International Journal of Business Marketing and Management (IJBMM)
Volume 8 Issue 3 May-June 2023, P.P. 53-60
ISSN: 2456-4559
www.ijbmm.com

Characteristics and simulation analysis of nonlinear correlation coefficient on manifold

Hongcheng Tao [a] , Peibiao Zhao*[a]

[a] School of Mathematics & Statistics, Nanjing University of Science and Technology, China Email:
yysf6131@163.com; pbzhao@njust.edu.cn

Abstract: In this paper, by using the basic method of differential geometry, combined with the optimization theory and the basic technique of data analysis, the definition, basic properties and statistical characteristics of nonlinear correlation coefficients on manifolds are studied and given, test the rationality and validity of the nonlinear correlation coefficient defined in this paper. Therefore, the study of this paper has certain theoretical value and potential practical significance.

Keywords: nonlinear correlation coefficient, manifold, statistical characteristics

I. Introduction

In recent years, manifold learning, as a more and more popular research direction, has received widely attention by many experts and scholars. Manifold learning, roughly speaking, is a set of sample data points of observable points in a high-dimensional Euclidean space, which is understood to lie on a low-dimensional manifold embedded in a high-dimensional Euclidean space, then, the data points on the high-dimensional space are projected to the low-dimensional space by a suitable dimension reduction method, and the basic characteristics of the data points on the original manifold are preserved as much as possible. In this way, we can eliminate the superfluous information, realize dimensionality reduction, and study the inherent law and basic characteristics of data easily. The correlation coefficient is an important inherent law of data. Considering that manifolds are themselves generalizations of curves and surfaces in higher dimensional space, there are often strong Linear independence between their variables, therefore, it is necessary to study the nonlinear correlation coefficients between variables on manifolds systematically. As far as we know, the research on the correlation coefficient on manifold is still in the initial stage, and the related research results are few.

So some scholars began to study the correlation coefficient, initially with Ting Wang (2011) [1] who defines a new correlation coefficient, it can be used not only in nonlinear cases, but also in linear cases, and can be regarded as a general correlation coefficient. Ruiming Liu (2013)[2] proposed a tracking framework based on a combination of template matching and Emmerich Kálmán prediction. The projection coefficient obtained from principal component analysis is used as template, and the nonlinear correlation coefficient is used to measure the matching degree. Zhou Hongfang (2021)[3] studies the feature selection process in machine learning, and proposes a feature selection method based on mutual information and correlation coefficient, the absolute value of the correlation coefficient between the two features is used as the weight of the redundancy expressed by mutual information in the evaluation criteria. Experimental results show that the method can effectively remove the redundant information in the evaluation criteria.Antonella Plaia(2021)[4] considered the importance of the commutative elements belonging to the top (or bottom) in ranking (position weight) , studied the consistency of ranking with relation, and proposed the position-weighted rank correlation coefficient, used to compare rankings and relationships.Dong Xiaomeng (2010) [5]proposed a generalized correlation coefficient which can describe the correlation degree between variables or vectors by using the fourth-order moment method in order to solve the linear correlation problem of correlation coefficient.

In manifold learning, Chen Changyou(2010)[6] believed that existing algorithms such as ISOMAP try to ensure that data points are equidistant embedded on manifolds, but in practical applications, such as face and gait recognition, the recognition effect is not so satisfactory, he proposed a two-stage Bernhard Riemann manifold distance approximate projection (TRIMAP) algorithm based on tensor, which can quickly calculate the approximate optimal projection for a given tensor data set, at the same time, the experiments were carried out based on human gait database and face recognition database. The experimental results show that the recognition ability of TRIMAP algorithm is better than other common algorithms. Lee S.-M (2007) [7]transformed the probability distribution by mapping it to a hypersphere through isometric transformation. The sphere constructs a Bernhard Riemann manifold with a simple geodetic distance. Then, a Freychet mean is estimated on a Bernhard Riemann manifold for principal component analysis in a plane tangent to the mean. I. Ya. Savka(2010) [8]studied the non-local two-point boundary conditions for secondorder partial differential equation with constant linear correlation coefficients, the unique solvable condition of the problem on the scale of Sobolev space is established, and the metric theorem of the lower estimate of the small denominator on the linear manifold is proved.

