大数据相关分析综述

梁吉业[1)] 冯晨娇[1),2)] 宋 鹏[1),3)]

[1)](山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)

[2)](山西财经大学应用数学学院 太原 030006)

[3)](山西大学经济与管理学院 太原 030006)

摘 要 大数据时代,相关分析因其具有可以快捷、高效地发现事物间内在关联的优势而受到广泛的关注,并有效地应用于推荐系统、商业分析、公共管理、医疗诊断等领域.面向非线性、高维性等大数据的复杂特征,结合现有相关分析方法的语义分析,文中从统计相关分析、互信息、矩阵计算、距离4个方面对大数据相关分析的现有研究成果进行了梳理.在对统计学中的经典相关分析理论进行归纳、总结的基础上,文中从大规模数据的通用性和均等性视角阐述了基于互信息的两个变量间非线性相关分析理论,从高维数据可计算的角度分析了基于矩阵计算的相关系数,从非线性、高维性数据的复杂结构方面解析了基于距离的相关系数.进一步地,该文在对已有相关分析方法进行分析与比较的基础上,围绕高维数据、多变量数据、大规模数据、增长性数据及其可计算方面探讨了大数据相关分析的研究挑战.

关键词 大数据;相关分析;相关系数;信息熵

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2016.00001

A Survey on Correlation Analysis of Big Data

LIANG Ji-Ye[1)] FENG Chen-Jiao[1),2)] SONG Peng[1),3)]

[1)](*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan* 030006)

[2)](*School of Applied Mathematics, Shanxi University of Finance & Economics, Taiyuan* 030006)

[3)](*School of Economics and Management, Shanxi University, Taiyuan* 030006)

Abstract In the big data time, correlation analysis has attracted much attention for its high-efficiency in analyzing inherent relation of things, and been effectively applied to many fields including recommender system, business analytics, public administration and medical diagnosis. Big data is usually nonlinear and high-dimensional. On the consideration of these complex characteristics and the semantic analysis for existing correlation analysis approaches, this paper gives a discussion of existing research findings of correlation analysis for big data. The discussion is analyzed from four aspects including statistical correlation analysis, mutual information, matrix calculation and distance. Based on summarizing classical correlation analysis theory in statistics, this paper firstly elaborates the nonlinear correlation analysis approaches between two stochastic variables induced by mutual information from the view of generality and equitability. Then, the correlation coefficient based on matrix calculation is analyzed in term of computability of high-dimensional data; and the distance correlation is analyzed from the point of complicated formation of nonlinear and high-dimensional data. Furthermore, on the account of analyzing and comparing

收稿日期:2015-04-17;**在线出版日期**:2015-09-25.**本课题得到国家自然科学基金**(61432011,U1435212,71301090)、**国家"九七三"重点基础研究发展规划项目基金**(2013CB329404)、**山西省高等学校创新人才支持计划**(2013052006)**资助**.**梁吉业,男,1962年生,博士,教授,中国计算机学会(CCF)理事,主要研究领域为粒计算、数据挖掘与机器学习.** E-mail: ljy@sxu.edu.cn. **冯晨娇,女,1977年生,博士研究生,讲师,主要研究方向为数据挖掘、统计学习方法、大数据相关分析.宋 鹏,男,1979年生,博士,副教授,主要研究方向为智能决策、数据挖掘.**

existing correlation analysis approaches, challenges of correlation analysis for big data are studied, namely high dimensional data, multivariable data, large-scale data, incremental data and its computability.

Keywords big data; correlation analysis; correlation coefficient; information entropy

1 引 言

随着信息与通信技术的迅猛发展,全球数据量呈现爆炸式增长.面对海量、复杂的数据,人们日益发现其是人类发展的重要经济资产,有效的数据分析与挖掘将推动国家、企业乃至整个社会的高效、可持续发展.

自 2008 年 9 月《Nature》出版"Big Data"专刊以来[1],大数据更是成为政府、学术界、实务界共同关注的焦点,如 2011 年《Science》出版的专刊"Dealing with Data"[2]和麦肯锡公司发布的报告"Big data: The next frontier for innovation, competition, and productivity"[3],2012 年达沃斯世界经济论坛上发布的报告"Big data, big impact: New possibilities for international development"[4]等.大数据分析与挖掘的研究成果也被广泛应用于物联网、舆情分析、电子商务、健康医疗、生物技术和金融等各个领域.

从国内研究进展来看,大数据研究也日益受到重视.李国杰院士和程学旗教授[5]围绕大数据的研究现状、科学问题、主要挑战以及发展战略进行了全面的分析与展望,为大数据的进一步深入研究提供了重要的研究思路;从具体研究进展来看,围绕大数据环境下的数据仓库架构[6]、大数据降维[7]、相关性分析[8]、海量数据应用[9]等方面的研究工作不断涌现,形成了一批重要的研究成果[10-11].同时,中国计算机学会于 2013 年出版了《中国大数据技术与产业发展白皮书》[12],2014 年出版了《中国大数据技术与产业发展报告》[13],并在大数据的发展背景、典型应用、技术进展、IT 产业链与生态环境以及发展趋势等方面进行了详细的阐述、分析与论证.

毋庸置疑,大数据研究之所以备受关注,本质原因在于其具有巨大的潜在价值[14].因此,可以肯定,大数据分析与挖掘技术,作为探测数据价值的关键手段,在大数据研究中具有极其重要的位置.

值得强调的是,在大数据分析与挖掘任务中,相关分析的研究受到更为广泛的关注和重视.事实上,相关分析的研究从 1888 年 Galton[15]关注人类身高与前臂长度的关系开始,就已经引起了人们的注意.然而,从人类的思维方式来看,人们并不仅仅满足于发现相关关系,而是在相关关系的基础上进一步探索因果关系,即在发现两个相关事物的基础上进一步探究哪一事物为因哪一事物为果.尽管因果关系的准确发现仍然非常困难,但人们可以通过设立假设、实验验证等反复尝试的繁琐手段探索这一难题[16].显然,在传统的统计抽样背景下,这一繁琐的任务是可以接受的.但是,大数据时代,人们更加关注数据总体,并试图从数据总体中直接获取有价值的信息,而用于发现因果关系的传统的反复尝试方法就变得异常困难.与此相反,从亚马逊的推荐系统[17]、谷歌的流感预测[18]等诸多现实案例中,人们日益发现,与以往相比,大数据时代相关关系的探索具有更加重要的价值.

特别是,近年来大数据相关分析的应用成果不断涌现,使得相关分析的研究成为学界、实务界关注的热点问题.在大数据相关分析的诸多应用领域中,最为引人注目的是推荐系统[19].基于相关系数给出用户相似性、物品相似性的度量,进而进行产品推荐;进一步地,相关系数还是推荐系统的一类重要评价指标.2009 年发表在《PNAS》上的文献"Predicting Social Security Numbers from public data"[20],则以美国公众数据为研究对象,研究发现公民的社会安全号码(Social Security Number)与其出生时间、出生地具有显著的相关关系,研究成果揭示了个人隐私泄露的可能诱因.文献[21]则面向药物基因组大数据,基于协方差矩阵的稀疏建模与奇异值分解,探测与癌症相关的重要基因组.此外,大数据相关分析在灾害应急管理[22]、医疗诊断[23]等领域也有着广泛的应用.

当然,大数据研究中,"相关关系"与"因果关系"的争论从未停止,李国杰院士和程学旗教授在文献[5]中也进行了详细的分析.文中指出"因果关系本质上是一种相互纠缠的相关性",并进一步强调"大数据的关联分析是不是'知其然而不知其所以然',其中可能包含深奥的哲理,不能贸然下结论".需要进一步说明的是,尽管争论还将存在,但是不可否认

的是，大数据的相关分析能够满足人类的众多决策需求，因此，有效的发现与度量相关关系具有重要的研究价值. 从科学层面来看，面对混杂的大数据，相关分析可以帮助人们更快捷、更高效地发现事物间的内在关联[16]，其本身不仅仅是一项重要的数据分析与挖掘任务，同时也为数据的深度分析与挖掘进而发现事物内在规律提供“导航”功能；因此，尽管大数据分析中有“相关关系”与“因果关系”之争论，但不可否认的是，大数据相关分析必然是大数据分析与挖掘的关键科学问题之一. 从应用层面来看，商业企业作为大数据应用的重要领域，其核心目标是实现利润的增长，因此，其在数据分析与挖掘中的核心任务是探测何种经营策略与利润增长具有相关性，而并不必然要求探究经营策略与利润增长之间的因果关系，即“从数据到价值”的商业范式，而非“从数据到信息再到知识”的科学范式[5]. 综合来看，可以肯定，大数据相关分析已经成为大数据分析与挖掘的核心科学问题与关键应用技术.

所谓相关关系，是指 2 个或 2 个以上变量取值之间在某种意义下所存在的规律，其目的在于探寻数据集里所隐藏的相关关系网[5]. 从统计学角度看，变量之间的关系大体可分两种类型：函数关系和相关关系. 一般情况下，数据很难满足严格的函数关系，而相关关系要求宽松，所以被人们广泛接受. 需要进一步说明的是，研究变量之间的相关关系主要从两个方向进行：一是相关分析，即通过引入一定的统计指标量化变量之间的相关程度；另一个是回归分析. 由于回归分析不仅仅刻画相关关系，更重要的是刻画因果关系，因此本篇文章讨论的相关关系为前者.

