Sample Expansion Approach and Simulation Analysis for Low-Capacity Samples Based on Distance-Trend Double Effects

Shaoqian Huang+, Qi Zhou+, and Hongqing Wang

***Abstract*—In real-world scenarios, samples are sometimes unavailable in large quantities. Although there may be access to sufficient samples, these samples often have high acquisition costs. Additionally, general statistical methods require large sample sets to guarantee a certain accuracy. Therefore, low-capacity sample sets must be expanded. In this paper, from the perspective of sample component evolution and the distance between a pair of samples, a new approach for expanding the sample capacity is proposed. Through simulation, this sample expansion technique based on the distance-trend double effects (DTDE) is shown to be effective and accurate. In addition, we introduce two special cases and further analyze the proposed algorithm.**

***Index Terms*—sample capacity expansion, distance-trend effects, neural network, copula, simulation.**

I. Introduction

WITH the further development of society, the complexity of industrial problems has been gradually increasing. Various sources of interference and random variations in external conditions make it increasingly difficult to model and analyze general industrial problems through analytical means. Particularly for high-dimensional modeling problems involving large and complex systems, traditional methods cannot achieve accurate results and are usually accompanied by insurmountable computational complexity. However, the use of experimental data for modeling and simulation can restrain some conditions in field environments and explore the information and rules behind the data. In real-world cases, given the limitations imposed by test conditions, collection methods and human or material resource requirements, the acquisition of data is often accompanied by high costs. For example, certain research involving one-time or special tests, such as bulb lifetime tests or nuclear tests, cannot obtain large amounts of data; in the former situation, the product is destroyed after the test, and in the latter, it is impossible to repeat the test. Therefore, there is important practical significance to expanding sample capacity in the above situations.

Manuscript received March 06, 2019; revised August 10, 2019. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant No. 2015ZCQ-LY-01.

Shaoqian Huang is a graduate with the Department of Mathematics, Beijing Forestry University, Beijing, 100083 China e-mail: (huangsq3614@foxmail.com).

Qi Zhou is a graduate with the Department of Mathematics, Beijing Jiaotong University, Beijing, 100044 China e-mail: (17121641@bjtu.edu.cn).

Hongqing Wang is a professor with the Department of Mathematics, Beijing Forestry University, Beijing, 100083 China e-mail: (wanghq@bjfu.edu.cn).

Corresponding author: Hongqing Wang (phone: 8615652936218).

+ These authors equally contributed to the work.

The importance of high sample capacity in research methods is self-evident. On the one hand, traditional statistical models can provide good approximations of parameters, and the corresponding conclusions can be theoretically guaranteed through the law of large numbers and the central limit theorem in the case of large sample size. On the other hand, machine learning algorithms, which exhibit better performance in terms of spatial search and function generalization than statistical models, also need large amounts of data when exploring the laws between inputs and outputs. In addition, the sample size influences the accuracy of results, as indicated in [1]. Therefore, sample expansion is necessary.

Traditional sample expansion techniques mainly include interpolation, noise injection, data sampling and virtual sample generation. An interpolation method was first established for sample expansion in [2]. The basic idea is to construct an approximation function through several discrete points on the interval and then to expand the sample capacity by solving the function value of the unknown point using an approximation function. As indicated in [3], [4], [5], researchers have designed noise injection methods to generate new samples by adding Gaussian noise. Data sampling uses a variety of sampling techniques to obtain more samples. However, the above three methods do not effectively use the original small amount of information contained in the low-capacity sample set. Therefore, it is difficult to establish a homogeneous high-capacity sample set using these methods. By contrast, virtual sample generation is a way of filling information intervals between a pair of original samples to produce new samples, which better utilizes the original information. Virtual sample expansion technologies mainly include distribution-based capacity expansion, domain-based prior knowledge expansion, perturbation-based sample capacity generation, and SVM-based sample expansion. Distribution-based capacity expansion first fits a probability distribution and then performs sample expansion through various sampling techniques. The bootstrap method, which obtains self-help samples with back-sampling, was introduced in [6] and [7]. However, because it is difficult to establish the probability distribution based on a low volume of samples, this method has low application value. Sample expansion technology based on prior domain knowledge is suitable for generating samples in professional fields. Prior knowledge, which is extremely valuable, is applied to obtain appropriate constraints, thereby transforming sample expansion into an optimization problem. Two forms of prior knowledge were given in [8]. However, due to the complexity of real-world conditions, prior knowledge is often difficult to obtain, which reduces the applicability of this method. In

[9], virtual sample generation, i.e., perturbation-based sample generation, was regarded as a virtual measurement process with measurement error that obeys a Gaussian distribution. However, due to its high computational complexity and the difficulty of generalizing and changing the distribution of measurement error, the information from the original low-capacity sample set us ignored. The basic idea of SVM-based sample expansion is to generate samples near the limit state function using an approximation method, thereby improving the classification ability of SVM. In [10], Basudhar et al. established an explicit design space decomposition method. However, this method has low efficiency and suffers from large errors on high-dimensional samples. In [11], Mao et al. performed sample generation by calculating the posterior probability. However, changing the dimensionality of the original sample and introducing new sample attributes are inappropriate for the expansion of high-dimensional samples. The prior knowledge obtained from a small training set was used to create virtual samples in [12]. This method strongly relies on prior knowledge obtained from a low capacity of original samples, but it has not been proposed as a formal procedure. In [13], Li et al. considered the domain of each attribute or variable in a low-capacity sample set to form a new virtual sample set. However, the expansion strategy is a trial-and-error process by nature, resulting in sensitivity to the type of sample set. The particle swarm optimization-based virtual sample generation (PSOVSG) approach was proposed to generate virtual samples in [14]; however, the shape parameter of the core TMIE function, which is not determined by a mature optimization method, strongly affects the algorithm's performance. In [15], an oversampling method based on the Euclidean distance was proposed, but this method is greatly influenced by dimension, and for high-dimensional random sample generation, the dependency between components is not considered. In [16], from the perspective of fault prediction with uncertainty quantification, recursive model resampling bootstrap (RMRB) was proposed. RMRB relies heavily on fuzzy inference, especially in the case of insufficient prior information, which makes it difficult to reasonably generate samples. In addition to the classic methods above, new approaches based on traditional statistical methods have been proposed. In [17], a new method called importance sampling was developed by employing a neural network. Although it is suitable for nonlinear data, the use of excessive neurons greatly increases the computational complexity. In [18], several sampling algorithms based on determinantal point processes (DPP) were introduced; however, memory cost is relatively high, and thus there is a high requirement for computer performance. In addition, machine learning algorithms are being applied to generate and expand samples. For example, in [19], an approach based on extreme learning machine (ELM) was proposed; however, for this kind of approach, some theories are immature, and the application effect is often related to the specific research problem, leading to poor generalizability. Moreover, for certain issues, such as face recognition [20] and chemistry analysis [21], some approaches have been proposed from the perspective of the professional field but cannot be extended to other fields.

In the real world, many phenomena are related; thus, many sample sets contain both independent variables and dependent variables. Based on the literature review above, no algorithms can effectively generate such samples. Traditional methods that fit the corresponding distribution function based on the original low-capacity samples may be inefficient and difficult to employ due to the low capacity and complex dependence structure of the sample components. Therefore, a sample expansion technology that is easy to implement programmatically and has strong generalizability for correlation analysis, regression analysis, and various statistical modeling methods involving independent variables and dependent variables must be developed. The difficulty of implementing sample expansion is how to maintain the relationship between the independent and dependent variables in the newly generated samples. In addition, for specific research problems in industrial production, the external conditions for sample generation change due to the monitoring fluctuations and time delay caused by the observations. Samples produced under the same external conditions are homogeneous, i.e., small in distance and similar in component evolution. Likewise, samples produced under different external conditions are theoretically different in terms of distance and trend. In this paper, for low-volume samples (X, Y) containing both independent variables $X = (X_1, X_2, \ldots, X_p)$ and a dependent variable Y, from the perspective of sample component evolution and the distance between a pair of samples, we propose an approach called DTDE to expand the sample capacity. Then, simulations are implemented based on various kinds of probability distributions to show that DTDE is effective and accurate for generating new samples from the original samples.

