2(u,v), \cdots, C_m(u,v)$ *are copulas and* $\alpha_1, \alpha_2, \cdots, \alpha_m$ *are linear coefficients of* $C_M(u,v)$.

We prove that $C_M(u,v)$ is a copula, which will be introduced in a extended version due to the length limitation. To self-adaptively estimate the linear coefficients of the proposed mixture copulas, we utilize Pearson χ^2 statistic proposed by Hu to measure the goodness of fit of each copula [14], i.e., $M = \sum_i^k \sum_j^k \frac{(A_{i,j} - B_{i,j})^2}{B_{i,j}}$, where $A_{i,j}$ and $B_{i,j}$ denote the number of observed and predicted frequencies in cell (i,j) of a contingency table, respectively. The details on the contingency table can be referred to as Hu [14]. M follows χ^2 distribution with $(k-1)^2$ degree of freedom. Given a total of m base copulas, $C_1, C_2, \cdots, C_m$, we can attain m results, $M_{C_1}, M_{C_2}, \cdots, M_{C_m}$. Because M_{C_i} follows χ^2 distribution, we can get probability β_i from M_{C_i}. In fact β_i is the probability that there is no significant difference between copula C_i and the data. Because of the additivity of chi-squared distribution, let $\alpha_i = \frac{\beta_i}{\sum_{i=1}^{m} \beta_i}$ be the mixing coefficient of C_i, then the proposed self-adaptive mixture-copula is

$$C_M(u,v) = \sum_{i=1}^{m} \alpha_i C_i(u,v) = \frac{1}{\sum_{i=1}^{m} \beta_i} \sum_{i=1}^{m} \beta_i C_i(u,v) \tag{5}$$

Since our self-adaptive method chooses coefficients automatically, it can deal with more general distributions. By contrast, the ordinary mixture copula manually selects the coefficients, heavily depending on human experience and need lots of tuning for each new group of data [14]. Consequently, it is only applicable to some specific distributions.

To measure the statistical consistency or reproducibility based on the proposed self-adaptive mixture copula, we here consider the situation with two rows of observations for simplification. The reason is that when observations subject to independent identically distribution, it is not difficult to expand to multivariate situation if we use a pairwise analysis to them.

Formally, let the two rows of observations, $(x_{1,1}, x_{1,2}), \cdots, (x_{n,1}, x_{n,2})$, be two replicates of n random signals. We assume that the observations consist of a more reproducible group and a less reproducible one, and π_0 and π_1 denote the proportion of the less reproducible group and the more reproducible group, respectively. Let parameter K_i be an indicator to identify whether or not a signal belong to the more or less reproducible group. $K_i = 1$ if the *i-th* signal belong to the more reproducible group, and $K_i = 0$ if it is in the less reproducible group.

Obviously, the signals in the more or less reproducible group have different probability distributions. We assume that the dependence structures of the two observations of signals in the more and less reproducible groups are induced by $\boldsymbol{z_1} = (z_{1,1}, z_{1,2})$ and $\boldsymbol{z_0} = (z_{0,1}, z_{0,2})$, respectively. According to Sklar's Theorem, any multivariate probability distribution can be divided into its marginal probability distributions and its copula. In other words, it provides a way to infer the reproducibility among several random variables without knowing their marginal distributions in advance. Thus, we construct our parametric model as follows:

Let $K_i \sim \mathit{Bernoulli}(\pi_1)$ and $(z_{i,1}, z_{i,2})$ be distributed as $(z_{i,1}, z_{i,2}) \mid K_i = k \sim S_k(u, v), k = 0, 1$, and

$$S_k(z_{i,1}, z_{i,2}) = C(F_1(z_{i,1}), F_2(z_{i,2}); \lambda_k), \quad k = 0, 1 \tag{6}$$

where $F_1(z_{i,1})$ and $F_2(z_{i,2})$ are the marginal distributions of $z_{i,1}$ and $z_{i,2}$, respectively. And λ_k denotes the relevant parameter of copula. Then considering (6), the total distribution function is

$$\begin{aligned} S(z_{i,1}, z_{i,2}) = P\{Z_{i,1} \le z_{i,1}, Z_{i,2} \le z_{i,2}\} &= \pi_0 S_0(z_{i,1}, z_{i,2}) + \pi_1 S_1(z_{i,1}, z_{i,2}) \\ &= \sum_{k=0,1} \pi_k C(F_1(z_{i,1}), F_2(z_{i,2}); \lambda_k) \end{aligned} \tag{7}$$