相关分析的研究成果中，最具影响力的是早在 1895 年由 Pearson 提出的积矩相关系数（也称皮尔逊相关系数）[24]. 在长达 100 多年的时间里，相关分析得到实践的检验，并广泛地应用于机器学习、生物信息、信息检索、医学、经济学与社会统计学等众多领域和学科. 进入大数据时代，作为度量事物之间协同、关联关系的有效方法，大数据相关分析由于其计算简捷、高效，必将具有更强的生命力. 但是，由于大数据具有数据规模大、数据类型复杂、价值密度低等特征，因此，如何找到有效且高效的相关分析计算方法与技术则成为大数据分析与挖掘任务中亟待解决的关键问题.

目前，常见的大数据相关分析分为两类：一类是面向高度复杂的数据关系（换言之，大数据的现实背景往往是非线性复杂系统[25]），传统的线性相关分析方法显然难以刻画变量之间的非线性等复杂关系，因此，研究者基于互信息和距离测度探索了变量间的非线性等复杂相关关系[26-27]；另一类是面向高维数据（如基因数据、天文数据），利用协方差矩阵内在的稀疏性特征，建立基于稀疏性约束的参数估计方程，通过快速求解来提升处理数据的能力. 类似于文献[28-29]利用稀疏技术处理 PCA 和回归方程. 这两类问题经常混杂在一起，也就是说我们经常见到的是高维复杂数据，需要同时进行维数约简和非线性描述. 那么，从不同的角度采用不同的研究方法就得到大数据相关分析的各种模型.

围绕前述非线性、高维性问题，本文试图从统计相关分析、互信息、矩阵计算和距离 4 个方面对现有研究成果进行梳理、分析与总结. 统计相关分析主要刻画变量（包括两个变量、多变量）间的线性相关关系，基于互信息的相关分析考查两个变量间的非线性相关关系，基于矩阵计算的相关分析围绕高维性探讨随机向量间的线性相关关系，基于距离的相关关系则同时考虑了非线性、高维性问题，研究了高维数据的非线性相关关系.

当然，从语义上来看，上述 4 个方面的研究也有所不同. 相关分析是指一个变量的变化是否会影响另一个变量变化. 经典的统计学方法正是在这一含义下提出了消减误差比例[30]（Proportionate Reduction in Error，PRE）（详见 2.2 节），并以此为基础，提出了各种相关系数. 基于互信息的相关分析是从信息量角度来刻画相关性，即已知某一变量情况下，另一变量信息量的变化程度. 基于矩阵计算的相关分析由于考查的仍是线性相关关系，因此，从语义上来看，其与经典的统计学方法一致；所不同的是，其目标在于刻画高维变量间的相关关系. 基于距离的相关分析则是从分布函数角度来刻画，即在已知某一变量的情况下，另一变量分布函数的变化程度.

从大数据相关分析的研究进展来看，尽管其研究成果尚不丰富，仍处于起步阶段，但对现有成果的梳理与总结，可以为关注大数据相关分析这一大数据分析中关键问题的研究者提供借鉴.

本文第 2 节总结统计学中的各种相关系数；第 3 节阐述基于互信息的两个变量间的非线性相关分析方法；第 4 节分析基于矩阵计算的随机向量间的相关分析方法；第 5 节探讨基于距离的高维数据的非线性相关分析；最后是研究展望与总结.

2 统计学中的相关分析理论

事实上,相关分析在统计研究中早已有所讨论,只是相对于大数据分析而言具有一定的局限性.在统计学中,相关系数种类繁多,我们首先给出相关系数的定义,然后介绍关于不同类型变量以及随机向量的相关系数表示.

2.1 相关的定义及性质

在19世纪80年代,Galton[15]通过研究人类身高遗传问题首次提出了相关的概念.文中指出相关关系可以定义为"一个变量变化时,另一个变量或多或少也相应地变化",而测量这种相关关系的统计量则称为相关系数.相关关系有强弱之分,大多数的相关系数是用0代表不相关,用1代表全相关.介于0～1之间的数,数值越大相关性越强,数值越小相关性越弱.另外,关系有方向之分,若一个变量增加,另一个变量也增加称为正相关,用正数表示同方向;若一个变量增加,另一个变量减少则称为负相关,用负数表示反方向.

2.2 两个变量之间的相关系数

本节我们将用 X,Y 代表两个随机变量.当 X,Y 均为一维变量时,分别用 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_n$ 表示随机变量的取值,n 称为样本容量,样本均值记为 $\bar{x},\bar{y}$.事实上,在机器学习中,当 X,Y 均为一维变量时,可将其看作是样本的两个特征,$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 是 n 个样本在两个特征上的取值.进一步地,用 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$,$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_q)$ 表示 p 维、q 维随机向量,其中,$X_1,X_2,\cdots,X_p$ 为 $\boldsymbol{X}$ 的 p 个特征,$Y_1,Y_2,\cdots,Y_q$ 为 $\boldsymbol{Y}$ 的 q 个特征,更详细的描述见本文第4节.

统计上常用消减误差比例衡量两个变量之间的相关性.假设不知道 X 值预测 Y 值时产生的误差用 E_1 表示,如果知道 X 值来预测 Y 值时产生的误差用 E_2 表示,则

$$PRE=(E_1-E_2)/E_1 \tag{1}$$

PRE 的值越大说明已知 X 值对预测 Y 值的帮助越大,也就是说 X 与 Y 的相关性越大.因此 PRE 是一种适用于任何类型数据的相关系数.

2.2.1 两个定类变量之间的相关系数

定类变量即名义变量,是指变量的值是研究对象的名称或符号.每个值代表一个类别,这些值之间没有大小、次序之分,是平等的.如对于性别这个变量而言其取值为男、女两类.

计算这类变量之间的相关性通常需借助列联表.列联表又称交互分类表,是一种将样本按照两个或两个以上的特征分类后的交叉频数表.假设有两个特征 X 和 Y,特征 X 有 k 类,用 $X_{(i)}$ 表示第 i 类,$i=1,2,\cdots,k$;特征 Y 有 l 类,用 $Y_{(j)}$ 表示第 j 类,$j=1,2,\cdots,l$.对于 n 个样本,用 n_{ij} 代表既属于特征 X 的第 i 类又属于特征 Y 的第 j 类的样本频数.由此可以得到一个 k 行 l 列的列联表,见表1.

表1 列联表的一般形式

	$Y_{(1)}$	$Y_{(2)}$	…	$Y_{(l)}$
$X_{(1)}$	n_{11}	n_{12}	…	n_{1l}
$X_{(2)}$	n_{21}	n_{22}	…	n_{2l}
…	…	…	…	…
$X_{(k)}$	n_{k1}	n_{k2}	…	n_{kl}

以最简单的2×2列联表为例,假设我们对性别(特征 X)与本科毕业生就业难易程度(特征 Y,包括就业好和就业难两类)感兴趣.在调查的200名本科毕业生中,就业好的学生中有80名是男性,15名是女性;另一方面,就业难的学生有20名是男性,85名是女性.见表2.

表2 性别与就业难易程度相关分析的2×2列联表

	$Y_{(1)}$/就业好	$Y_{(2)}$/就业难	
$X_{(1)}$/男	$n_{11}=80$	$n_{12}=20$	100
$X_{(2)}$/女	$n_{21}=15$	$n_{22}=85$	100
	95	105	200

在众多定类变量的相关系数中,Q 系数是计算两个变量相关性的最简单方法,仅适用于2×2列联表,公式如下[31]:

$$Q=(n_{11}n_{22}-n_{12}n_{21})/(n_{11}n_{22}+n_{12}n_{21}) \tag{2}$$

我们可以取两种特殊情况理解 Q 系数的构造.若 $n_{12}=n_{21}=0$,则 Q 系数为1;若 $n_{11}=n_{22}=0$,则 Q 系数为−1.显然这两种情况都表明性别与就业是完全相关的,而正负号在这里表明了两个特征所属类别中具有相关性的类别的不同.比如,在上述例子中 Q 系数为1代表男性易于就业,女性难就业,Q 系数为−1则代表女性易于就业,男性难就业.

λ 系数可以计算任意两个定类变量的相关性,适用于任意维数的列联表,公式如下[31]:

$$\lambda=\frac{\sum_i f_{xi}+\sum_j f_{yj}-(F_x+F_y)}{2n-(F_x+F_y)} \tag{3}$$

其中:f_{xi} 是第 i 行的众数(即频数的最大值);f_{yj} 是第 j 列的众数;F_x 是边际行众数;F_y 是边际列众数.仍以表2为例,$f_{x1}=80$,$f_{x2}=85$,$f_{y1}=80$,$f_{y2}=85$,

$F_x=100$，$F_y=105$，则 $\lambda=0.641$. 这一系数表明了男同学的就业易于女同学，当然，从样本数据取值情况来看，也与这一结果相吻合.

除了 Q 系数，λ 系数之外，还有 χ^2 检验，φ 系数，C 系数，V 系数. 由于 λ 系数计算前提条件宽松，计算相对简单，且具有消减误差比例意义，而成为较为常用的一种衡量定类变量相关性的统计量. 其它相关系数由于篇幅关系这里不一一介绍，细节可参见文献[31-32].

2.2.2　两个定序变量之间的相关系数

定序变量即等级变量，变量取值具有序的意义，换言之，其取值有等级或次序之分. 如高校教师职称分为助教、讲师、副教授、教授 4 个等级. 在定序变量的相关系数度量中，常用的概念有同序对、异序对. 具体定义是：如果某对样本在两个特征上的相对等级是一致的，即对于一对样本 (x_1,y_1) 和 (x_2,y_2) 而言，在序上 x_1 优于 x_2，同时 y_1 优于 y_2，则称为同序对，同序对数用 n_s 表示；相反则称之为异序对，异序对数用 n_d 表示. 下面是两个定序变量之间常用的相关系数[31,33].