This paper is organized as follows. In section II, we establish the overall approach of DTDE and illustrate each important step. In section III, a simulation method is used to verify the algorithm's performance. In section IV, we consider two special cases to further analyze the algorithm. Finally, in section V, we conclude the study and provide research directions for future study.

II. Methods in the designed approach

In the following context, (X, Y) represents a sample, and (x, y) represents an observation of the corresponding sample.

The essential characteristics of a sample are determined by its probability distribution, which gives the probability that the sample takes a certain value when the joint distribution is fixed. Therefore, in terms of probability, for a pair of samples from the same distribution, such as $X_1 = (X_{11}, X_{12}\ldots, X_{1p})$ and $X_2 = (X_{21}, X_{22}, \ldots, X_{2p})$, the size relationship between components of X_1 is consistent with that of X_2, which means that sample components follow a similar trend and that the Euclidean distance between samples is relatively small. The sample itself is reflected in the value of each component and the relationship between the sample components, as shown in Fig.1.

In section II-A and section II-B, we will provide methods for measuring the distance and trend similarity between samples, respectively, to generate the independent variables of new samples. In section II-C, we present the method for generating the corresponding dependent variable of the new samples.

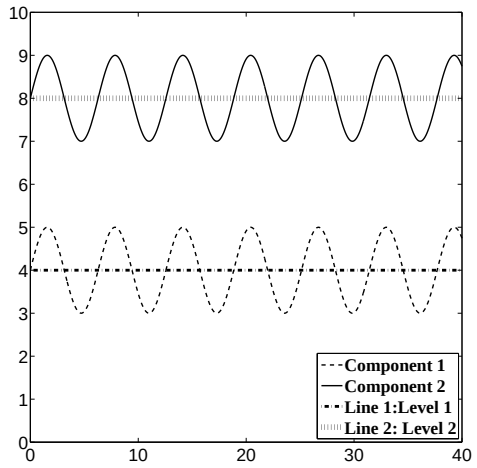


Fig. 1. Sample structure determined by component values and the dependence structure between components. Assume that a sample (X_1, X_2) contains two subelements, component 1 (X_1) and component 2 (X_2). Lines 1 and 2 represent the average level determined by the marginal distributions of X_1 and X_2, respectively. The two curves, which represent observations of component 1 (X_1) and component 2 (X_2), have the same variation trend, indicating strong dependence between them. If there is a probability distribution, the value of (X_1, X_2) is completely determined by the marginal distribution and the dependency.

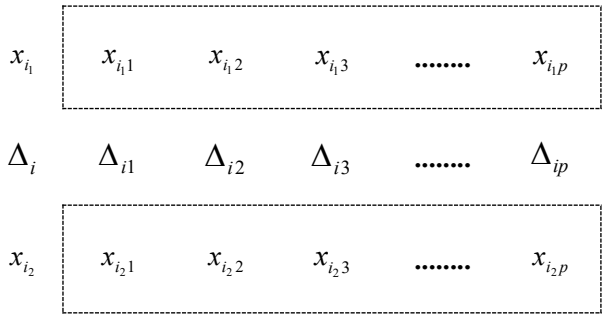


Fig. 2. Generation of independent variables of new samples $\triangle_i$. (i_1, i_2) is a pair of numbers among $1, \ldots, n$, and $\triangle_{ij}$ is the average of $x_{i_1,j}$ and $x_{i_2,j}$ $(j = 1, \ldots, p)$.

A. Measuring Distance based on Dynamic Clustering (DC)

Firstly, we use dynamic clustering [22] to classify the independent variables of the original samples so that the original sample set can be divided into several subsets that satisfy the requirement that the distance between each pair of samples in each subset is sufficiently small.

Dynamic clustering can dynamically adjust the classification by modifying cluster errors so that the clustering accuracy is relatively high. In addition, compared with other methods, this algorithm has lower computational complexity. The dynamic clustering algorithm is shown in Fig.12.

The clustering algorithm ensures that any pair of samples in the same class is close in distance from the perspective of independent variables of samples. Therefore, if new samples are generated in one class, then the generated samples will move to the center of each class in terms of the distance scale.

B. Matching Trend Based on Gray Correlation Analysis (GCA)

The clustering process considers only the distance between samples. In this subsection, we will consider the trend, i.e., the dependency between sample components. In terms of probability, the random samples generated by different joint distributions may be close in distance but totally different in dependence structure between sample components. Therefore, based on the clustering results, the samples in each subset should be further distinguished from the trend. The Gray correlation degree (GCD) [23] can reflect trend similarity between two columns of data. In the same class or subset, DC ensures a relatively small distance between any two samples. Furthermore, if two samples have a large GCD, under extreme conditions, the two samples can be considered to overlap with each other. Therefore, they can be regarded as the same and naturally homogeneous. Finally, for such a pair of samples with small distance and large GCD, we can average the independent variables of the original samples to generate the independent variables of the new samples. A diagram of the independent variable generation process is shown in Fig.2.

Thus, we should sort the samples in the same class according to a certain rule to guarantee that there is a large GCD between adjacent samples and perform averaging of independent variables between adjacent samples. The sample is no longer considered to be moved or reset once it enters the sequence during the sorting process. Therefore, the sorting rule ensures only that the GCD between any two adjacent samples in the sequence is relatively large. The sorting rule is as follows.

1) Pair the samples in accordance with the GCD and make the two most-correlated samples enter into the sequence. Record the samples as X_l and X_r.

2) Select the sample with the largest GCD with X_l, excluding samples that are already in place (only X_r at this time). Record this sample as X_l^*. Select the sample with the largest GCD with X_r, excluding samples that are already in place (only X_l at this time). Record this sample as X_r^*.

3) If the GCD between X_l^* and X_l is larger than that between X_r^* and X_r, we place X_l^* to the left first. Otherwise, we place X_r^* to the right first. At this point, the outermost samples of the sequence are X_l^* and X_r (or X_l and X_r^*).

4) Repeat 2) and 3) until all samples enter the sequence.

Example 1. To illustrate our rules, we present an example. We generate 5 random samples using an 8-dimensional normal distribution and apply our rules to sort them. The calculation process is shown in Steps 1 to 5.

Sample x_1: (1.118381, 1.099983, -0.07087, -0.10037, 0.333038, 1.115501, 1.731224, 0.882705)

Sample x_2: (1.813595, 3.130268, 1.855258, 0.470508, 0.73944, 2.820913, 2.653026, 2.628306)

Sample x_3: (2.279882, 1.244179, 1.697396, 0.411669, 0.550887, 1.158872, 2.01521, 1.603087)

Sample x_4: (1.165409, 2.200386, 1.074408, -0.23164, 0.903754, 0.022006, 1.731615, 2.748088)

Sample x_5: (0.587883, -0.09481, -0.94886, -0.81807, 0.209335, -2.222352, 0.119873, -0.20111)

Step 1. Calculate the GCDs between the 5 samples.

$$\mathbf{X} = \begin{pmatrix} 1 & 0.8486 & 0.8222 & 0.7564 & 0.5176 \\ 0.8486 & 1 & 0.8765 & 0.8453 & 0.5604 \\ 0.8222 & 0.8765 & 1 & 0.8101 & 0.6094 \\ 0.7564 & 0.8453 & 0.8101 & 1 & 0.8419 \\ 0.5176 & 0.5640 & 0.6094 & 0.8419 & 1 \end{pmatrix}$$

Step 2. The GCD between x_2 and x_3 is 0.8765, which is the largest among all GCDs. Therefore, x_2 and x_3 enter the sequence first. The sorted result is x_2x_3.