In this equation, our actual observations $(x_{i,1}, x_{i,2})$ are used to estimate the cumulative distribution function of $(z_{i,1}, z_{i,2})$. From (3) we can get the density functions of $S_0(z_{i,1}, z_{i,2})$ and $S_1(z_{i,1}, z_{i,2})$ as follows:

$$s_k(z_{i,1}, z_{i,2}) = c(F_1(z_{i,1}), F_2(z_{i,2}); \lambda_k) f_1(z_{i,1}) f_2(z_{i,2}), \quad k = 0, 1, \tag{8}$$

where $f_1(z_{i,1})$ and $f_2(z_{i,2})$ are the density functions of $F_1(z_{i,1})$ and $F_2(z_{i,2})$, respectively. From (7) and (3), therefore, the density function of $S(z_{i,1}, z_{i,2})$ is

$$s(z_{i,1}, z_{i,2}) = \sum_{k=0,1} \pi_k c(F_1(z_{i,1}), F_2(z_{i,2}); \lambda_k) f_1(z_{i,1}) f_2(z_{i,2}) \tag{9}$$

Up to now, our model is parametrized by $\boldsymbol{\theta} = (\pi_0, \lambda_0, \lambda_1)$ and F_1, F_2. The parameters can be attained using maximum likelihood estimation as:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} s(x_{i,1}, x_{i,2}) = \prod_{i=1}^{n} (f_1(x_{i,1}) f_2(x_{i,2}) \sum_{k=0,1} \pi_k c(F_1(x_{i,1}), F_2(x_{i,2}); \lambda_k)). \tag{10}$$

Note that selecting different Archimedean copulas in (6) will lead to different forms of $S(z_{i,1}, z_{i,2})$. Since

$$C_M(u, v; \boldsymbol{\theta}) = \pi_0 C(u, v; \lambda_0) + (1 - \pi_0) C(u, v; \lambda_1) \tag{11}$$

is also a copula, the formula (7) can be rewritten as $S(z_{i,1}, z_{i,2}) = C_M(F_1(z_{i,1}),$ $F_2\ (z_{i,2}); \boldsymbol{\theta})$, where the form of the final copula $C_M(u, v)$ is determined by the selection of Archimedean copulas in (6).

When several Archimedean copulas $C_{(1)}(u, v)$, $C_{(2)}(u, v)$, $\cdots$, $C_{(m)}(u, v)$ are substituted into (6), the corresponding parameters $\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}, \cdots, \boldsymbol{\theta}_{(m)}$ can be estimated from (10), where $\boldsymbol{\theta}_{(i)} = (\pi_{(i),0}, \lambda_{(i),0}, \lambda_{(i),1})$. After that, we substitute the obtained copulas $C_{M(1)}(u, v; \boldsymbol{\theta}_{(1)})$, $C_{M(2)}(u, v; \boldsymbol{\theta}_{(2)})$, $\cdots$, $C_{M(m)}(u, v; \boldsymbol{\theta}_{(m)})$ into the method of generating self-adaptive mixture-copula. Because $C_{M(i)}(u, v; \boldsymbol{\theta}_{(i)})$ is also a mixture-copula, we get the linear coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_m)$ and attain the final mixture-copula as:

$$C_R(u, v) = \sum_{i=1}^{m} \alpha_i C_{M(i)}(u, v; \boldsymbol{\theta}_{(i)}) \tag{12}$$

where α_i is the linear coefficient. Since $C_{M(i)}(u, v; \boldsymbol{\theta}_{(i)})$ is also a mixture-copula, $C_R(u, v)$ can be decomposed into

$$\begin{aligned} C_R(u, v) &= \sum_{i=1}^{m} \alpha_i \pi_{(i),0} C_{(i)}(u, v; \lambda_{(i),0}) + \sum_{i=1}^{m} \alpha_i (1 - \pi_{(i),0}) C_{(i)}(u, v; \lambda_{(i),1}) \\ &= C_{R,0}(u, v) + C_{R,1}(u, v) \end{aligned} \tag{13}$$

where $C_{R,0} = \sum_{i=1}^{m} \alpha_i \pi_{(i),0} C_{(i)}(u, v; \lambda_{(i),0})$ and $C_{R,1} = \sum_{i=1}^{m} \alpha_i (1 - \pi_{(i),0}) C_{(i)}(u,$ $v; \lambda_{(i),1})$ are the more and less reproducible groups, respectively.