(1) γ 系数.

$$\gamma=(n_s-n_d)/(n_s+n_d) \tag{4}$$

这个公式的直观含义是对于所有的样本对，其同序对数和异序对数之差与同序对数和异序对数之和的比例. 比如：若 $n_d=0$，则 $\gamma=1$，即对于两个特征来说，它们所有样本对都是同序的，则我们认为两个特征是完全正相关；反之，若 $n_s=0$，则 $\gamma=-1$，即对于两个特征来说，它们所有样本对都是异序的，则我们认为两个特征是完全负相关.

(2) 斯皮尔曼(Spearman)相关系数[34].

$$r_s=1-\left(6\sum_{i=1}^{n}d_i^2\right)/(n(n^2-1)) \tag{5}$$

其中，$d_i=x_i'-y_i'$，x_i'，y_i' 为样本 i 在两个特征下排序后的等级值.

斯皮尔曼相关系数是由英国统计学家斯皮尔曼根据皮尔逊相关系数的概念推导而来，其统计意义可以看作是皮尔逊相关系数的特例.

就定序变量间的相关系数而言，γ 具有消减误差比例意义，r_s 平方后具有消减误差比例意义，因而对定序变量进行相关性衡量时多采用这两个相关系数. 当然在衡量定序变量相关性时，还有 dy 系数、肯德尔系数(常用的有 3 种，分别记为 $tau\text{-}a$，$tau\text{-}b$，$tau\text{-}c$)等其它相关系数，它们类似 γ 系数，仅仅是对分母做了一些修正，由于它们没有消减误差比例意义，因此使用较少[35].

2.2.3　两个定距变量之间的相关系数

定距变量即数值变量，变量之间具有数量差别，可以进行加减乘除运算. 度量其相关性的最为常用的相关系数是皮尔逊相关系数(又称积距相关系数)[24]

$$\rho=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}=\frac{E(X-\bar{X})(Y-\bar{Y})}{\sqrt{E(X-\bar{X})^2E(Y-\bar{Y})^2}} \tag{6}$$

其样本相关系数为

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{7}$$

从式(7)来看，其是两个随机变量样本取值标准化后乘积之和. 为了更好地理解皮尔逊相关系数的含义，可从几何角度给出形象的解释. 通常描述两个随机变量 X,Y 相关关系的几何模型是样本散点图，即样本在两个变量轴构成的直角坐标系平面上的分布图. 对上述空间进行变换，将每一个样本作为一个数轴，相应地，对于 n 个样本则产生 n 个数轴，进而构造一个 n 维空间. 这个空间只包含两个点，即每一个变量对应一个点. 在这个高维空间上，这两个点可以视为两个向量的端点(如图 1 所示). 通过上述方法，可以将 n 个样本 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 转换为两个向量 $\boldsymbol{\xi}=(x_1,x_2,\cdots,x_n)$，$\boldsymbol{\eta}=(y_1,y_2,\cdots,y_n)$. 不失一般性，假设两个向量已中心化，夹角 θ 的余弦值为

$$\cos\theta=\frac{\langle\boldsymbol{\xi},\boldsymbol{\eta}\rangle}{\|\boldsymbol{\xi}\|\cdot\|\boldsymbol{\eta}\|}=\frac{\sum_{i=1}^{n}x_iy_i}{\sqrt{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2}}=r \tag{8}$$

其中：$\langle\cdot,\cdot\rangle$ 代表向量内积；$\|\cdot\|$ 代表向量的长度.

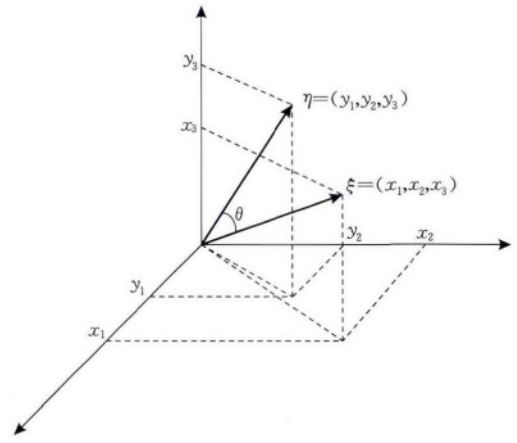

图 1　3 个样本 $(x_1,y_1),(x_2,y_2),(x_3,y_3)$ 生成的三维空间图

当 $\theta=0°$，代表 $\boldsymbol{X},\boldsymbol{Y}$ 两个向量夹角为 0，即两个向量同方向重合，说明两个向量正线性相关，则 $r=\cos\theta=1$；当 $\theta=90°$，代表 $\boldsymbol{X},\boldsymbol{Y}$ 两个向量夹角为 90°，即两个向量垂直，说明两个向量线性无关，则 $r=\cos\theta=0$；当 $\theta=180°$，代表 $\boldsymbol{X},\boldsymbol{Y}$ 两个向量夹角为 180°，即两个向量异方向重合，说明两个向量负线性相关，则 $r=\cos\theta=-1$.

当然，皮尔逊相关系数同样存在不足：(1) 当变量不服从正态分布时，即使是对大样本而言，r 也有相当大的偏差[36]；(2) r 的计算易受异常点的影响，且影响较为显著[37]. 基于上述原因，相关学者对皮尔逊相关系数进行了改进[38-39]. 但是，皮尔逊相关系数公式具有丰富的内涵，文献[40]分别从几何、代数、三角等不同角度给出了 13 种理论解释，同时，其平方后也具有消减误差比例意义，因此成为认可度最高的用于刻画定距变量相关性的相关系数.

2.2.4 两个变量之间相关系数层次图

层次图(如图 2 所示)说明：以圆盘代表变量类型，依次为定类变量、定序变量、定距变量. 上半圆中的相关系数具有消减误差比例意义，下半圆中的相关系数不具有消减误差比例意义. 箭头代表对前一个相关系数的改进.

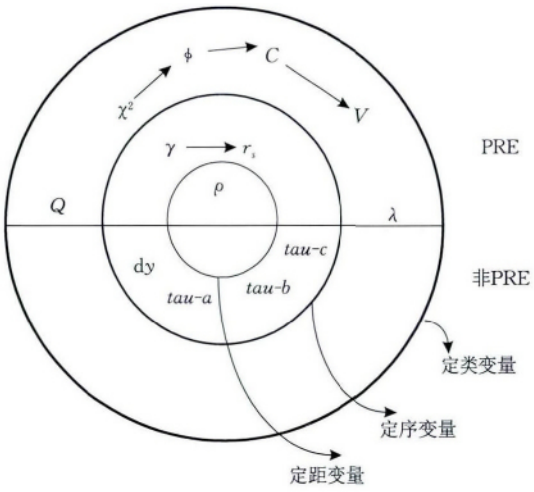

图 2 两个变量之间相关系数层次图

2.2.5 混合变量之间的相关系数

混合变量之间的相关系数通常采用两种方法计算. 一种是消减误差比例 PRE. 例如变量 X 是定类变量，共有 k 类，每类的样本数分别为 $n_1,n_2,\cdots,n_k$，且 $n_1+n_2+\cdots+n_k=n$，变量 Y 是定距变量. 若未知 X，则全部误差平方和为 $E_1=\sum_{j=1}^{n}(y_j-\bar{y})^2$，其中 y_j 代表变量 Y 的第 j 个样本；若已知 X，则可将样本划分为 k 类，每类的类内均值为 $\bar{y}_{(i)}(i=1,2,\cdots,k)$，从而可以计算类内误差平方和 $E_2=\sum_{i=1}^{k}\sum_{j=1}^{n_k}(y_{(i)j}-\bar{y}_{(i)})^2$，其中，$y_{(i)j}$ 代表 Y 的第 i 类中的第 j 个样本，将 E_1 和 E_2 代入式(1)即得消减误差比例 PRE. 另一种方法是降级处理，比如定类变量和定序变量之间的相关系数，可以把定序变量降级为定类变量，应用定类变量之间的相关系数进行计算，当然这必然导致信息损失.

2.3 多变量相关系数

多变量相关系数包括多个变量中固定其它变量时任意两个变量的偏相关系数、一个变量对多个变量的复相关系数、多个变量对多个变量的典型相关系数. 需要指出的是，这些相关系数均是对定距变量的线性相关关系(式(7))的推广.

2.3.1 偏相关系数

设 $X,Y,\zeta_1,\zeta_2,\cdots,\zeta_p$ 是 $p+2$ 个随机变量. 当 $p=1$ 时，偏相关系数是指剔除第 3 个变量 ζ_1 对 X,Y 的影响后 X,Y 的相关系数：

$$r_{XY\cdot1}=(r_{XY}-r_{X1}r_{Y1})/(\sqrt{1-r_{X1}^2}\sqrt{1-r_{Y1}^2}).$$

其中 r_{XY},r_{X1},r_{Y1} 分别代表 X 和 Y，X 和 ζ_1，Y 和 ζ_1 的皮尔逊相关系数. 推广上式，当 $p=2$ 时，偏相关系数即为剔除 ζ_1,ζ_2 对 X,Y 的影响后 X,Y 的相关系数，公式如下：

$$r_{XY\cdot12}=(r_{XY\cdot1}-r_{X2\cdot1}r_{Y2\cdot1})/(\sqrt{1-r_{X2\cdot1}^2}\sqrt{1-r_{Y2\cdot1}^2}).$$

以此类推，有 $r_{XY\cdot123}=(r_{XY\cdot12}-r_{X3\cdot12}r_{Y3\cdot12})/(\sqrt{1-r_{X3\cdot12}^2}\sqrt{1-r_{Y3\cdot12}^2})$ 等等.