Step 3. Excluding x_3, x_1 has the largest GCD with x_2 (0.8486). Excluding x_2, x_1 has the largest GCD with x_3 (0.8222). Because 0.8486 is greater than 0.8222, x_1 enters the sequence and is set on the left of x_2. The sorted result is $x_1x_2x_3$.

Step 4. Excluding x_2 and x_3, x_4 has the largest GCD with x_1 (0.7564). Excluding x_1 and x_2, x_4 has the largest GCD with x_3 (0.8101). Because 0.8101 is greater than 0.7564, x_4 enters the sequence and is placed to the right of x_1. The sorted result is $x_1x_2x_3x_4$.

Step 5. Excluding x_2, x_3 and x_4, x_5 has the largest GCD with x_1 (0.5176). Excluding x_1, x_2 and x_3, x_5 has the largest GCD with x_4 (0.8419). Because 0.8419 is greater than 0.5176, x_5 enters the sequence and is placed to the right of x_4. The sorted result is $x_1x_2x_3x_4x_5$.

Remark 1. During clustering, samples from different distributions with small GCDs may be classified into the same class because of the small distance between them. When averaging only such samples, the newly generated samples do not actually couple with the original samples. To solve this problem, based on the distance scale, DTDE further considers the effect of component evolution, as illustrated in **Example 2.**, between a pair of samples, which guarantees that the generated samples are homogenous with the original ones.

Example 2. A situation is presented to illustrate the effect of GCA. Three samples are given as follows.

Original data: (1.3967, 1.4354, 0.4843, -0.0131, 0.3435, 1.0595, 1.6750, 0.9734)

Data with similar trend: (1.5725, 1.3643, 0.3448, -0.0183, 0.2998, 1.1219, 1.8082, 0.7717)

Data with small distance: (1.7371, 1.5632, 0.5710, 0.0842, 0.4086, 1.2772, 1.9655, 1.0344)

As shown in Fig.3, although the distance coincidence of Lines 1 and 3 is higher—i.e., the distance between the two polylines is smaller—Line 2 is closer to Line 1 in terms of the trend. As a result, even though the distance between them is larger than that between Lines 1 and 3, the trend of their components is more coincident. Therefore, Lines 1 and 2 can be averaged based on the DTDE method, but Lines 1 and 3 fail.

Remark 2. In this remark, we discuss a heuristic consideration. Our approach of averaging adjacent samples is based on such a logic thought; that is, two identical samples are averaged equal to the original sample. Two samples in the same class are close in terms of distance, and a large GCD indicates that the trend between them is similar. Under extreme conditions, the two samples overlap. As shown in Fig.4, the average of the two samples, which is represented by the broken curve, is equivalent to the average of two identical samples, which means the original samples themselves are generated. It is obvious that the new generated samples are in the class and do not destroy the structure of the original data.

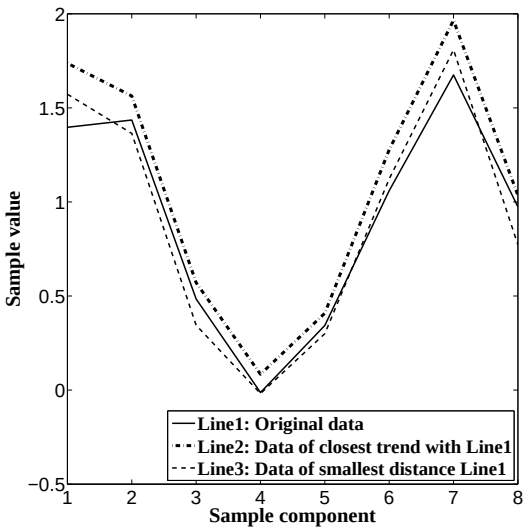


Fig. 3. Illustration of **Example 2**. Differentiate different data columns by different line types

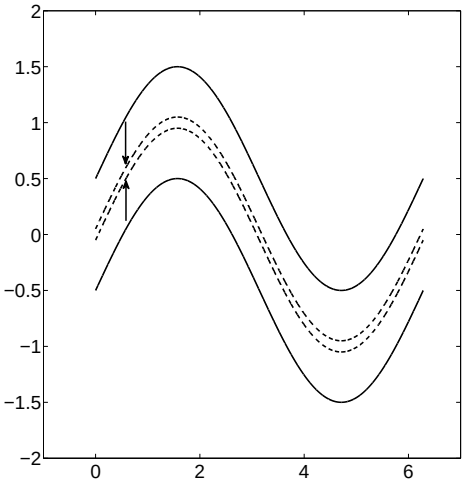


Fig. 4. Limiting case of adjacent samples.

C. Supplemental Samples Based on One-Dimensional Kernel Density Estimation (KDE)

A completely new sample includes both the independent variables and the dependent variable. In this subsection, we generate the corresponding dependent variable of the new samples (DVNSs); however, due to the complex nonlinear relationship between the independent variables and the dependent variable, if we generate the dependent variable by simply averaging the counterparts of adjacent samples, the function relationship between the dependent variables and the independent variable is ignored, which is clearly unreasonable and generates pseudo-samples. Therefore, the probability density is introduced, and DVNSs are filled based on the possibility of occurrence. By estimating the kernel density function [24] based on the dependent variable of the original samples, the average density of two dependent

variable samples whose corresponding independent variables are adjacent is calculated as the density of the corresponding DVNS. Finally, by calculating the inverse density function, we obtain the DVNSs. Thus, new samples containing independent variables and the dependent variable are obtained through the above procedure. The process is shown in Fig.13.

The probability density reflects the possibility that a random variable takes a certain value. Adjacent samples have a high GCD and a small distance, which means that they are relatively similar or even equivalent in an extreme case. The occurrence probability of similar samples is almost equivalent under similar external conditions. Thus, it is reasonable to estimate the possibility of the dependent variable appearing in a new sample based on the average density.

Remark 3. *A comparison of one-dimensional KDE and multidimensional KDE for directly generating samples.* Compared with one-dimensional KDE, multidimensional KDE, which estimates the multivariate distribution function or joint density function based on the original samples, has higher computational complexity. In addition, to ensure high accuracy, the estimation of multidimensional KDE needs a large number of samples to provide enough information, including the dependence structure between components and marginal distributions, which may violate the information of the original low-capacity samples. However, one-dimensional KDE can avoid the above drawbacks and obtain higher accuracy for low-capacity samples.

III. Simulation method and analysis

A. Simulation Method

First, we give the joint samples (X_i, Y_i) $(i = 1, \ldots, n)$ using the same distribution. In terms of probability, all samples from the same distribution concentrate so that they are close in distance. Therefore, we skip the process of dynamic clustering. In addition, the same distribution guarantees that the trends of all samples should be similar so that the GCD of the sample pairs is relatively large in theory. However, because of the uncertainty, sample pairs following different trends are likely to appear. As a result, we apply GCA to obtain more accurate results to select adjacent samples.