Once $C_R(u, v)$ is determined, it is easy to update the total distribution function of the data as

$$S_R(z_{i,1}, z_{i,2}) \equiv S_{R,0}(z_{i,1}, z_{i,2}) + S_{R,1}(z_{i,1}, z_{i,2}), \tag{14}$$

and thus

$$\begin{aligned} &s_R(z_{i,1}, z_{i,2}) = \\ &c_{R,0}(F_1(z_{i,1}), F_2(z_{i,2})) f_1(z_{i,1}) f_2(z_{i,2}) + c_{R,1}(F_1(z_{i,1}), F_2(z_{i,2})) f_1(z_{i,1}) f_2(z_{i,2}) \end{aligned} \tag{15}$$

where s_R ,c_R and $s_{R,k}$ are density functions of S_R, C_R and $s_{R,k}$, respectively.

Finally, we can estimate the irreproducibility of each signal based on:

$$P\{K_i = 0 \mid (x_{i,1}, x_{i,2})\} = \frac{\pi_{R,0} s_{R,0}(x_{i,1}, x_{i,2})}{\sum\limits_{k=0,1} \pi_{R,k} s_{R,k}(x_{i,1}, x_{i,2})} \tag{16}$$

$P\{K_i = 0 \mid (x_{i,1}, x_{i,2})\}$ describes the probability that a signal's two observations $(x_{i,1}, x_{i,2})$ are irreproducible, i.e., the irreproducibility of $(x_{i,1}, x_{i,2})$. To estimate

Algorithm 1. The Proposed SaMic Approach

```
**Input:** data $(x_{i,1}, x_{i,2})$, size $n$, copulas $c_j(u, v; \lambda)$ and size $m$
Get $\hat{F}_1(x_{i,1})$ and $\hat{F}_2(x_{i,2})$
**for** $j = 1$ **to** $m$ **do**
  Initialize $\pi_0^{(0)}, \lambda_1^{(0)}$, $k = 0$
  **repeat**
    Initialize $noChange = false$
    $\pi_0^{(k+1)} = \arg\min_{\pi_0} \prod_{i=1}^{n} \{\pi_0 + (1 - \pi_0) * c_j(\hat{F}_1(x_{i,1}), \hat{F}_2(x_{i,2}), \lambda_1^{(k)})\}$
    $\lambda_1^{(k+1)} = \arg\min_{\lambda_1} \prod_{i=1}^{n} \{\pi_0^{(k+1)} + (1 - \pi_0^{(k+1)}) * c_j(\hat{F}_1(x_{i,1}), \hat{F}_2(x_{i,2}), \lambda_1)\}$
    **if** $|\pi_0^{(k+1)} - \pi_0^{(k)}| < \varepsilon$ and $|\lambda_1^{(k+1)} - \lambda_1^{(k)}| < \varepsilon$ **then** $noChange = true$ **end if**
  **until** $noChange = true$
  let $M_j$ be Pearson $\chi^2$ statistic of $c_j$
**end for**
get $C_R$ from (5), (11) and (12), and irreproducibility from (14), (15) and (16)
```

the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\theta}_{(i)}$ and the indicators K_i of each signal, we propose an effective two-stage one-dimensional optimization method. It's worth noting that we use empirical marginal CDFs $\hat{F}_1(x_{i,1})$ and $\hat{F}_2(x_{i,2})$ instead of using raw data directly, where $\hat{F}_j(x_{i,j}) = \frac{rank_j(x_{i,j})}{n+1}$ $j = 0, 1$, since it makes experiments comparable and the rank statistic tends to cope better with real-world systematic biases and errors. The pseudo-code of our estimation procedure is shown in Alg. 1.

Compared with the IDR proposed by Li [1], a remarkable advantage of our algorithm is that it is more effective since it only needs to do one-dimensional optimization search for no more than $2km$ times, where k is the threshold of iterations. So its asymptotic time complexity is $O(mn)$. The actual running time of our algorithm is also affected by the selected threshold of precision and iterations.

4 Experiments

To evaluate the effectiveness of our proposed SaMiC approach, we compare it with IDR proposed by Li [1] in two simulated data with remarkably different marginal distributions and reproducibility structures and one real biological data. Note that although IDR has four values to be initialized, we found that they have less influence to the analysis of the final results. We chose Frank Copula, Clayton Copula and Gumbel Copula that all from Archimedean family as base copulas since these copulas have high potential of extending from bivariate Archimedean copulas to multivariate ones.