另一种推广是若 ζ_1 只对 X 有影响，称为半偏相关系数，公式如下：

$$r_{Y(X\cdot1)}=(r_{XY}-r_{Y1}r_{X1})/\sqrt{1-r_{X1}^2}.$$

若只对 Y 有影响，由于公式类似，此处省略.

2.3.2 复相关系数

复相关系数描述的是 Y 与 p 个变量 $\zeta_1,\zeta_2,\cdots,\zeta_p$ 之间的相关性. 当 $p=2$ 时，公式如下：

$$R_{Y\cdot12}^2=r_{Y1}^2+r_{Y2\cdot1}^2(1-r_{Y1}^2),$$

等价于

$$R_{Y\cdot12}^2=(r_{Y1}^2+r_{Y2}^2-2r_{Y1}r_{Y2}r_{12})/(1-r_{12}^2).$$

同理，可以推广至 3 个及以上自变量对 Y 的影响.

2.3.3 典型相关系数

典型相关系数是研究两个随机向量之间相关性的一种多元统计方法[41]. 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$，$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_q)$ 分别是 p 维、q 维随机向量，不失一般性，设 $\boldsymbol{X},\boldsymbol{Y}$ 已经中心化，典型相关分析旨在

寻找两个投影向量 w_x，w_y，使得数据在该投影方向上的皮尔逊相关系数最大：

$$\rho_d = \max_{w_x, w_y} \frac{w_x' \boldsymbol{\Sigma}_{XY} w_y}{\sqrt{w_x' \boldsymbol{\Sigma}_{XX} w_x}\ \sqrt{w_y' \boldsymbol{\Sigma}_{YY} w_y}} \tag{9}$$

这可以通过优化方法获得

$$\max_{w_x, w_y} w_x' \boldsymbol{\Sigma}_{XY} w_y$$
$$\text{s.t. } w_x' \boldsymbol{\Sigma}_{XX} w_x = w_y' \boldsymbol{\Sigma}_{YY} w_y = 1,$$

其中：$\boldsymbol{\Sigma}_{XX} = E\boldsymbol{X}\boldsymbol{X}'$，$\boldsymbol{\Sigma}_{XY} = E\boldsymbol{X}\boldsymbol{Y}'$，$\boldsymbol{\Sigma}_{YY} = E\boldsymbol{Y}\boldsymbol{Y}'$，$E$ 为数学期望.

典型相关分析的基本思想是寻找两个线性变换来抽取特征，把原始的数据压缩到一个低维的子空间中，使得数据在该子空间中的线性关系最大. 从现有研究来看，典型相关分析已经被广泛应用于电子通信、医学、生物信息、遥感、航天、经济管理等研究领域[42].

2.4　大数据中的统计相关分析

大数据的非线性、高维性以及海量性等复杂特征为经典的统计相关分析方法带来了新的挑战. 这不仅仅包括如何有效度量相关关系，还包括如何从大数据中有效识别伪相关.

2.4.1　典型相关分析的新进展

在众多的统计相关系数中，典型相关系数由于能够考查随机向量间的相关关系，而在大数据时代受到了更多的关注.

从典型相关分析的计算方法来看，本质上是将问题求解转化为矩阵的特征值与特征向量的求解，其中，矩阵的运算涉及随机向量的协方差矩阵以及协方差矩阵的逆. 传统的统计分析中，存在一个重要假设，即协方差矩阵是可计算的；而主要的求解困难在于协方差矩阵的逆，如小样本问题导致矩阵奇异. 针对这一困难，文献[43]利用奇异值分解，采用伪逆来解决协方差矩阵奇异的问题. 然而，在高维情况下，无论协方差矩阵还是协方差矩阵的逆，其计算耗时都将非常巨大；同时，存在的矩阵奇异问题也将导致逆矩阵的不可计算. 从现有研究进展来看，正则化方法是一类主要的解决手段. 关于高维协方差矩阵的计算，本文将在 4.2 节详细阐述，这里仅就高维情况下协方差矩阵的逆的求解进行分析. 实际上，正则化方法类似于岭回归，通过在协方差矩阵上添加参数倍的单位矩阵（即 $\boldsymbol{G} = \boldsymbol{\Sigma} + \lambda \boldsymbol{E}$，其中 $\boldsymbol{E}$ 为单位矩阵），从而用正则矩阵 $\boldsymbol{G}$ 代替协方差矩阵 $\boldsymbol{\Sigma}$，进而有效避免矩阵的不可逆问题. 可以看出，参数 λ 的估计是保证正则化方法有效的关键问题. 文献[44]基于均方误差最小准则给出了参数 λ 的估计方法，且无需进行分布假设，同时还避免了类似于 Bootstrap、Cross-Validation 的复杂计算过程，进而保证了参数估计的计算效率；文献[45-46]则面向高维协方差矩阵，从正则矩阵正定性、计算效率的提升给出了系列的改进，为高维情况下典型相关系数的计算提供了有效的求解技术.

此外，经典的典型相关系数只能度量随机向量间的线性相关关系，对于大数据中常见的非线性相关而言，其仍然存在局限. 因此，相关学者开展了典型相关分析从线性到非线性推广的研究. 文献[47]基于互信息方法对典型相关分析进行了推广. 文献[48]基于核化原理，通过非线性映射，将样本映射到高维特征空间，从而提出了核典型相关判别分析方法，并针对抽样样本数的确定问题，基于在线稀疏思想给出了一个具有较高计算效率的自适应学习算法，可适用于大规模数据分析. 文献[49]则针对刻画非线性相关关系的核典型相关分析研究了收敛率的问题. 当然，这些模型、算法仍然受到自身方法的约束：对于互信息方法而言，其密度函数的估计是难点所在；就核方法而言，如何选择恰当的核函数及相应参数显然是另一个研究挑战.

2.4.2　伪相关

所谓伪相关是指两个并不具有相关关系的变量却具有高的样本相关系数的一种统计现象[50]. 显然，伪相关将导致错误的统计推断甚至误导科学发现. 就伪相关的产生原因而言，是由于其他未见因素（也称“第三变量”）的影响，而得出具有误导性的两个变量之间在统计上的相关系数[51]. 统计学上典型的例子是：荷兰统计数据显示，在连续的几个春季中，鹳鸟巢的数量与人类婴儿出生数量之间呈现正相关关系；但事实上，这两者之间并不存在相关关系，而是皆与数据观测之前 9 个月的天气相关.

在经典的统计学相关分析方法中，偏相关系数是发现伪相关关系的重要手段. 从偏相关系数的定义来看，其考查的是剔除其他因素影响后两个变量之间的相关关系. 就鹳鸟巢数量与人类婴儿出生数量这一伪相关实例而言，当剔除“数据观测之前 9 个月的天气”这一影响因素后，鹳鸟巢数量与人类婴儿出生数量并不呈现相关关系.

值得强调的是，大数据情境下，数据的海量性、高维性将大大增加伪相关发生的可能性[50]；进一步地，海量性、高维性也使得伪相关的发现变得更加困难. 而本质上来看，伪相关的识别关键在于“第三变量（可以是单个变量，也可以是多个变量）”的探寻.

因此,大数据时代,如何从复杂庞大的数据集中快速、准确的发现“第三变量”是伪相关判别的重要瓶颈.毕竟,面向整体数据集的反复计算是我们难以接受的.那么解决这一问题的可能路径是什么?笔者认为,大数据时代的来临,并不意味着经典统计学中统计抽样、假设检验都应摒弃;而应借鉴经典的统计抽样思想,设计有效的拆分与融合策略(如:如何有效的保持整体性质等),从而在保证计算效率的条件下实现研究目标.其主要思想是,在特定的数据分析与挖掘任务下,按照某种策略,将大数据集拆分为若干小数据集,然后在每一个小数据集上进行数据分析,在此基础上,将每个小数据集上的数据分析结果融合,进而形成一个整体数据集上的推断.笔者文献[7]的研究成果面向大数据的降维问题开展了数据集的拆分与融合策略研究.策略的设计考虑到了小数据集与大数据集分布的近似性、各小数据集间的信息传递性、数据分析时样本的整体覆盖性等.研究结果表明,所提出的拆分与融合策略,既显著提高了大数据集的计算效率,又获得了满意的降维结果.我们认为,这一尝试可为大数据情境下高效发现“第三变量”进而识别伪相关关系提供可行的数据挖掘技术.当然,本质上来看,这种拆分与融合策略有望对大数据的“海量性”这一困难问题提供可行的求解路径.

3 基于互信息的相关分析

就大数据而言,数据关系往往呈现非线性等复杂特征.因此,经典的线性相关分析方法难以有效探测数据的内在结构与规律.从现有研究进展来看,基于互信息的度量准则,由于其具有能够有效刻画非线性相关关系的优势,而日益受到重视.

我们知道,对于信息系统而言,信息熵是有效刻画信息含量(信息结构、不确定性等)的度量工具.熵最早由德国物理学家 Clausius 于 19 世纪 50 年代提出,用于度量能量在空间中分布的均匀程度,若能量分布越均匀,则熵值越大.在此基础上,信息论之父 Shannon(1948)[52]借鉴熵的概念,将信息中排除冗余信息的平均信息量定义为“信息熵”,并给出了信息熵的数学形式化表示.