The algorithm proposed in this paper is suitable for low-capacity samples, and there is a functional relationship between the independent variables and the dependent variables in the sample. Therefore, the joint distribution of (X, Y) given in simulation must guarantee the functional relationship between Y and X, which implies a strong dependence structure between Y and X. In our simulation experiment, we use a linear dependency. For the joint distribution $(X_1, X_2, \ldots, X_n, Y)$, the Gaussian copula [25], which characterizes the linear relationship between each pair of variables, is given in (1), where R is the correlation matrix of the corresponding marginal normal distribution, including $X_i (i = 1, \ldots, p)$ and Y.

For our simulation, normal distributions are taken as the margins, and a Gaussian copula is used to describe the dependency. Then, a joint distribution is set up. The binary joint distribution is taken as an example. Different linear correlation coefficients show different geometric relationships, as shown in Fig. 7. In our simulation process, we assigned a large correlation coefficient between Y and X_i $(i = 1, \ldots, p)$ to ensure a strong linear dependence structure between them.

We use the neural network (NN) model to verify whether the generated samples are coupled with the original samples. Specifically, we can take the independent variables of the original samples as inputs and their dependent variables as the output to train the network. Then, we predict the dependent variables corresponding to independent variables of the new samples utilizing the trained network. Next, we compare the error between the predicted results with the new generated dependent variables, as shown in Fig.14. If the approach in this paper is effective, then the value of the dependent variable supplemented by our algorithm should be close to the predicted value of the neural network. In other words, the relative error will be small.

Remark 4. If we want to train a high-precision neural network, we need to provide enough samples. Therefore, we give a random sample set with a large sample size for simulation. Although the algorithm is based on low-capacity samples, to illustrate the effectiveness of the algorithm, we use a high-capacity random sample set to perform the simulation analysis. The complexity of the sorting rule is greatly increased as the sample size increases. To improve operational efficiency, we determine a threshold value to generate new samples. Specifically, given a threshold, for a pair of samples, if the GCD is greater than or equal to this threshold, then we use the pair to generate new samples; otherwise, the pair is not used to generate new samples.

B. Simulation Analysis

Assume $X = (X_1, X_2 \ldots, X_8)$. We randomly generate elements of a correlation matrix using a uniform distribution. To strengthen the correlation between X_i $(i = 1, \ldots, 8)$ and Y, we use $U(0.6, 1)$ to generate their correlation coefficients. The correlation coefficients between the independent variables are generated from $U(0, 1)$. Similarly, we assume that all marginal distributions obey a normal distribution, and the parameters of the mean and variance are generated from $U(0, 1)$.

First, the neural network is trained by 2000 original samples, and the relative prediction error of the neural network is calculated, as shown in Fig. 8(a). Most of the predicted values are very close to the true values, indicating that the neural network is of high quality and has good prediction accuracy. We calculate the relative error between the values predicted by the neural network and those supplemented by DTDE. The result is shown in Fig. 8(b). As shown in the figure, most of the relative errors are concentrated around 0, and only a few relative errors are large. The mean relative error is 0.0126, which indicates that the DTDE algorithm is significantly effective.

In addition, we performed simulation analysis in three cases of weak linear correlation, medium correlation and strong linear correlation, respectively. As shown in Fig.11, we give the relative error graphs in three cases. We find that the accuracy of the algorithm increases as the correlation strengthens, which shows that the algorithm become more effective on sample expansion when the dependent variables and independent variable have a strong dependence structure. Therefore, the results are consistent with our initial idea of designing DTDE.

$$Copula = \int_{-\infty}^{\Phi^{-1}(u_1)} \dots \int_{-\infty}^{\Phi^{-1}(u_n)} \frac{1}{|R|^{\frac{1}{2}}} exp\{-\frac{1}{2} s^T (R^{-1} - I)s\} ds_1 \dots ds_n \quad (1)$$

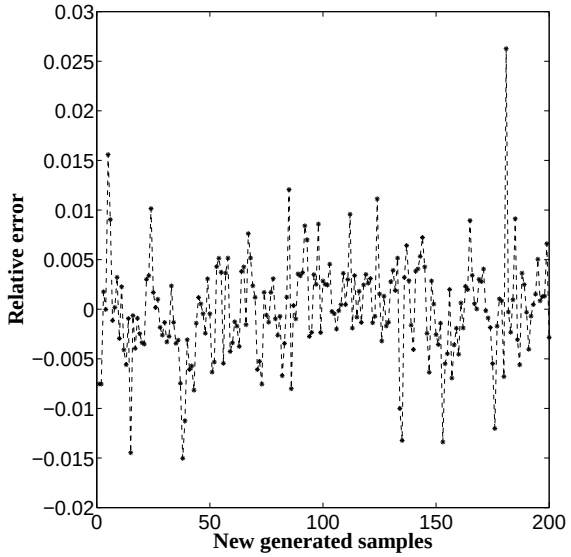


Fig. 5. Relative error of DTDE under a reverse order; the mean relative error is 0.0037.

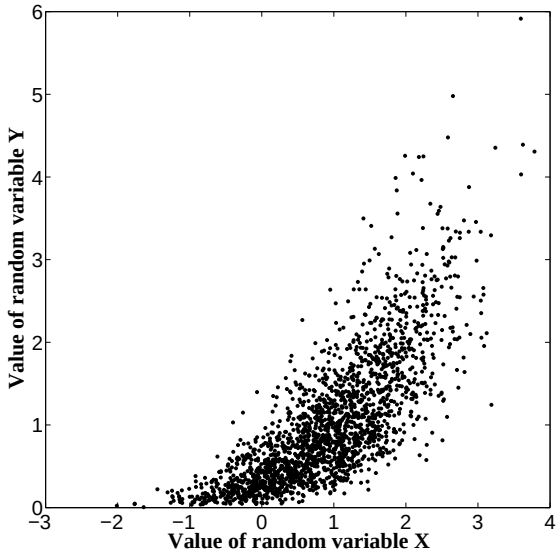


Fig. 6. Geometric nonlinear relationship between Y and X_i ($i \in (1, \dots, p)$) when $\rho = 0.99$.

IV. Discussion

For the algorithm proposed in this paper, we discuss two special cases for further analysis.

A. Case 1. Changing the Order of DC and GCA.

In this subsection, we keep the basic methods unchanged. The original samples are first classified by the GCD and then sorted under the rule with the distance as a measure to generate new samples. Since we use samples from the same distribution, the dependence structure between sample components is identical. Therefore, based on the measurement of GCD, we treat the identically distributed samples as a class. Next, we sort the samples according to distance using the sorting rules stated in section II-B. For computational convenience, a distance threshold is also taken. When the distance between two samples is smaller than this threshold, the samples are considered as a pair of adjacent samples; otherwise, they are not used to generate new samples. The result of the relative error for the simulation is shown in Fig.5.

The order change has no significant effect on the similarity measurement between samples due to the high prediction precision shown in Fig.5. Therefore, our proposed algorithm essentially considers the similarity between samples, which has nothing to do with the order of measurement, i.e., distance and trend.