This paper mainly studies how to measure the nonlinear correlation between variables on a manifold, summarizes and classifies the common manifold dimensionality reduction methods, and selects the suitable dimensionality reduction methods. The main reasons for dimensionality reduction are: 1. The Galway number of the original observation space sample will be greatly redundant, if the sample does not deal with the classifier is prone to "Dimension disaster" 2. The non-linear correlation coefficient proposed in this paper is mainly suitable for the random variables with strong non-linear relationship, and is not good for the random variables with strong linear relationship, therefore, it is necessary to adopt appropriate dimensionality reduction methods to preserve the nonlinear structure of the original manifold and find the random variables with strong nonlinear relations. 3. Manifold is different from Euclidean space, and its properties are very different from those of Euclidean space, so it can not be calculated directly by coefficient on Euclidean space.

The paper is structured as follows. In the second part, we introduce the basic definition of non-linear correlation coefficient $SEVP$, prove and give some basic properties of the corresponding non-linear correlation coefficient $SEVP$ .The third part is the simulation analysis, the results show that the proposed definition is reasonable and effective. In the last section, we summarize the theories and models covered in this paper and give directions for further in-depth research on nonlinear correlation coefficient on manifold.

II. The nonlinear correlation coefficient on manifold

At present, there are few researches on the nonlinear correlation coefficient, and some commonly used rank correlation coefficients are not suitable for our research needs, we need a kind of coefficient which is fully suitable for measuring the nonlinear relationship between variables, using the variance decomposition formula:

$$\operatorname{Var}(Y)=\operatorname{Var}\left(E\left(Y|\ X\right)\right)+E\left(\operatorname{Var}\left(Y|\ X\right)\right) \tag{2.1}$$

We define a new nonlinear correlation coefficient and prove some of its basic properties.

Definition 1. Let X and Y be random vectors on the probability space (Ω, F, P), and the nonlinear correlation coefficient $SEVP$ be in the following form:

$$SEVP=\frac{\operatorname{Var}\left(E\left(Y|\ X\right)\right)}{\operatorname{Var}\left(Y\right)}-p_{xy}^{2} \tag{2.2}$$

Theorem 2.1. $SEVP \in [0,1]$

Proof. First prove $SEVP..0$, because the coefficient has translation invariance, so suppose $E(X)=0, E(Y)=0, f(x,y)$ be its joint probability density, $f^{x}(x)$ is an edge probability density function for the random variable X ,to prove $SEVP..0$, just need proof:

$$\frac{\operatorname{Var}\left(E\left(Y|\ X\right)\right)}{\operatorname{Var}\left(Y\right)}..p_{xy}^{2} \tag{2.3}$$

The left expansion of the formula (2.3) gives:

$$\frac{\operatorname{Var}\left(E\left(Y|\ X\right)\right)}{\operatorname{Var}\left(Y\right)}=\frac{E\left(E^{2}\left(Y|\ X\right)\right)-(E\left(Y\right))^{2}}{\operatorname{Var}\left(Y\right)}=\frac{\int\left(g^{y/x}\left(x\right)\right)^{2} f^{x}\left(x\right)dx}{\operatorname{Var}\left(Y\right)} \tag{2.4}$$

Among them, $g^{y/x}\left(x\right)=E\left(Y|\ X=x\right)=\dfrac{\int yf\left(x,y\right)dy}{f^{x}\left(x\right)}$

And then the deformation of the right can be:

$$\begin{aligned} p_{xy}^2 &= \frac{(E(XY)-E(X)E(Y))^2}{\operatorname{Var}(X)\operatorname{Var}(Y)} \\ &= \frac{E^2(XY)}{\operatorname{Var}(X)\operatorname{Var}(Y)} \\ &= \frac{E^2(XY)}{\operatorname{Var}(X)\operatorname{Var}(Y)} \\ &= \frac{\left(\iint xyf(x,y)\,dxdy\right)^2}{\operatorname{Var}(X)\operatorname{Var}(Y)} \\ &= \frac{\iint (xyf(x,y)\,dxdy)^2}{\int x^2 f_x(x)dx\operatorname{Var}(Y)} \end{aligned} \tag{2.5}$$