考虑有 n 个可能结果的随机变量 X,其概率分布为 $P(X=x_i)=p_i, i=1,2,\cdots,n$.则其信息熵的定义为 $H(X)=-\sum_{i=1}^{n} p_i \log p_i$.

设随机向量 (X,Y) 的联合概率分布为 p_{ij},则 (X,Y) 的二维联合熵为

$$H(X,Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}\log p_{ij}.$$

假定 X 和 Y 的边际分布分别为 $p_{i\cdot}$ 和 $p_{\cdot j}$,可定义在已知 Y 的条件下 X 的条件熵

$$H(X/Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}\log\frac{p_{ij}}{p_{\cdot j}},$$

同理,可得在已知 X 的条件下 Y 的条件熵为

$$H(Y/X)=-\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}\log\frac{p_{ij}}{p_{i\cdot}}.$$

信息论认为,系统越有序,则信息熵越小;相反地,系统越混乱,则信息熵越大.因此,信息熵可以作为系统不确定性程度(或者说有序化程度)的度量标准.一般而言,$H(X)-H(X|Y)$(等价于 $H(Y)-H(Y|X)$)表示已知 $Y(X)$ 的情况下 $X(Y)$ 信息量的变化程度.显然,若变化程度较小则表明 $Y(X)$ 对 $X(Y)$ 的影响较小,也就是说,X 与 Y 相关性弱.反之,说明 X 与 Y 相关性强.这个差值称为互信息,记为 $I(X,Y)$.

互信息作为相关分析的度量,其最大优势在于能有效刻画变量之间的非线性关系.在大数据相关分析中,最具影响力的研究成果是 Reshef 等人于 2011 年发表在《Science》上的论文“Detecting novel associations in large data sets”[26].研究中通过互信息定义了两个变量之间的最大信息系数(Maximal Information Coefficient,*MIC*),用来衡量两个变量之间的相关性.这一系数包含两个重要性质:通用性(Generality)、均等性(Equitability).具体而言,传统的相关系数往往是针对特定的函数类型(如线性、指数、周期性函数)测量变量之间的相关性程度,而 *MIC* 可测量任何函数形式的相关性,包括叠加函数,因此,具有通用性.均等性则是指,对于具有相等 *MIC* 取值的不同函数形式的数据而言,当给予同等程度的噪音,*MIC* 的取值仍然保持相等,这在传统的统计方法中是很难做到的.

MIC 的直观理解是:对于变量 X,Y 的散点图而言,存在某种网格覆盖散点图;根据各散点在网格中子格内的频率来计算变量 X,Y 之间的相关系数.具体而言,首先对数据集进行网格划分,即分别在 x 轴、y 轴上进行行、列划分从而形成具有 x 列、y 行的网格 G;对给定的有序对集合 D,D 中的每一个序对将被置于某一个子格中,可以容许某些子格为空.然后,对于给定的 x 列、y 行,计算任意划分的网格

对应的最大互信息 $I^*(D,x,y)=\max I(D|_G)$. 其中，$I(D|_G)$(简记为 I_G)是 $D|_G$的互信息，$D|_G$的概率分布是通过网格 G 中每个子格中散点的频率给出；$I^*(D,x,y)$代表在给定的 x 列、y 行情况下的最大互信息. 最后，针对任意的 x 列、y 行，基于每个 $I^*(D,x,y)$标准化得到特征矩阵：

$$\boldsymbol{M}(D)_{x,y}=\frac{I^*(D,x,y)}{\log\min\{x,y\}},$$

在此基础上定义信息最大相关系数(MIC)：

$$MIC=\max_{xy<B(n)}\{\boldsymbol{M}(D)_{x,y}\} \tag{10}$$

其中，$xy<B(n)$是指网格分割细度小于 $B(n)$.

从 MIC 的定义来看，网格划分存在无穷多种情况. 为了提高计算效率，本文引入两个参数：$B(n)$和 c. $B(n)$是搜索网格大小的上界，定义适当的 $B(n)$是很重要的. $B(n)$太高，将会导致每一个点具有一个网格，进而使得变量不相关时，MIC 的取值却不为零；$B(n)$太低，意味着只能搜索一些简单的模型. 本文默认 $B(n)$取值为 $n^{0.6}$. 而另一个参数 c 则与网格划分方式相关. 在给定 x 列、y 行的情况下，可先将所有点按照纵轴均分为 y 行(如图 3(a)所示)，并将列划分为 $c\cdot x$ 等份(大量实验表明，$c=15$ 为最佳值)(如图 3(b)所示)，再将 $c\cdot x$ 等份合并为 x 列(如图 3(c)所示)；类似的，将所有点按照横轴均分为 x 列，并将行划分为 $c\cdot y$ 等份($c=15$)，再将 $c\cdot y$ 等份合并为 y 行；在遍历所有可能的划分情况下，计算给定 x 列、y 行情况下的最大互信息 $I^*(D,x,y)$. 需要注意的是，这里提到的“均分”是指样本点个数的均分，而非样本取值的均分. 以 $x=2,y=2$ 为例，如图 3 所示.

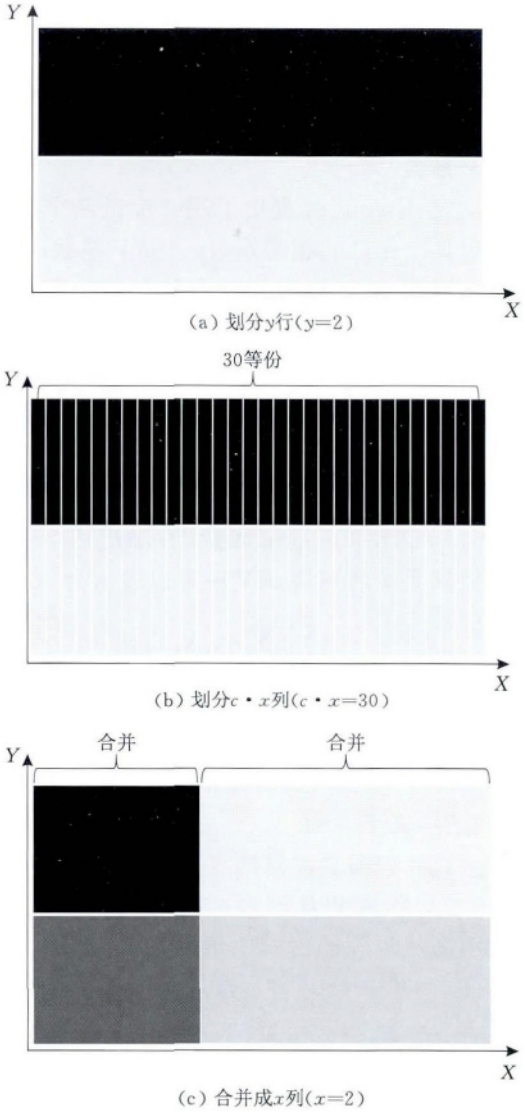

图 3　MIC 的网格划分示意图

根据 MIC 的定义，可以证明其具有如下性质.

(1) MIC 是每个 $I^*(D,x,y)$标准化后的最大值，因此其值在$[0,1]$；

(2) 由于互信息具有对称性，因此，$MIC(X,Y)=MIC(Y,X)$；

(3) 因为 I_G的取值仅依赖于数据点的排序分布，故 MIC 在保序变换下具有不变性；

(4) MIC 的极限性质，即当样本 $n\to\infty$时：

① 对于无噪音的非常数函数关系，$MIC\to 1$；

② 对于无噪音关系(包括无噪音函数的叠加)，$MIC\to 1$；

③ 当两个变量独立时，$MIC\to 0$.

本文通过大量的实验说明该方法比经典的皮尔逊相关系数、斯皮尔曼相关系数等方法更细致地描述了两个变量之间的相关关系，尤其均等性是任何已有相关系数都无法替代的.

然而，该论文仅针对两个数值型变量的情况开展了研究. 我们试想，由于任意两类随机变量、两组随机向量之间均可以计算互信息，因而，可以考虑将该方法推广到任意两类变量之间和向量之间的相关性的度量.

4　基于矩阵计算的相关系数

大数据研究中，变量的高维特性是相关分析的另一个重要挑战. 相应地，面向高维变量，基于矩阵计算的相关分析方法成为一种自然的研究范式.

本节用 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$，$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_q)$分别表示 p 维，q 维随机向量. 其中，$X_1,X_2,\cdots,X_p$为 $\boldsymbol{X}$ 的 p 个特征，$Y_1,Y_2,\cdots,Y_q$为 $\boldsymbol{Y}$ 的 q 个特征. 两个随机向量在 n 个样本下的矩阵为 $\boldsymbol{A}=(x_{ij})_{n\times p}$与

$\boldsymbol{B}=(y_{ij})_{n\times q}$. 用 X^i 代表 $\boldsymbol{A}$ 的第 i 行,表示第 i 个样本的 p 个变量(特征)的取值. 用 X_j 代表 $\boldsymbol{A}$ 的第 j 列,表示 n 个样本的第 j 个变量(特征)的取值. 类似地,有 Y^i,Y_j.

4.1 *RV* 系数

1973 年 Escoufier 提出了用于度量两个随机向量之间更具泛化性的相关系数——RV 系数[53].