B. Case 2. Strong Nonlinear Correlation between Dependent Variables and the Independent Variable.

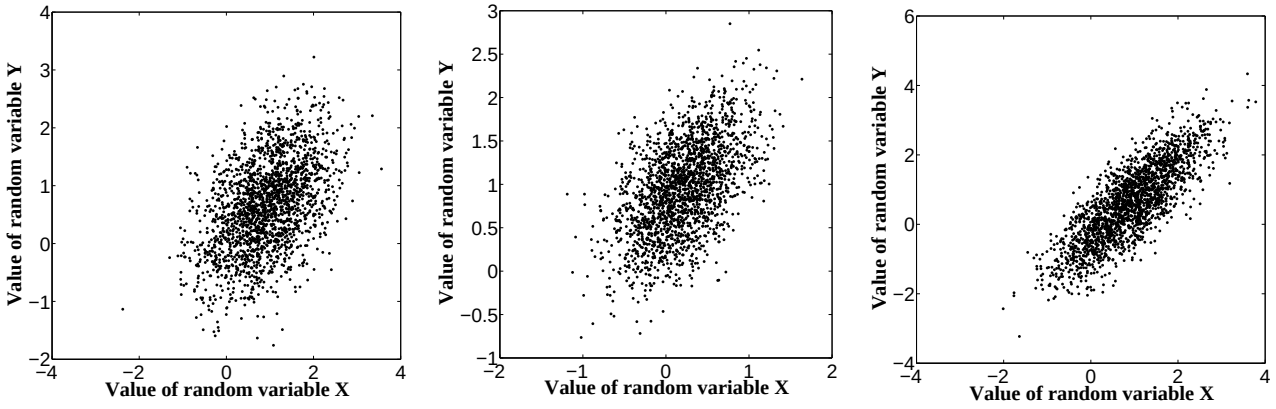
In this subsection, we consider a strongly geometric nonlinear relationship between the dependent variables and independent variable. The marginal distribution of the independent variable remains a normal distribution. However, the margin of the dependent variable transforms into $Gamma(2, 0.5)$, which results in the geometric nonlinear correlation between Y and the single independent variable. In this case, we found that the algorithm still achieves high precision, with a mean relative error of only 0.0356. In Fig.6, we present one of the geometric relationships.

In addition, we perform simulation analysis for several strong geometric nonlinear relationships by changing the distribution of Y, as shown in Fig.9. In terms of the mean relative errors, all results have high accuracy, as shown in Fig.10. The results show that our proposed algorithm is also applicable to complex nonlinear relationships and is generalizable to more complex relationships.

V. Conclusions

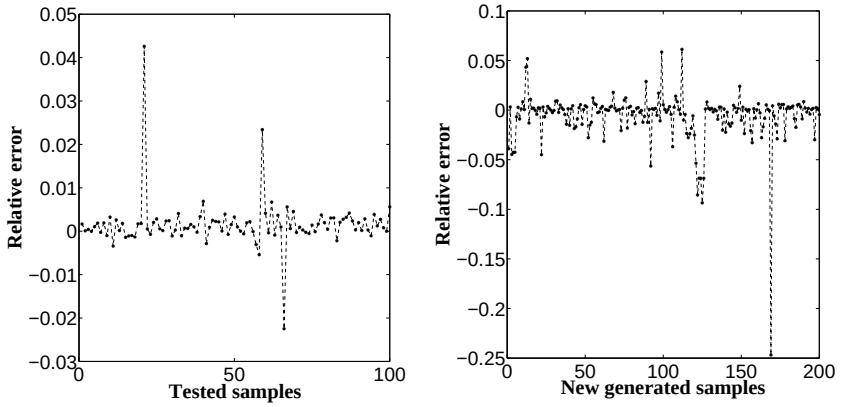
In this paper, considering distance and trend double factors, an approach for sample expansion technology is established. The algorithm can both avoid the estimation of a high-dimensional probability distribution and fully utilize the information from the original low-capacity samples to generate new samples. In addition, this technology produces high-capacity homogeneous samples and can expand the application range of some statistical methods that have higher sample size requirements. Finally, the simulation analysis proves the validity and accuracy of the proposed algorithm.

A topic to address in future studies is information expansion (IE). We hope to propose an algorithm that can enhance the coverage rate of the original information with continuously derived new samples. If the original low-capacity samples are concentrated locally, then the homogeneous new samples will better reflect the local information. However, the information outside the local area cannot be reflected. Therefore, IE has theoretical research value. In addition, we present an operational sample sorting rule. In the future, we



(a) (b) (c)

Fig. 7. Geometric linear relationship between Y and X_i $(i \in (1, \ldots, p))$ under different correlation coefficients ρ and normal margins. (a)$\rho = 0.33$. (b)$\rho = 0.66$. (c) $\rho = 0.99$.



(a) (b)

Fig. 8. (a) Relative error of NN; the mean relative error is 0.0027. (b) Relative error of DTDE; the mean relative error is 0.0126.

hope to strictly define the system GCD so that the rule can maximize the system GCD under this definition.

REFERENCES

[1] D. C. Li, Y. S. Lin, Using virtual sample generation to build up management knowledge in the early manufacturing stages, European Journal of Operational Research, vol. 175, no. 1, pp413-434, 2006

[2] R. Mao, H. Zhu, L. Zhang, A. Chen, A new method to assist small data set neural network learning, in: Sixth International Conference on Intelligent Systems Design and Applications, vol. 1, pp17–22, 2006

[3] R. M. Balabin, S. V. Smirnov, Interpolation and extrapolation problems of multivariate regression in analytical chemistry: benchmarking the robustness on near-infrared (NIR) spectroscopy data, The Analyst, no. 7, pp1604-1610, 2012

[4] P. Niyogi, F. Girosi, T. Poggio, Incorporating prior information in machine learning by creating virtual examples, Proceedings of the IEEE, vol. 86, no. 11, pp2196–2209, 1998

[5] T. I. Tsai, D. C. Li, Utilize bootstrap in small data set learning for pilot run modeling of manufacturing systems, Expert Systems with Applications, vol. 35, no. 3, pp1293-1300, 2008

[6] J. Yang, X. Yu, Z. Q. Xie, J. P. Zhang, A novel virtual sample generation method based on gaussian distribution, Knowledge-Based Systems, vol. 24, no. 6, pp740-748, 2011

[7] A. Basudhar, S. Missoum, Adaptive explicit decision functions for probabilistic design and optimization using support vector machines, Computers & Structures, vol. 86, no. 19-20, pp1904–1917, 2008

[8] G. Bloch, F. Lauer, G. Colin, Y. Chamaillard, Support vector regression from simulation data and few experimental samples, Information Sciences, vol. 178, no. 20, pp3813-3827, 2008

[9] G. Y. Chao, T. I. Tsai, T. J. Lu, H. C. Hsu, B. Y. Bao, W. Y. Wu, M. T. Lin, T. L. Lu, A new approach to prediction of radiotherapy of bladder cancer cells in small dataset analysis, Expert Systems with Applications, vol. 38, no. 7, pp7963–7969, 2011.

[10] Z. S. Chen, B. Zhu, Y. L. He, L. A. Yu, A PSO based virtual sample generation method for small sample sets: Applications to regression datasets, Engineering Applications of Artificial Intelligence, vol. 59, pp236–243, 2017

[11] A. di Bella, L. Fortuna, S. Graziani, G. Napoli, M. G. Xibilia, Development of a soft sensor for a thermal cracking unit using a small experimental data set, in: 2007 IEEE International Symposium on Intelligent Signal Processing, pp1-6, 2007

[12] K. I. J. Ho, C. S. Leung, J. Sum, Convergence and Objective Functions of Some Fault/Noise-Injection-Based Online Learning Algorithms for RBF Networks, IEEE Transactions on Neural Networks, vol. 21, no. 6, pp.938–947, 2010

[13] A. Kuhl, L. Krüger, C. Wöhler, U. Kreßel, Training of classifiers using virtual samples only, in: ICPR, vol. 3, pp418–421, 2004

[14] D. C. Li, L. S. Chen, Y. S. Lin, Using functional virtual population as assistance to learn scheduling knowledge in dynamic manufacturing environments, International Journal of Production Research, vol. 41, no. 17, pp4011-4024, 2003

[15] K. Savetratanakaree, K. Sookhanaphibarn, S. Intakosum, R. Thawonmas, Borderline over-sampling in feature space for learning algorithms in imbalanced data Environments, IAENG International Journal of Computer Science, vol. 43, no. 3, pp363-373, 2016

[16] D. P. Li, Failure prognosis with uncertain estimation based on recursive models re-sampling bootstrap and ANFIS, IAENG International Journal of Computer Science, vol. 43, no. 2, pp253–262, 2016

[17] Q. Zheng, M. Zwicker, Learning to importance sample in primary sample space, arXiv preprint arXiv:1808.07840.