To prove that the left is greater than the right, just prove:

$$\int x^2 f_x(x)dx\int\left(g^{y/x}(x)\right)^2 f^x(x)dx..\left(\iint xyf(x,y)dxdy\right)^2 \tag{2.6}$$

And because:

$$g^{\frac{y}{x}}(x)=E(Y|\ X=x)=\frac{\int yf(x,y)dy}{f^x(x)} \tag{2.7}$$

The substitution of the formula (2.7) into the formula (2.6), whose left-hand form can be converted to:

$$\int x^2 f_x(x)dx\int\frac{\left(\int yf(x,y)dy\right)^2}{f^x(x)}dx \tag{2.8}$$

Using the integral form of Cauchy's Cauchy-Schwarz inequality, we get:

$$\int x^2 f_x(x)dx\int\frac{\left(\int yf(x,y)dy\right)^2}{f^x(x)}dx..\left(\iint xyf(x,y)dxdy\right)^2 \tag{2.9}$$

So we have $SEVP..0$, because $\frac{\operatorname{Var}(E(Y|\ X))}{\operatorname{Var}(Y)}$ and p_{xy}^2 are all between 0 and 1,so the maximum value of the correlation coefficient of $SEVP$ must not exceed 1, hence $SEVP\in[0,1]$.

Consider a smooth m dimensional manifold whose tangent plane should in fact be m dimensional, the above m principal curvature is arranged from small to large in absolute value, reflecting the high to low degree of linearization of the manifold along the m principal direction at that point, in fact, it is a space where the manifold is more linearized at that point, and the space that the linear space is stretched by the tangent vector m at that point, that is, the tangent space of the manifold at that point, it is shown that the tangent space of a manifold is highly linearized, and the projection onto the tangent space contains more linear structures on the original manifold.

But our goal is to preserve the nonlinear structure of the manifold as much as possible, since the projection onto the tangent space of the manifold will contain more linear structure, then the projection onto the orthogonal complement of the tangent space may contain more nonlinear structures. In theory, if we can find a linear space that is perpendicular to the tangent space of all points on the manifold, so, by projecting onto that space, we can preserve as many nonlinear structures as possible. Intuitively, that is, to have as large an angle as possible with all locally tangent spaces.

So now the problem comes down to a given set of sample points $X=(X_1,X_2,X_3,\ldots X_d),X$ on a m dimensional manifold, how to find the tangent space angle with all points and the largest one k dimensional linear space, for each point x_i there is a tangent space, may be remembered as span (U_i), where $U_i^T U_i = I_m$, the base of the linear space we need to solve for is V_i, where $V_i^T V_i = I_k$, the optimal problem is as follows:

$$\min_{V\in \mathrm{R}^{d\times k},V^T}\sum_{V=I}^{n}\left\|V^T U_i\right\|_F^2 = \min_{V\in \mathrm{R}^{d\times k},V^T V=I}\sum_{i=1}^{n}\mathrm{tr}\left(V^T U_i U_i^T V\right) \qquad (2.10)$$

where $\mathrm{tr}\,(\,)$ represents the sum of the diagonal elements of the matrix.

Then the original problem is transformed into the eigenvector $t_1,\cdots,t_k$, which corresponds to the minimum eigenvalue k of the matrix $\sum_{i=1}^{n} V_i V_i^T$, the k eigenvector is the base of the linear space with the largest angle between the solution and all tangent spaces.

Now we know that a m dimensional manifold is embedded in a d dimensional space by an unknown function $f(\tau),\tau\in R^m$, where $m<d$, a known set of sample points $X=\left(X_1,X_2,X_3,\ldots,X_d\right),X_i\in R^{d\times 1}$, has:

$$X_i = f(\tau_i), i=1,\cdots,n \qquad (2.11)$$

$\tau_i\in R^{m\times 1}$ is the result of the dimensionality reduction of X_i. the goal of nonlinear dimensionality reduction is to reconstruct τ_i corresponding to X_i without explicitly building the f function. Assuming that f is smooth enough, do a Taylor expansion at a given τ :

$$f\left(\bar{\tau}\right) = f(\tau)+J_f(\tau)\cdot\left(\bar{\tau}-\tau\right)+O\left(\left\|\bar{\tau}-\tau\right\|_2^2\right) \qquad (2.12)$$