两个矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的协方差和方差分别定义为 $tr(\boldsymbol{AA}'\boldsymbol{BB}')$, $tr(\boldsymbol{AA}')^2$, $tr(\boldsymbol{BB}')^2$(其中 $tr(\cdot)$是矩阵$(\cdot)$的迹,定义为矩阵主对角线元素的和). 鉴于上述定义,RV 系数以皮尔逊相关系数的方式重新构造,即得

$$RV(\boldsymbol{A},\boldsymbol{B})=\frac{tr(\boldsymbol{AA}'\boldsymbol{BB}')}{\sqrt{tr(\boldsymbol{AA}')^2tr(\boldsymbol{BB}')^2}} \tag{11}$$

从矩阵元素的视角,RV 还可以表示为

$$RV=\frac{\sum_{k=1}^{p}\sum_{l=1}^{q}r^2(X_k,Y_l)}{\sqrt{\sum_{k,l=1}^{p}r^2(X_k,X_l)}\sqrt{\sum_{k,l=1}^{q}r^2(Y_k,Y_l)}} \tag{12}$$

其中 $r(X_k,Y_l)$是 X_k,Y_l的样本相关系数. 因此,RV 是式(7)的广义平方和.

该公式我们需要注意两点:

(1)当矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 同为 $n\times p$ 阶矩阵时,该公式可以理解为从矩阵的内积的角度定义. 矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的内积为$\langle\boldsymbol{A},\boldsymbol{B}\rangle=tr(\boldsymbol{A}'\boldsymbol{B})$,按照式(8),$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的相关系数为

$$r(\boldsymbol{A},\boldsymbol{B})=\frac{tr(\boldsymbol{A}'\boldsymbol{B})}{\sqrt{tr(\boldsymbol{A}'\boldsymbol{A})tr(\boldsymbol{B}'\boldsymbol{B})}} \tag{13}$$

这是最简单的一种矩阵相关系数,RV 系数是对不同型矩阵的改进,用 $\boldsymbol{A}'\boldsymbol{A}$ 和 $\boldsymbol{B}'\boldsymbol{B}$ 分别代替 $\boldsymbol{A}$ 和 $\boldsymbol{B}$,代入式(13)即得式(11). 由于 $\boldsymbol{A}'\boldsymbol{A}$ 和 $\boldsymbol{B}'\boldsymbol{B}$ 是对称矩阵保证了 RV 系数具有旋转不变性.

(2)RV 是测量 $\boldsymbol{A}'\boldsymbol{A}$ 和 $\boldsymbol{B}'\boldsymbol{B}$ 紧密程度的测度,RV 的取值范围是$[0,1]$. 当 RV 越接近 1 时,说明对于这 n 个样本,用 $X(Y)$代替 $Y(X)$越合理.

文献[54]则指出对于小样本而言 RV 相关系数偏高,当样本增大时,RV 将趋近于 0. 本文认为这种误差是由于 $\boldsymbol{AA}'$和 $\boldsymbol{BB}'$对角线元素引起的,由此提出了用 $\boldsymbol{AA}'-\mathrm{diag}(\boldsymbol{AA}')=\widetilde{\boldsymbol{AA}}'$代替 $\boldsymbol{AA}'$,类似定义 $\widetilde{\boldsymbol{BB}}'$,得到修正 RV 系数,记为 RV_{mod}:

$$RV_{\mathrm{mod}}=tr(\widetilde{\boldsymbol{AA}}'\widetilde{\boldsymbol{BB}}')/\sqrt{tr(\widetilde{\boldsymbol{AA}}')^2tr(\widetilde{\boldsymbol{BB}}')^2}.$$

实际上,与 RV 系数相比,RV_{mod}系数还可以衡量变量间的负相关,即 $RV_{\mathrm{mod}}\in[-1,1]$. 从元素角度看,$RV_{\mathrm{mod}}$是将 $r(X_k,Y_l)$修正为

$$r^2_{\mathrm{mod}}(X_k,Y_l)=r^2(X_k,Y_l)-\left(\frac{1}{n-1}\right)^2\sum_{i=1}^{n}(x_{ik}y_{il})^2,$$

则

$$RV_{\mathrm{mod}}=\frac{\sum_{k=1}^{p}\sum_{l=1}^{q}r^2_{\mathrm{mod}}(X_k,Y_l)}{\sqrt{\sum_{k,l=1}^{p}r^2_{\mathrm{mod}}(X_k,X_l)}\sqrt{\sum_{k,l=1}^{q}r^2_{\mathrm{mod}}(Y_k,Y_l)}}.$$

进一步地,文献[55]指出 RV,RV_{mod}是有偏估计,给出了一种无偏估计 RV_{adj}:

$$RV_{\mathrm{adj}}=\frac{\sum_{k=1}^{p}\sum_{l=1}^{q}r^2_{\mathrm{adj}}(X_k,Y_l)}{\sqrt{\sum_{k,l=1}^{p}r^2_{\mathrm{adj}}(X_k,X_l)}\sqrt{\sum_{k,l=1}^{q}r^2_{\mathrm{adj}}(Y_k,Y_l)}},$$

其中,$r^2_{\mathrm{adj}}(X_k,Y_l)=1-\frac{n-1}{n-2}(1-r^2(X_k,Y_l))$.

事实上,RV 系数具有优良的泛化性. 文献[56]给出了 RV 系数的广义模型,即寻找 $p\times s$ 矩阵 $\boldsymbol{L}$ 和 $q\times s$ 矩阵 $\boldsymbol{M}$,使得 $RV(\boldsymbol{AL},\boldsymbol{BM})$达到最大. 以 2.3.3 节的典型相关系数为例,运用矩阵的表现形式,即寻找满足 $\boldsymbol{L}'\boldsymbol{A}'\boldsymbol{AL}=\boldsymbol{E}$ 和 $\boldsymbol{M}'\boldsymbol{B}'\boldsymbol{BM}=\boldsymbol{E}$ 的矩阵 $\boldsymbol{L}$ 和 $\boldsymbol{M}$,使得 $RV(\boldsymbol{AL},\boldsymbol{BM})$达到最大. 可见,典型相关系数是 RV 系数广义模型的特例.

RV 系数广义模型泛化性的优势在于可以根据需要设计矩阵 $\boldsymbol{L}$ 和 $\boldsymbol{M}$. 实际上,经典的线性多元统计分析方法(如主成分分析、典型相关分析、多元回归分析、线性判别)均可视为 RV 系数广义模型的特例,其区别在于 $\boldsymbol{L}$、$\boldsymbol{M}$ 的构造条件不同. 更多细节请参看文献[56].

4.2 协方差矩阵的改进

大数据时代,诸多领域产生了大量的高维数据. 例如:基因数据、天体物理数据、图像数据等等. 这些数据有一个共同的特点是样本的维数(特征)远远大于样本个数,即特征要素和样本量可能都趋于无穷大的增长,而特征要素相比于样本量呈指数级增长. 以 Web 文本为例,它的维度(属性)通常可以达到成百上千维,甚至更高;研究文本分类问题时,即便对相关文本进行全部采样,所获得的样本量仍然小于特征维数. 通常,我们将这类问题称之为高维数据问题,即 $p,q\gg n$. 高维数据问题带来的主要挑战是解的不确定性问题,即已知的信息量难以获取唯一解. 因此,诸多学者致力于探索面向高维数据分析的新方法.

传统的高维数据相关分析是以协方差矩阵为基础构造相应的数学模型. 正如 4.1 节中的 $\boldsymbol{AA}'$、$\boldsymbol{BB}'$在中心化后本身就是协方差矩阵. 因此协方差矩阵

的估计的精准性直接影响随机向量相关系数的计算. 然而，经典的样本协方差矩阵的估计方法难以适用于高维数据. 具体而言，不失一般性，设 $E\boldsymbol{X}=0$，$\boldsymbol{X}$ 的协方差矩阵为 $\boldsymbol{\Sigma}_p=E\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}=(\sigma_{ij})_{p\times p}$. 在统计学中，往往用样本协方差矩阵 $\hat{\boldsymbol{\Sigma}}_p=(\hat{\sigma}_{ij})_{p\times p}$ 来估计它. 这里，当 $p<n$ 时，$\hat{\boldsymbol{\Sigma}}_p$ 具有无偏性同时是正定矩阵. 但是，当维数 p 增大或 p/n 大时，样本协方差矩阵不再具有这些性质，现已有随机矩阵理论证明了样本协方差矩阵的这一缺陷[57-58]. 换言之，对于高维数据而言，经典的样本协方差估计不再是协方差矩阵的优良估计.

因此，相关学者研究了适应于高维数据的协方差矩阵估计，估计方法大致可分为两类. 一类是针对变量(特征)具有一定自然顺序的高维数据，如纵向数据，对于这类数据，变量之间间隔越远则相关性越弱. 另一类则是面向变量(特征)不具有自然顺序的高维数据. 对于第一类情况可以正则化协方差矩阵，比如条带(Banding)估计[59]或渐变(Tapering)估计[60]. 对于第二类情况，由于变量之间不存在自然顺序，也就无法排序(比如基因表达序列)，这时需要应用变量排序的不变性阈值估计方法[61].

文献[59]针对第一类情况，运用条带估计方法直接将样本协方差矩阵的每一项正则化，即

$$\hat{\boldsymbol{\Sigma}}_p\equiv\hat{\boldsymbol{\Sigma}}_{k,p}=([\sigma_{ij}I(|i-j|\leq k)])_{p\times p}\quad(14)$$

其中，$0\leq k<p$(k 为条带宽度，表示矩阵的稀疏程度)，$I(\cdot)$是示性函数，

$$I(|i-j|\leq k)=\begin{cases}1, & |i-j|\leq k\\ 0, & |i-j|>k\end{cases}.$$

通过条带估计方法，可将原协方差矩阵转化为稀疏矩阵，矩阵形如图 4 所示.