[18] N. Tremblay, S. Barthelme, P. O. Amblard, Optimized algorithms to sample determinantal point processes, arXiv preprint arXiv:1802.08471.

[19] Y. L. He, P. J. Wang, M. Q. Zhang, Q. X. Zhu, Y. Xu, A novel and effective nonlinear interpolation virtual sample generation method for enhancing energy prediction and analysis on small data problem: A case study of Ethylene industry, Energy, vol. 147, pp418–427, 2018

[20] J. Zeng, X. Zhao, Y. Zhai, J. Gan, Z. Lin, C. Qin, A novel expanding sample method for single traning sample face recognition, in: 2017

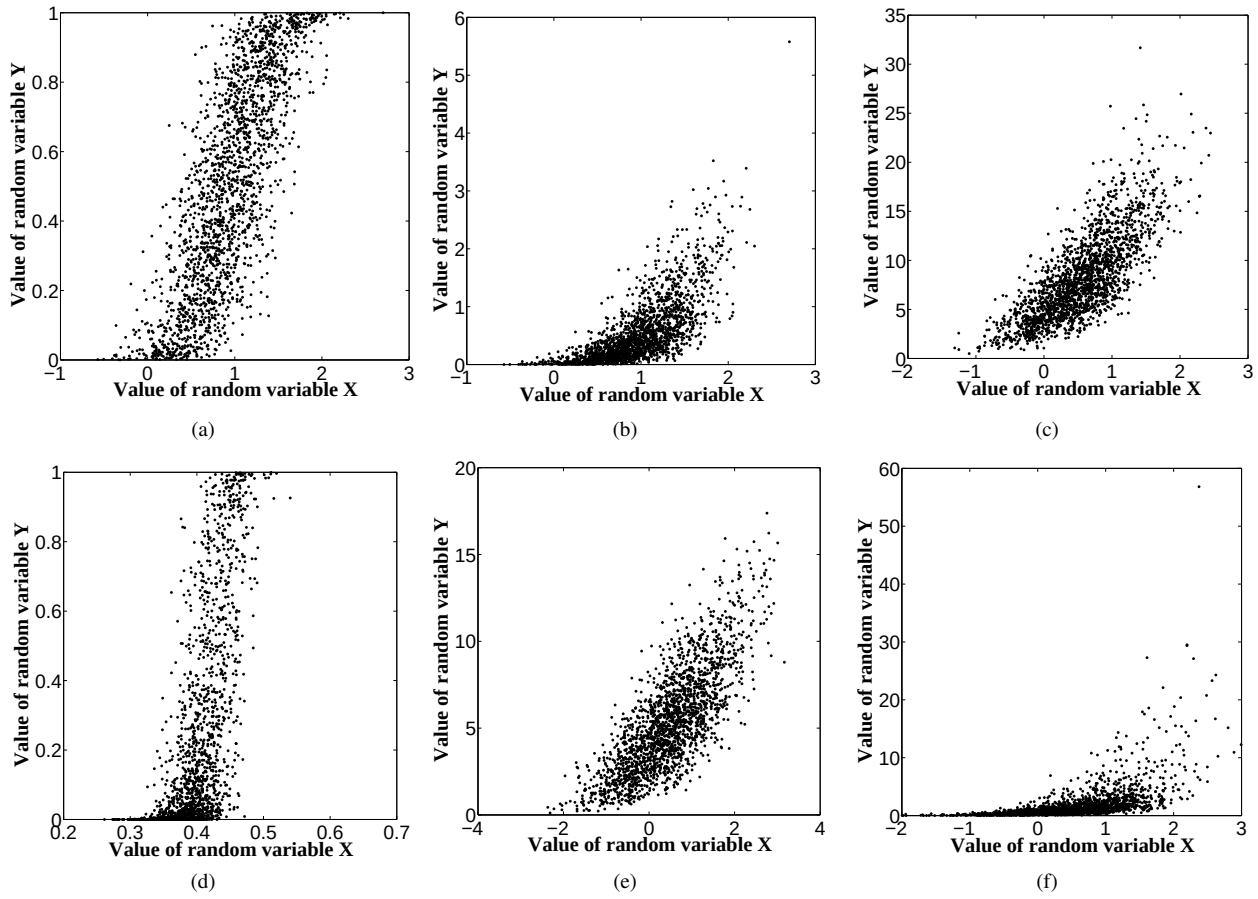


Fig. 9. Several functional relationships between Y and X_i $(i \in (1, \ldots, p))$ when $\rho = 0.99$. (a) $Y \sim U(0,1)$ and $X_i \sim N(\mu_1, \sigma_1)$. (b) $Y \sim exp(2)$ and $X_i \sim N(\mu_2, \sigma_2)$. (c) $Y \sim \chi(8)$ and $X_i \sim N(\mu_3, \sigma_3)$. (d) $Y \sim Beta(0.4, 0.6)$ and $X_i \sim N(\mu_4, \sigma_4)$. (e) $Y \sim Weibull(2, 6)$ and $X_i \sim N(\mu_5, \sigma_5)$. (f) $Y \sim F(3, 5)$ and $X_i \sim N(\mu_6, \sigma_6)$. μ_j, σ_j $(j = 1, ..., 6)$ come from $U(0, 1)$

International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), pp33–37, 2017

[21] Y. H. Liu, K. F. Huang, D. C. Lee, Precise and accurate boron and lithium isotopic determinations for small sample-size geological materials by MC-ICP-MS, Journal of Analytical Atomic Spectrometry, vol. 33, no. 5, pp846–855, 2018

[22] M. E. Celebi, H. A. Kingravi, P. A. Vela, A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Systems with Applications, vol. 40, no.1, pp 200–210, 2013

[23] R. Congjun, Z. Yong, Z. Zhongcheng, Multi-attribute auction method based on grey relational degree of hybrid sequences, Journal of Grey System, vol. 21, no.2, pp175-184, 2009

[24] B. W. Silverman, Density estimation for statistics and data analysis, Routledge, 2018.

[25] Y.Fang, A Bayesian Approach to Inference and Prediction for Spatially Correlated Count Data Based on Gaussian Copula Model, IAENG International Journal of Applied Mathematics, vol. 44, no. 3, pp126-133, 2014

Shaoqian Huang He was born in HeBei, China, in 1994. He received a bachelor's degree in statistics from North China University of Science and Technology. His main research interests are statistics, copula theory, change point analysis and simulation.

Qi Zhou He was born in HeBei, China, in 1993. He received a bachelor's degree in statistics from North China University of Science and Technology. His main research interests are reliability, signature theory and application, statistics and simulation.

Hongqing Wang He was born in ShanXi, China, in 1971. He received a Ph.D in Science from Yamagata University of Japan. He is a professor at the College of Science, Beijing Forestry University, China. His main research interests are statistical algorithms.