Here $J_f(\tau)\in R^{d\times m}$ is the Jacobi matrix of f at τ, if remember:

$$f(\tau)=\begin{pmatrix} f_1(\tau)\\ \vdots \\ f_d(\tau)\end{pmatrix} \qquad (2.13)$$

There are:

$$J_f(\tau)=\begin{pmatrix} \partial f_1/\partial\tau_1 & \cdots & \partial f_1/\partial\tau_m \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \partial f_d/\partial\tau_1 & \cdots & \partial f_d/\partial\tau_m \end{pmatrix} \qquad (2.14)$$

f tangent space τ_τ at τ is generated from the m column vector of $J_f(\tau)$, with the highest dimension $m,\tau_\tau=\mathrm{span}\left(J_f(\tau)\right)$, the vector $\tau-\bar{\tau}$ is the coordinates of $f(\tau)$ in the affine space $f(\tau)+\tau_\tau, J_f(\tau)$ can not be solved because the exact form of f is not known, if $V_\tau\in R^{d\times m}$ is a Orthonormal basis matrix of τ_τ, then:

$$J_f(\tau)\left(\bar{\tau}-\tau\right)=V_\tau\theta_\tau^* \qquad (2.15)$$

And then there's:

$$\theta_\tau^* = V_\tau^T J_f(\tau)\left(\bar{\tau}-\tau\right)\equiv P_\tau\left(\bar{\tau}-\tau\right) \qquad (2.16)$$

And because when $O\left(\left\|\bar{\tau}-\tau\right\|_2^2\right)$ goes to 0 , there are:

$$f(\tau)-f(\tau)\approx J_f(\tau)(\tau-\tau) \qquad (2.17)$$

Namely:

$$f\left(\bar{\tau}\right)\approx f(\tau)+V_\tau\theta_\tau^* \qquad (2.18)$$

Let $Y_i=\left[x_{i1},\cdots,x_{is}\right]$ be x_i the nearest neighbor of s as measured by Euclidean distance (including x_i itself) for each sample point, the neighborhood of $x_{i_j}\approx x_i+V_i\theta_i$ and replace x_i with some x_i^* the optimization process can then be written:

$$\min_{x_i^*,V_i,\Theta}\sum_{j=1}^{s}\left\|x_{i_j}-\left(x_i^*+V_i\theta_i\right)\right\|_2^2=\min_{x_i^*,V_i,\Theta}\left\|Y_i-\left(x_i^*e^T+V_i\Theta_i\right)\right\|_F^2 \tag{2.19}$$

Where V_i is the standard orthogonal matrix of m column, $\Theta_i=\left(\theta_1,\cdots,\theta_s\right)\in R^{m\times s}$, thus V_i is the approximate orthogonal basis of m of the approximate tangent space of the manifold at x_i, the optimal value of x should be $\overline{x}_i$, and the optimal value of V_i should be the left singular vector corresponding to the singular value of m before $Y_i\left(I-\frac{1}{s}ee^T\right)$.

Definition 2. The initial sample point $X=\left(X_1,X_2,X_3,\ldots X_d\right)$ on a manifold embedded in dimensional Euclidean space, $\mathrm{M}\subseteq R^d$ is a submanifold of m Dimension, $\pi_{\mathrm{M}}:R^d\to\mathrm{M}$ is an orthogonal projection transformation that projects a point on X to M, $\tau:R^d\to F$ is a linear projection transformation that projects a point on X to a tangent space F, its base solution is as shown above, a_i is the solution of the equation $X_i=\tau_F\left(X\right)a_i$, for any $i,j\in\left\{1,\ldots,d\right\}$, The manifold form defining the correlation coefficient of $SEVP$ is:

$$SEVP=\frac{\mathrm{Var}\left(E\left(\tau_F\left(X\right)a_j\mid\ \tau_F\left(X\right)a_i\right)\right)}{\mathrm{Var}\left(\tau_F\left(X\right)a_j\right)}-R_iR_j\int_{\mathrm{M}}S_{i,j}\left(x\right)S_{j,i}\left(x\right)\mathrm{d}P_{\mathrm{M}} \tag{2.20}$$