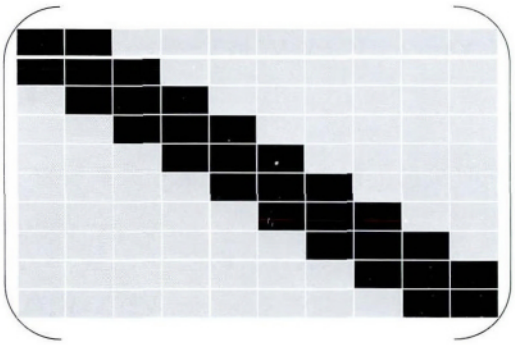

图 4　条带协方差矩阵估计示意图

图中黑色方块代表该位置数据非零，灰色方块代表该位置数据为零. 进一步地，本文在谱范数的基础上证明了，这种正则化方法可以达到一个理想的收敛速度. 换言之，协方差的估计值与真实值之间的距离与 $\left(\frac{\log p}{n}\right)^{\alpha/(2(\alpha+1))}$ 同阶，其中 α 是刻画矩阵稀疏的参数.

实际上，对于第一类情况而言，变量之间间隔越远，其相关性呈现逐渐减弱的特性. 显然，文献[59]基于示性函数的条带估计方法并未有效刻画这一特性. 因此，文献[60]对文献[59]进行了改进，通过设置权重 $w_{ij}=k_h^{-1}\{(k-|i-j|)_+-(k_h-|i-j|)_+\}$(其中，$k_h=k/2$，不失一般性可假设 k 是偶数，如图 5 所示)，从而定义

$$\hat{\boldsymbol{\Sigma}}=\hat{\boldsymbol{\Sigma}}_k=(w_{ij}\hat{\sigma}_{ij})_{p\times p}\quad(15)$$

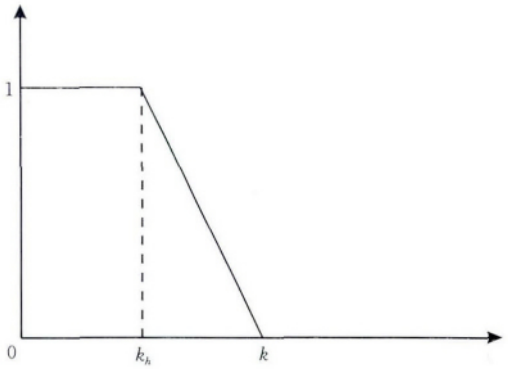

图 5　关于 $|i-j|$ 的权重函数图

基于渐变估计方法构造的稀疏矩阵如图 6 所示，文献[60]进一步证明了这种估计具有理想的收敛下界.

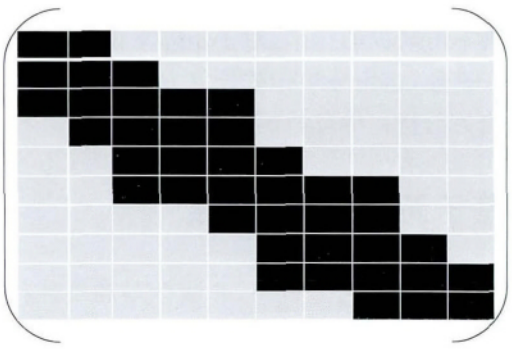

图 6　渐变协方差矩阵估计示意图

事实上，在相关系数的度量中(如典型相关分析)，有时需要对协方差矩阵的逆进行估计. 相应地，文献[61-62]围绕第一种情况下协方差矩阵的逆估计开展了研究，细节参看文献.

对于第二类情况而言，同样围绕稀疏矩阵的构造开展研究，提出了硬阈值估计[63](Hard Thresholding Estimation)、软阈值估计[64](Soft Thresholding Estimation)、平滑截尾绝对偏差估计[65](Smoothly

Clipped Absolution Deviation Estimation，SCADE）和极小极大凹性估计[66]（Minimax Concavity Estimation，MCE）. 事实上，硬阈值与软阈值各具优缺点. 硬阈值估计尽管是间断函数，但其具有无偏性；而软阈值估计满足连续性，但其估计偏差较大. 因此，在实际应用中，经常使用既保证连续（软阈值的优点）又更容易无偏（硬阈值的优点）的平滑截尾绝对偏差估计和极小极大凹性估计. 通过上述估计则将原协方差矩阵转化为如下稀疏矩阵（如图 7 所示）.

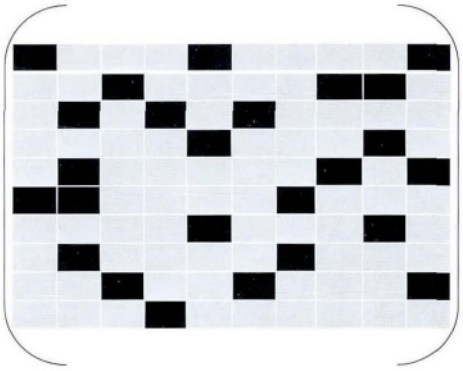

图 7 第二类协方差矩阵估计示意图

综合来看，基于矩阵计算的相关系数试图充分利用矩阵工具，将高维数据的关系罗列在矩阵表上，借助大数据高维稀疏的特点，利用正则化方法简化数据表从而进行数据相关分析. 从技术上来讲，基于矩阵计算的相关系数本质上仍然是皮尔逊相关系数的构造思想，但通过矩阵理论将简化协方差矩阵表，进而降低计算复杂度. 进一步地，分布式计算是大数据处理不可缺少的技术，对于矩阵来说，探索数据的块分布现象，运用分块矩阵计算方法，可为高维大数据的相关分析提供可行的求解技术.

5 基于距离的相关系数

在大数据相关分析中，非线性与高维性往往是并存的. 2007 年 Székely 从特征函数的距离视角考察了两个随机向量之间的非线性相关系数[27]，为高维数据的非线性相关性分析提供了有效的度量准则.

由于 Y 的分布函数 $F(Y)$ 和已知 X 的条件下 Y 的条件分布函数 $F(Y\mid X)$ 的差异程度代表了 X 与 Y 相关程度的大小，为了便于计算，通常用 X 与 Y 的联合分布函数 $F(X,Y)$ 与各自分布函数乘积 $F(X)F(Y)$ 之差来代替. 因为特征函数可以由分布函数唯一确定，而特征函数又能与距离相联系，故而，用 X 与 Y 的联合特征函数 $f_{XY}(\boldsymbol{s},\boldsymbol{t})$ 与各自特征函数乘积 $f_X(\boldsymbol{s})f_Y(\boldsymbol{t})$ 之差来作为最终衡量相关程度的指标，具体细节如下：

对于实数向量 $\boldsymbol{s}=(s_1,s_2,\cdots,s_p)\in R^p$，它的欧氏范数为 $\|\boldsymbol{s}\|=(s_1^2+s_2^2+\cdots+s_p^2)^{1/2}$. 进一步定义 $\langle\boldsymbol{s},\boldsymbol{X}\rangle=s_1X_1+s_2X_2+\cdots+s_pX_p$ 为 $\boldsymbol{s}$ 与 $\boldsymbol{X}$ 的内积. 同理，可以定义 $\boldsymbol{t}=(t_1,t_2,\cdots,t_q)\in R^q$，$\|\boldsymbol{t}\|$，$\langle\boldsymbol{t},\boldsymbol{Y}\rangle$. 在此基础上，随机向量 $(\boldsymbol{X},\boldsymbol{Y})$ 的联合特征函数定义为

$$f_{\boldsymbol{XY}}(\boldsymbol{s},\boldsymbol{t})=E\exp[\mathrm{i}\langle\boldsymbol{s},\boldsymbol{X}\rangle+\mathrm{i}\langle\boldsymbol{t},\boldsymbol{Y}\rangle],$$

其中，i 为虚数单位. $\boldsymbol{X}$，$\boldsymbol{Y}$ 各自的特征函数为

$$f_{\boldsymbol{X}}(\boldsymbol{s})=f_{\boldsymbol{XY}}(\boldsymbol{s},0)=E\exp[\mathrm{i}\langle\boldsymbol{s},\boldsymbol{X}\rangle],$$

$$f_{\boldsymbol{Y}}(\boldsymbol{t})=f_{\boldsymbol{XY}}(0,\boldsymbol{t})=E\exp[\mathrm{i}\langle\boldsymbol{t},\boldsymbol{Y}\rangle].$$

我们知道，对于任意的 $\boldsymbol{s}\in R^p$，$\boldsymbol{t}\in R^q$，当且仅当 $f_{\boldsymbol{XY}}(\boldsymbol{s},\boldsymbol{t})=f_{\boldsymbol{X}}(\boldsymbol{s})f_{\boldsymbol{Y}}(\boldsymbol{t})$，$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 独立. 这意味着若等式 $f_{\boldsymbol{XY}}(\boldsymbol{s},\boldsymbol{t})=f_{\boldsymbol{X}}(\boldsymbol{s})f_{\boldsymbol{Y}}(\boldsymbol{t})$ 成立，则 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 独立，也就意味着 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 不相关；若 $f_{\boldsymbol{XY}}(\boldsymbol{s},\boldsymbol{t})=f_{\boldsymbol{X}}(\boldsymbol{s})f_{\boldsymbol{Y}}(\boldsymbol{t})$ 不成立，则 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 不独立，即具有线性或非线性关系. 依据此性质，设计新的不局限于线性关系的距离协方差和方差.