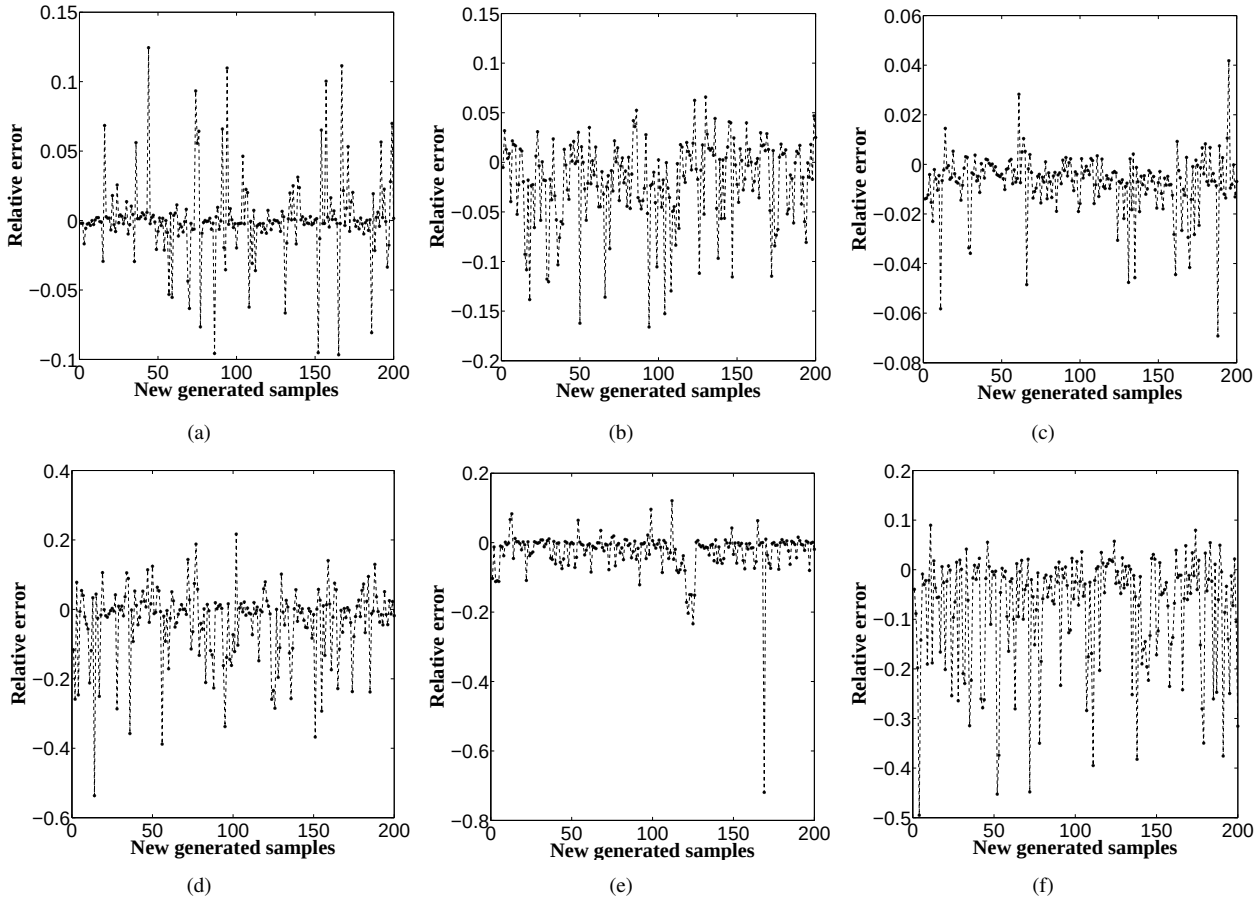


Fig. 10. Relative error of DTDE under several functional relationships between Y and X_i $(i \in (1, \ldots, p))$ when $\rho \in (0.6, 1)$. (a) $Y \sim U(0,1)$ and $X_i \sim N(\mu_1, \sigma_1)$. (b) $Y \sim exp(2)$ and $X_i \sim N(\mu_2, \sigma_2)$. (c) $Y \sim \chi(8)$ and $X_i \sim N(\mu_3, \sigma_3)$. (d) $Y \sim Beta(0.4, 0.6)$ and $X_i \sim N(\mu_4, \sigma_4)$. (e) $Y \sim Weibull(2, 6)$ and $X_i \sim N(\mu_5, \sigma_5)$. (f) $Y \sim F(3, 5)$ and $X_i \sim N(\mu_6, \sigma_6)$. Where μ_j, σ_j $(j = 1, ..., 6)$ come from $U(0, 1)$.

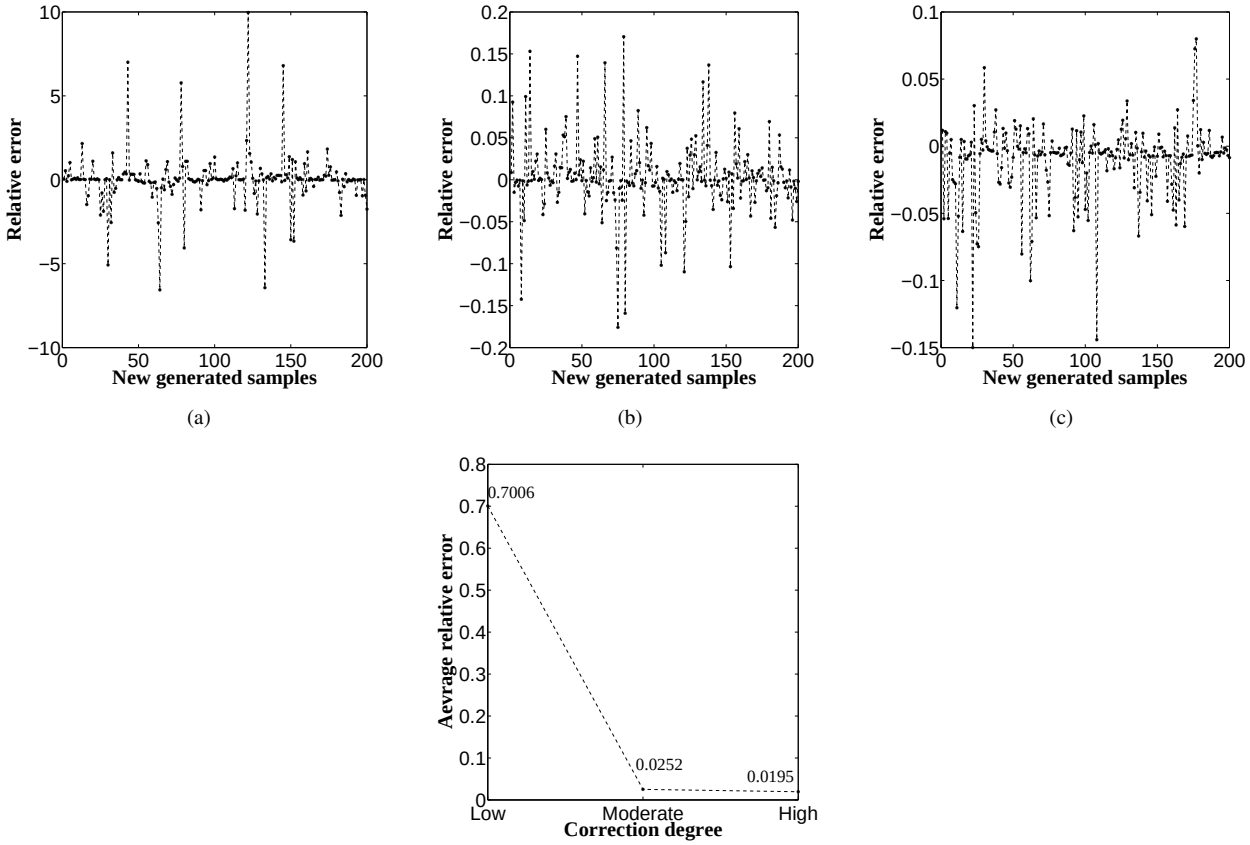


Fig. 11. Relative error of DTDE under several linear relationships controlled by the range of ρ between Y and X_i $(i \in (1, \ldots, p))$.(a)$\rho \in (0, \frac{1}{3})$.$Y \sim N(\mu_1, \sigma_1)$ and $X_i \sim N(\mu_2, \sigma_2)$. (b) $\rho \in (\frac{1}{3}, \frac{2}{3})$. $Y \sim N(\mu_3, \sigma_3)$ and $X_i \sim N(\mu_4, \sigma_4)$. (c) $\rho \in (\frac{2}{3}, 1)$. $Y \sim N(\mu_5, \sigma_5)$ and $X_i \sim N(\mu_6, \sigma_6)$. Where μ_j, σ_j $(j = 1, ..., 6)$ come from $U(0, 1)$.