Among them:

$$R_i:=1-\frac{\mathrm{Var}_i\left(X-\pi_{\mathrm{M}}\left(X\right)\right)}{\mathrm{Var}_i\left(X\right)}$$

$$S_{i,j}\left(x\right):=\frac{\partial}{\partial x_j}\left(x-\pi_{\mathrm{M}}\left(x\right)\right)\Big|_i$$

M is a k dimensional manifold and π is an orthogonal projection transformation.
When a_i and a_j can not be solved, the command:

$$\frac{\mathrm{Var}\left(E\left(\tau_F\left(X\right)a_j\mid\ \tau_F\left(X\right)a_i\right)\right)}{\mathrm{Var}\left(\tau_F\left(X\right)a_j\right)}=0 \tag{2.21}$$

Theorem 2.2. When the original sample data point X is on the Euclidean space, the manifold form of the correlation coefficient $SEVP$ is equal to the Euclidean space form.
Proof. Remember that the minimum affine Linear subspace of X Euclidean distance is L. When the original data sample point X is on an Euclidean space, the m submanifolds and L with the minimum Euclidean distance of X should be the same, there is $\mathrm{M}=L$, now $\beta\in R$, there is:

$$S_{i,j}\left(x\right)=\frac{\partial}{\partial x_j}\left(x-\pi_L\left(x\right)\right)\Big|_i\equiv\beta$$

$$S_{j,i}\left(x\right)=\frac{\partial}{\partial x_i}\left(x-\pi_L\left(x\right)\right)\Big|_j\equiv\frac{1}{\beta}$$

$$\rho^2_{X_i,X_j\mid\mathrm{M}}=R_iR_j\int_{\mathrm{M}}S_{i,j}\left(x\right)S_{j,i}\left(x\right)\mathrm{d}P_{\mathrm{M}} \tag{2.22}$$

$$
\begin{aligned}
&= R_i R_j \int_{\mathrm{M}} \mathrm{d}P_{\mathrm{M}} \\
&= R_i R_j \\
&= \rho^2_{X_i, X_j \mid L}
\end{aligned}
$$

And because:

$$\frac{\operatorname{Var}\left(E\left(\tau_F(X)a_j \mid \tau_F(X)a_i\right)\right)}{\operatorname{Var}\left(\tau_F(X)a_j\right)} = \frac{\operatorname{Var}\left(E\left(X_j \mid X_i\right)\right)}{\operatorname{Var}\left(X_j\right)} \tag{2.23}$$

III. Simulation analysis

In this chapter, we construct a group of sample sheets, and use the dimensionality reduction method to obtain $SEVP$ correlation coefficient.

Build the following sample point X :

$$a = 1.5 pai * \left(1 + 2 * \operatorname{rand}(1, N)\right)$$

$$\text{height} = h * \operatorname{rand}(1, N)$$

$$X = \left[a * \cos(a); height; a * \sin(a)\right]$$

rand $(1, N)$ is used to generate N random numbers between 0 and 1.Taking $h = 5, N = 5000$, it is obvious that the sample point X is a two-dimensional manifold embedded in a 3D. As shown in Figure 1 below, the dimension reduction effect is shown in Figure 2.