4.1 Simulated Data

In the first experiment, we generate two rows of 10,000 numbers which follow normal distributions $N(0, 1)$ and $N(2, 12)$, respectively. Then we consider these numbers as 10,000 signals' two observations to detect their (ir)reproducibilities. For the two rows of numbers generated from different distributions, there's little

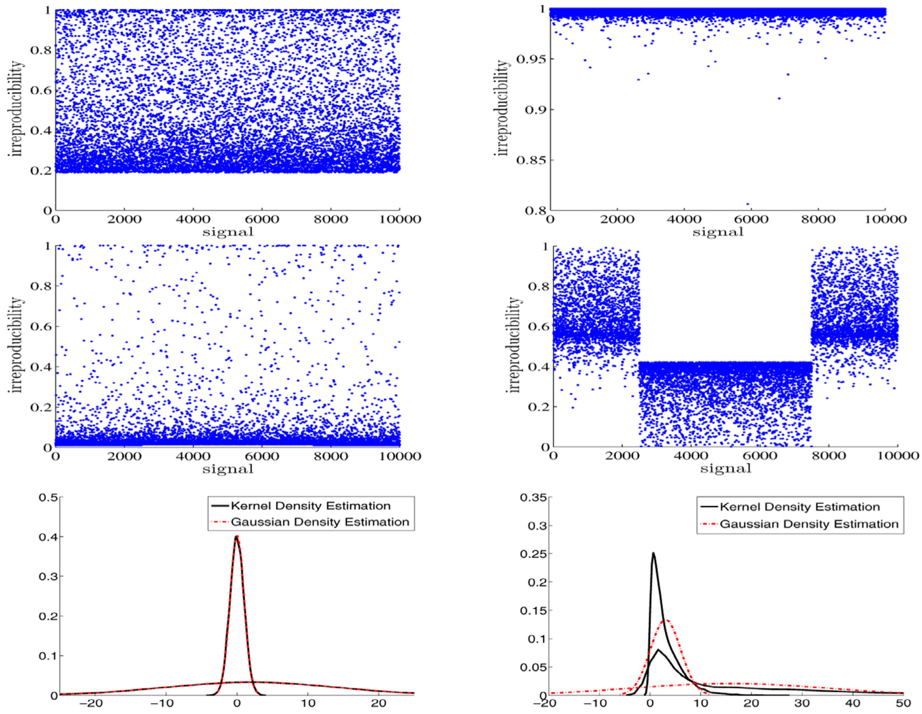

Fig. 1. Irreproducibilities (dot) of each signal observed by IDR (left) and SaMiC (right) in experiment 1 (Top) and experiment 2 (Middle). Bottom: Density Estimations of signals in experiment 1 (Left) and experiment 2 (Right).

chance that these signals' two observations are confidently consistent. So we expected the (ir)reproducibilities are low (high). Then we use both IDR and SaMiC to measure the (ir)reproducibilities of these signals. Note that both IDR and SaMiC output the irreproducibilities in $[0, 1]$. The results shown in Fig. 1 indicate that IDR has a lower recognition rate to discover the irreproducible signals. For example, if we regard those signals whose irreproducibilities are less than 0.5 are reproducible, then many irreproducible signals will be classified to be reproducible. In contrary, our SaMiC approach can correctly classify most irreproducible signals even when the cutoff value is set to be 0.9.

In the second experiment, we wish to test whether our SaMiC can be suitable for a more general distribution. Thus, we generated two rows of 10,000 random numbers combined from two different types of distributions. Firstly, we generated 5,000 random numbers $(t_1, t_2, \cdots, t_{5000})$ from Chi-square distribution $T \sim \Gamma(2, 14)$, and 10,000 random numbers $(a_1, a_2, \cdots, a_{10000})$ from beta distribution $A \sim \beta(3, 3)$. Let the first row be $(a_1, a_2, \cdots, a_{10000})$ and the second row be $(t_1, t_2, \cdots, t_{2500}, a_{2501}, a_{2502}, \cdots, a_{7500}, t_{2501}, t_{2502}, \cdots, t_{5000})$. From Fig. 1 it's obvious that IDR failed to distinguish the irreproducible signals. In