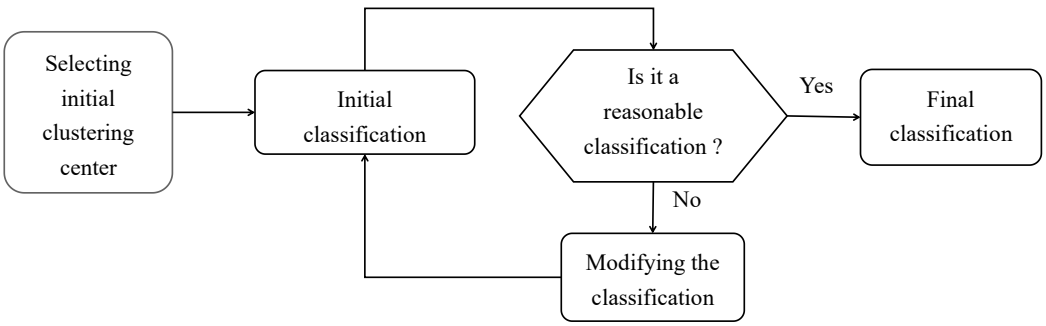


Fig. 12. Process of dynamic clustering.

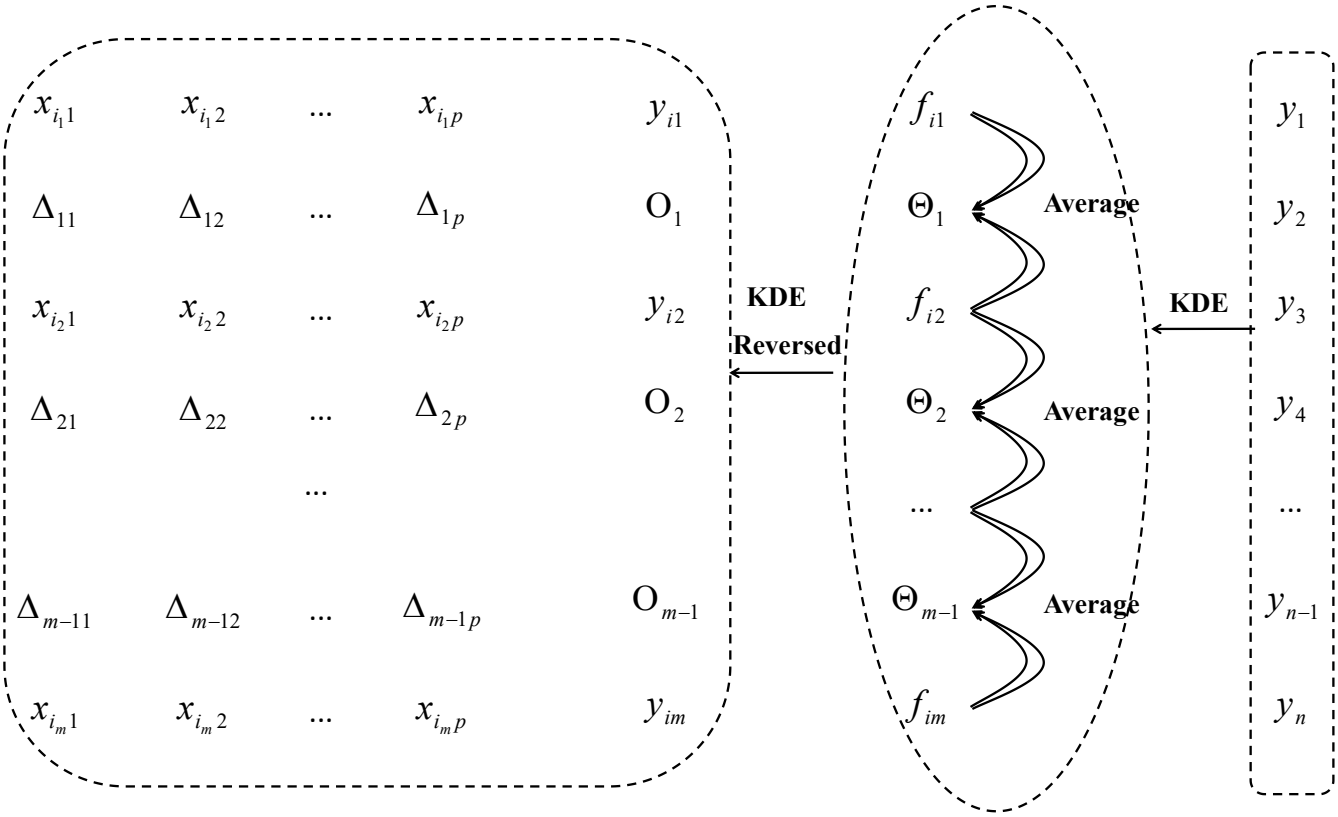


Fig. 13. Combination of the independent variables and dependent variable of new generated samples. $y_1, \ldots, y_n$ are the original samples of the dependent variable, and $(x_{i_1}, y_{i1}), \ldots, (x_{i_m}, y_{im})$ are adjacent samples in sense of DTDE for the independent variables. After estimation by KDE, the kernel density function of $y_1, \ldots, y_n$ is f. Let $\Theta_j = 1/2(f_{ij} + f_{i(j+1)})$ $(j = 1, ..., m-1)$ and $\bigcirc_j = f^{-1}(\Theta_j)$ $(j = 1, \ldots, m-1)$. Then, $(\triangle_k, \bigcirc_k)$ $(k = 1, \ldots, m-1)$ are the newly generated samples, where $f_{ij} = f(y_{ij})$.

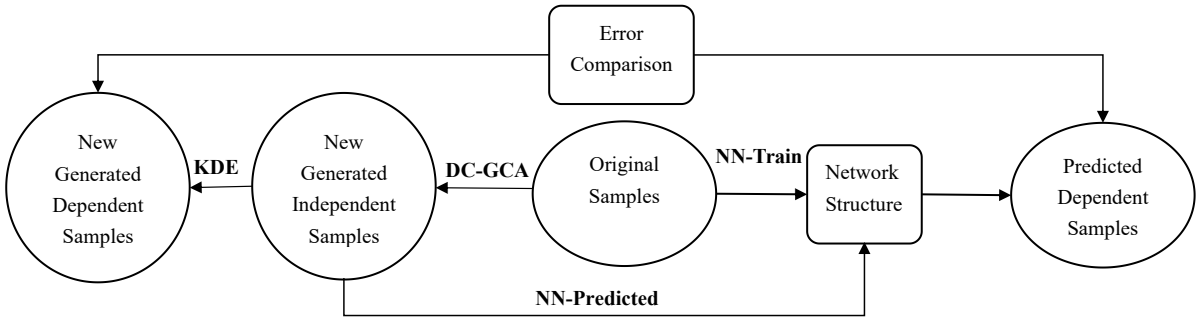


Fig. 14. Procedure of coupling verification between the newly generated samples and the original samples.