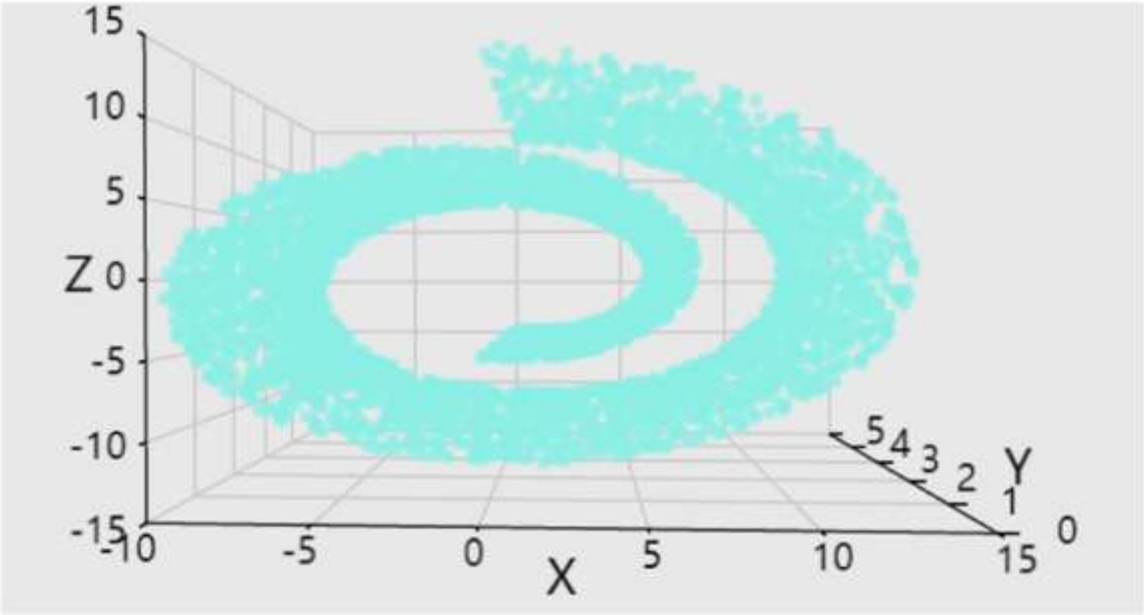

Figure 1. A scatter plot of 5000 sample points with a width h of 5

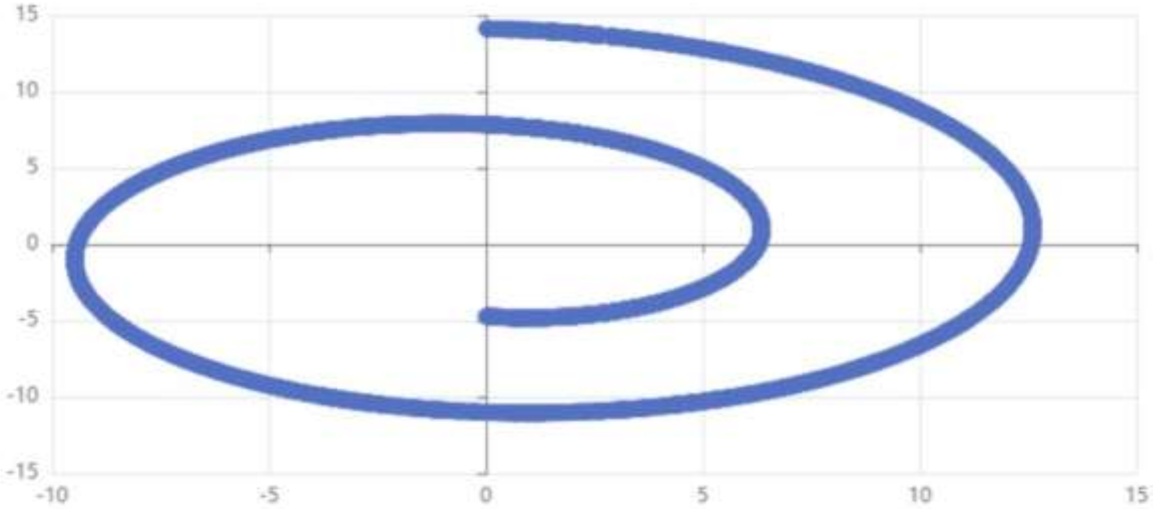

Figure 2. The result graph of dimension reduction by using tangent space

As you can see, the scatterplot in Figure 1 is a snail-like 3d scatterplot. The component of the sample point X on the X axis is very Linear independence to the component on the Z Axis, but the components on the Y Axis and the components on the X axis and the components on the Z Axis do not have a good correlation, figure 2 obviously preserves the snail-like structure of the original manifold and proves that the linear dimensionality reduction method has a good effect on preserving the nonlinear structure of the manifold. Make the original sample data point $X=\left(X_1,X_2,X_3\right)$, Using the random variable form of the correlation coefficient of $SEVP$ on a manifold, then the correlation coefficient of X_1 and X_3 is 0.845 and the data value is large. From the property of the correlation coefficient, it is known that there is a good nonlinear relationship between the variables.