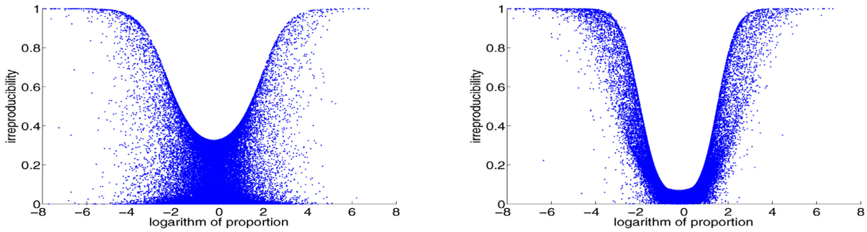

Fig. 2. HeLa-S3: The log-proportion vs. irreproducibility figure of IDR (**left**) and SaMiC (**right**)

contrast, SaMiC can estimate the (ir)reproducibilities of signals in experiment 2 with high confidence. The reason is that SaMiC makes less assumption to the dependence structure of observations, and the self-adaptive mixture copula is helpful to be suitable for general distributions.

We also use both kernel and Gaussian density estimation on these data of the two above experiments. As demonstrated in Fig. 1, the results from density estimation can show the differences between two rows of numbers only in an overall perspective. So it's hard to decide whether or not to trust some specific signals that are reproducible by using density estimation. In contrast, our method attains the (ir)reproducibility of each signal, which can distinguish bona fide signals from artificial signals.

4.2 Real Data

We also use real biological data to test the performance of SaMiC. The data can be downloaded from "http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19 &g=wg EncodeCshlLongRnaSeq" (selected categories: Cell Line = HeLa-S3, Location = cell, RNA Extract = Long PolyA+ RNA, View: Transcript Gencode V7). This data was generated by ENCODE project [20] and they are biological experiments to detect the expression level of HeLa-S3 cell's long RNA transcripts, which were sequenced by RNA-Seq. The downloaded data file contains 161,999 annotated transcript individuals' expression values — the normalized RPKM values. As need, each transcript has two values from different experiment replicates respectively. They are estimated by SaMiC and IDR to test their performance. Different from some classical data mining problems, it's difficult to verify the experiment results on real data because of lacking test data or labels. So we need to analyze the results in some indirect ways.

Intuitively, the larger (or smaller) the proportion of a signal's two observations is, the smaller probability that the signal is reproducible. In fact, SaMiC scores are different from proportions because proportions only consider local information while SaMiC scores rely on both the entire distribution and the dependence structure of the observations. Nevertheless, it still can demonstrate some differences between IDR and SaMiC by using figure of proportions versus

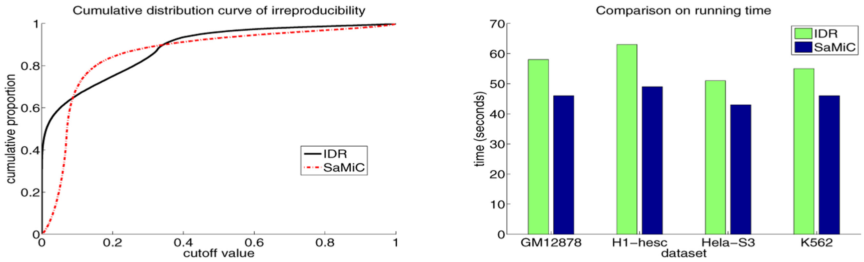

Fig. 3. Cumulative distribution curve (**left**) and comparison on running time (**right**)

irreproducibilities. As shown in Fig. 2, we draw this figure by putting logarithm of proportion on x-axis and irreproducibility on y-axis. From Fig. 2 we can see that SaMiC is more sensitive and has a stronger recognizing ability. Take signals in $(-\infty, -2] \cup [2, \infty)$ with irreproducibilities lower than 0.2 for example. The number of irreproducible signals estimated by IDR is remarkably larger than that estimated by SaMiC. It shows that SaMiC is more sensitive to (ir)reproducibility and can identify the irreproducible signals which are ignored by IDR.

For the convenience of subsequent data analysis, such as keeping specific proportion of data or choosing different critical values, we expect the irreproducibilities to be smooth. In order to compare IDR and SaMiC from this viewpoint, we produce the cumulative distribution curves of the results from both IDR and SaMiC. From Fig. 3 we observe that the curve of SaMiC is smoother than that of IDR. Besides, compared with IDR, SaMiC can provide more detailed data for keeping specific proportion of data. For example, if we want to get the signals with the lowest 20% irreproducibilities, it's easy while using SaMiC but unavailable while using IDR. This is because that there are almost 40% irreproducibilities that are 0 in the result of IDR, which means IDR fails to discriminate the signals with 40% lowest irreproducibilities while SaMiC succeeds. So SaMiC performs better on selecting the most reliable signals with a specific proportion.

Furthermore, we perform more experiments on another three different types of cells including GM12878, H1-hesc and K562, which are downloaded from the same website as HeLa-S3. For saving space, the detailed results can be found in future extended version. Based on these experiments, we give a comparison on running time. As shown in Fig. 3, SaMiC works faster than IDR on all of the four data. The computing environment is Intel Core2 2.93GHz with 4G memory.

5 Conclusion

In this paper, we have proposed a Self-adaptive Mixture Copula to measure the reproducibility of high-throughput deep-sequencing experiments, which is a difficult and challenging data mining problem since the number of samples is extremely small. The proposed SaMiC can effectively separate the dependence structure from joint distribution of signals without *priori* assumption. Compared with IDR, SaMiC can discover the irreproducible signals in a more reliable way.

SaMiC features no parameters that need to be tuned and can calculate the (ir)reproducibilities in an automatic way. It can self-adaptively choose the most suitable parameters for given data and is thus robust for different datasets. Besides, SaMiC works faster than IDR on all data we tested.

SaMiC can be used in all high-throughput deep-sequencing experiments that produce over one replicate to avoid reducing the confidence of experimental results. Actually, the reproducibility issue exists for a great number of researches so that the method of estimating reproducibility has a wide application.

In the future, we will compare SaMiC with other methods such as FDR. Furthermore, we will do more experiments with labeled data and investigate more application fields of SaMiC.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China (NSFC No. 61273299, 60975044) and Ministry of Education (No. 20120071110035).

References

1. Li, Q., Brown, J.B., Huang, H., Bickel, P.: Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics 5(3), 1752–1779 (2011)
2. Kole, E., Koedijk, K., Verbeek, M.: Selecting copulas for risk management. Journal of Banking & Finance 31(8), 2405–2423 (2007)
3. Frey, R., McNeil, A.: Dependent defaults in models of portfolio credit risk. Journal of Risk 6, 59–92 (2003)
4. Trivedi, P., Zimmer, D.: Copula modeling: an introduction for practitioners, vol. 1. Now Pub. (2007)
5. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 3, 229–231 (1959)
6. Deheuvels, P.: A Kolmogorov-Smirnov type test for independence and multivariate samples. Rev. Roumaine Math. Pures Appl. 26(2), 213–226 (1981)
7. Nelsen, R.B.: An introduction to copulas. Springer, New York (1999)
8. Oakes, D.: Multivariate survival distributions. Nonparametric Statistics 3(3-4), 343–354 (1994)
9. Genest, C., Ghoudi, K., Rivest, L.P.: A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82(3), 543–552 (1995)
10. Joe, H.: Asymptotic efficiency of the two-stage estimation method for copula-based models. Journal of Multivariate Analysis 94(2), 401–419 (2005)
11. Chen, X., Fan, Y.: Estimation of copula-based semiparametric time series models. Journal of Econometrics 130(2), 307–335 (2006)
12. Abegaz, F., Naik-Nimbalkar, U.V.: Modeling statistical dependence of markov chains via copula models. Journal of Statistical Planning and Inference 138(4), 1131–1146 (2008)
13. Klugman, S.A., Parsa, R.: Fitting bivariate loss distributions with copulas. Insurance: Mathematics and Economics 24(1-2), 139–148 (1999)
14. Hu, L.: Dependence patterns across financial markets: a mixed copula approach. Applied Financial Economics 16(10), 717–729 (2006)

15. Engle, R.F., Manganelli, S.: Caviar. Journal of Business and Economic Statistics 22(4), 367–381 (2004)
16. Patton, A.J.: Modelling asymmetric exchange dependence. International Economic Review 47(2), 527–556 (2006)
17. Embrechts, P., McNeil, A., Straumann, D.: Correlation: pitfalls and alternatives. RISK Magazine 12, 69–71 (1999)
18. Kim, J.M., Jung, Y.S., Sungur, E., Han, K.H., Park, C., Sohn, I.: A copula method for modeling directional dependence of genes. BMC Bioinformatics 9(225) (2008)
19. Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nussbaum, C., Myers, R., Brown, M., Li, W., et al.: Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9(9), R137 (2008)
20. Myers, R., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R., Bernstein, B., Gingeras, T., Kent, W., Birney, E., et al.: A user's guide to the encyclopedia of dna elements (ENCODE project consortium). PLoS Biol. 9(4), e1001046 (2011)