IV. Summary and prospect

Non-linear correlation coefficient is a very important correlation coefficient, in the era of big data, the phenomenon of data redundancy is very prominent, often in the form of high-dimensional, non-linear, therefore, it is necessary to reduce the dimension of high-dimensional data, remove the redundant information, study the correlation coefficient between variables, and mine the essential characteristics of data. In this paper, a new nonlinear correlation coefficient $SEVP$ is proposed, and it is extended to the manifold, in this paper, a method to measure the nonlinear relationship between variables on a manifold is given, which is useful for us to study the nonlinear relationship between variables.

Because of the deficiency of professional knowledge and skill, this paper can not provide a simple and general formula of nonlinear correlation coefficient on manifold, we can only find a method to measure the Linear independence between variables on a manifold, and the steps are relatively complicated.

References

[1]. Ting W, Shiqiang Z. Study on linear correlation coefficient and nonlinear correlation coefficient in mathematical statistics [J]. Studies in Mathematical Sciences. 2011, 3(1): 58-63.

[2]. Liu R, Li X, Han L, et al. Track infrared point targets based on projection coefficient templates and non-linear correlation combined with Kalman prediction [J]. Infrared Physics & Technology. 2013, 57: $68-75$.

[3]. Zhou H, Wang X, Zhu R. Feature selection based on mutual information with correlation coefficient [J]. Applied Intelligence. 2022: 1-18.

[4]. Plaia A, Buscemi S, Sciandra M. Consensus among preference rankings: a new weighted correlation coefficient for linear and weak orderings [J]. Advances in Data Analysis and Classification. 2021, 15(4): 1015-1037.

[5]. Dong Xiaomeng. Nonlinear generalized correlation coefficient and its R language implementation [J]. Science, Technology and Engineering. 2010, 10(16): 3942-3943+3950.

[6]. Chen C, Zhang J, Fleischer R. Distance approximating dimension reduction of Riemannian manifolds [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2009, 40(1): 208-217.

[7]. Lee S M, Abbott A L, Araman P A. Dimensionality reduction and clustering on statistical manifolds [C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007: 1-7.

[8]. Il'kiv V S, Savka I Y. Nonlocal two-point problem for partial differential equations with linearly dependent coefficients [J]. Journal of Mathematical Sciences. 2010, 167(1).

[9]. Thomas P E, Fulton R W. Correlation of ectodesmata number with nonspecific resistance to initial virus infection [J]. Virology. 1968, 34(3): 459-469.

[10]. Zhang Z Y, Zha H Y. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment [J]. Journal of Shanghai University. 2004, (4): 406-424.

[11]. Chen K X, Ren J Y, Wu X J, et al. Covariance descriptors on a gaussian manifold and their application to image set classification [J]. Pattern Recognition. 2020, 107: 107463.

[12]. Faraki M, Harandi M T, Porikli F. A comprehensive look at coding techniques on riemannian manifolds [J]. IEEE Transactions on Neural Networks and Learning Systems. 2018, 29(11): 5701-5712.

[13]. Chen M, Lian Y, Chen Z, et al. Sure explained variability and independence screening [J]. Journal of Nonparametric Statistics. 2017, 29(4): 849-883.

[14]. Jian M, Guo J, Zhang C, et al. Semantic manifold modularization-based ranking for image recommendation [J]. Pattern Recognition. 2021, 120: 108100.

[15]. Fan Y, Wang G, Dong Q, et al. Tetrahedral spectral feature-Based bayesian manifold learning for grey matter morphometry: Findings from the Alzheimer's disease neuroimaging initiative $\left[J\right]$. Medical Image Analysis. 2021, 72: 102123.

[16]. Kontolati K, Alix-Williams D, Boffi N M, et al. Manifold learning for coarse-graining atomistic simulations: Application to amorphous solids [J]. Acta Materialia. 2021, 215: 117008.