定义随机向量 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的距离协方差 $V(\boldsymbol{X},\boldsymbol{Y})$，方差 $V^2(\boldsymbol{X})$，$V^2(\boldsymbol{Y})$. 公式如下：

$$\begin{aligned}V^2(\boldsymbol{X},\boldsymbol{Y})&=\|f_{\boldsymbol{XY}}(\boldsymbol{s},\boldsymbol{t})-f_{\boldsymbol{X}}(\boldsymbol{s})f_{\boldsymbol{Y}}(\boldsymbol{t})\|_w^2\\&=\int_{R^{p+q}}|f_{\boldsymbol{XY}}(\boldsymbol{s},\boldsymbol{t})-f_{\boldsymbol{X}}(\boldsymbol{s})f_{\boldsymbol{Y}}(\boldsymbol{t})|^2w(\boldsymbol{s},\boldsymbol{t})\mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{t},\end{aligned}$$

$$V^2(\boldsymbol{X})=V^2(\boldsymbol{X},\boldsymbol{X})=\|f_{\boldsymbol{XX}}(\boldsymbol{s},\boldsymbol{t})-f_{\boldsymbol{X}}(\boldsymbol{s})f_{\boldsymbol{X}}(\boldsymbol{t})\|_w^2,$$

$$V^2(\boldsymbol{Y})=V^2(\boldsymbol{Y},\boldsymbol{Y})=\|f_{\boldsymbol{YY}}(\boldsymbol{s},\boldsymbol{t})-f_{\boldsymbol{Y}}(\boldsymbol{s})f_{\boldsymbol{Y}}(\boldsymbol{t})\|_w^2,$$

其中，$w(\boldsymbol{s},\boldsymbol{t})$ 是权重函数，它的选择需要满足 3 个条件，即保证被积函数可积性；$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 独立时，相关系数为零；$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 同比例变化时，相关系数不变. 在此基础上，定义距离相关系数

$$R^2(\boldsymbol{X},\boldsymbol{Y})=\begin{cases}\dfrac{V^2(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{V^2(\boldsymbol{X})V^2(\boldsymbol{Y})}}, & V^2(\boldsymbol{X})V^2(\boldsymbol{Y})>0\\0, & V^2(\boldsymbol{X})V^2(\boldsymbol{Y})=0\end{cases}\tag{16}$$

同时给出了相应的样本距离相关系数为

$$R_n^2(\boldsymbol{X},\boldsymbol{Y})=\begin{cases}\dfrac{V_n^2(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{V_n^2(\boldsymbol{X})V_n^2(\boldsymbol{Y})}}, & V_n^2(\boldsymbol{X})V_n^2(\boldsymbol{Y})>0\\0, & V_n^2(\boldsymbol{X})V_n^2(\boldsymbol{Y})=0\end{cases}\tag{17}$$

其中：

$$V_n^2(\boldsymbol{X},\boldsymbol{Y})=\frac{1}{n^2}\sum_{k,l=1}^{n}u_{kl}v_{kl},$$

$$V_n^2(\boldsymbol{X})=V_n^2(\boldsymbol{X},\boldsymbol{X})=\frac{1}{n^2}\sum_{k,l=1}^{n}u_{kl}^2,$$

$$V_n^2(\boldsymbol{Y})=V_n^2(\boldsymbol{Y},\boldsymbol{Y})=\frac{1}{n^2}\sum_{k,l=1}^{n}v_{kl}^2,$$

$$u_{kl}=a_{kl}-\bar{a}_{k\cdot}-\bar{a}_{\cdot l}+\bar{a}_{\cdot\cdot},\ k,l=1,\cdots,n,$$

$$v_{kl}=b_{kl}-\bar{b}_{k\cdot}-\bar{b}_{\cdot l}+\bar{b}_{\cdot\cdot},\ k,l=1,\cdots,n,$$

$$a_{kl}=|X^k-X^l|_p,\ b_{kl}=|Y^k-Y^l|_q,$$

$$\bar{a}_{k\cdot}=\frac{1}{n}\sum_{l=1}^{n}a_{kl},\ \bar{a}_{\cdot l}=\frac{1}{n}\sum_{k=1}^{n}a_{kl},\ \bar{a}_{\cdot\cdot}=\frac{1}{n^2}\sum_{k,l=1}^{n}a_{kl},$$

$$\bar{b}_{k\cdot}=\frac{1}{n}\sum_{l=1}^{n}b_{kl},\ \bar{b}_{\cdot l}=\frac{1}{n}\sum_{k=1}^{n}b_{kl},\ \bar{b}_{\cdot\cdot}=\frac{1}{n^2}\sum_{k,l=1}^{n}b_{kl}.$$

需要强调的是，相关结论是在原假设 $f_{XY}=f_Xf_Y$ 的基础上推导出来的. 因此，在其实际应用中需要进行假设检验. 事实上，本文也给出了拒绝 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 独立假设的拒绝域 w：

$$w=\left\{V_n\,\middle|\,\sqrt{nV_n^2/S}\geq\Phi^{-1}(1-\alpha/2)\right\},$$

其中，$S=\frac{1}{n^2}\sum_{k,l=1}^{n}|X^k-X^l|_p\frac{1}{n^2}\sum_{k,l=1}^{n}|Y^k-Y^l|_q$，$\Phi(\cdot)$ 是标准正态分布的分布函数. 文献进一步证明了渐进性质，即若 $E|X|_p<\infty$，$E|Y|_q<\infty$，则 $\lim\limits_{n\to\infty}V_n(\boldsymbol{X},\boldsymbol{Y})=V(\boldsymbol{X},\boldsymbol{Y})$，$\lim\limits_{n\to\infty}R_n(\boldsymbol{X},\boldsymbol{Y})=R(\boldsymbol{X},\boldsymbol{Y})$ 都几乎处处成立，进而保证了样本估计总体的合理性. 需要进一步说明的是，这里的距离相关系数是一个有偏估计，当维数增大时，R_n 一致趋于 1. 因此，文献[67]进行了改进，得到了距离相关系数的无偏估计.

综合来看，以距离为基础的相关系数，从特征函数视角构造了相关性刻画的度量方法，因而，具有两个显著优点：(1) 所度量的相关性，不仅仅局限于线性相关关系；(2) 可以度量任意两个不同维数的随机向量的相关性. 基于上述优点，其被广泛应用于机器学习[68]、气候变化[69]、地理电离层时间序列分析[70]、核化学[71]等领域. 但是，距离相关系数涉及到高维向量间距离计算及矩阵点乘运算，计算耗时也非常巨大. 应用矩阵理论对数据降维或对矩阵分块，进而提高计算效率则是距离相关系数的未来研究方向.

6　研究展望

大数据表现出的非线性、高维性、海量性(大规模、快速增长)等特征，为大数据相关分析提出了严峻的挑战. 围绕这些挑战，各种大数据相关分析方法也不断涌现. 本文面向非线性、高维性问题，从统计相关分析、互信息、矩阵计算、距离 4 个方面对现有研究成果进行了分析，其特点比较见表 3. 此外，在探讨大数据的伪相关识别问题中，提出的保持整体性质的拆分与融合策略可为“海量性”问题提供新的解决途径.

表 3　大数据相关分析方法比较

方法	特点	发展
统计学中的相关分析理论	线性相关分析	典型相关系数的非线性、高维性推广
基于互信息的相关分析	非线性相关分析	*MIC* 的高维推广
基于矩阵计算的相关分析	线性、高维相关分析	分块矩阵计算
基于距离的相关分析	非线性、高维相关分析	分块矩阵计算

从现有研究成果来看，典型相关分析、基于矩阵计算的相关分析、距离相关系数围绕高维随机变量的相关分析开展了研究，然而，关于计算效率的问题仍显困难；就非线性而言，典型相关分析、距离相关系数也进行了探索，特别是基于互信息的相关分析，由于其具有的通用性、均等性两个重要性质，表现出良好的非线性相关的刻画能力，因而备受关注，但是，其仅仅是针对两个随机变量进行研究；就大数据的海量性(大规模、快速增长)而言，关于相关分析的重要研究成果仍不多见. 总体来看，尽管大数据相关分析已经取得重要进展，但围绕大数据的复杂特征，仍有诸多挑战亟待探索与解决.

(1) 高维数据的相关分析. 在探索随机向量间相关性度量的研究中，随机向量的高维特征导致巨大的矩阵计算量，这也成为高维数据相关分析中的关键困难问题. 面临高维特征空间的相关分析时，数据可能呈现块分布现象，如医疗数据仓库、电子商务推荐系统. 探测高维特征空间中是否存在数据的块分布现象，并发现各数据块对应的特征子空间，本质上来看，这是基于相关关系度量的特征子空间发现问题. 结合子空间聚类技术，发现相关特征子空间，并以此为基础，探索新的分块矩阵计算方法，有望为高维数据相关分析与处理提供有效的求解途径. 然而，面临的挑战在于：① 如果数据维度很高、数据表示非常稀疏，如何保证相关关系度量的有效性？② 分块矩阵的计算可以有效提升计算效率，但是，如何对分块矩阵的计算结果进行融合？

(2) 多变量数据的相关分析. 大数据相关分析中，非线性相关关系的度量是一个重要课题. 2011 年在《Science》上发表的论文“Detecting novel associations in large data sets”[26]，给出了两个随机变量之间非线性相关关系的度量准则. 然而，在现实的大数据相关分析中，往往面临多变量情况. 显然，发展多变量非线性相关关系的度量方法是我们面临的一个

重要挑战.

(3)大规模数据的相关分析.大数据时代,相关分析面向的是数据集的整体,因此,试图高效地开展相关分析与处理仍然非常困难.为了快速计算大数据相关性,需要探索数据集整体的拆分与融合策略.显然,在这种"分而治之"的策略中,如何有效保持整体的相关性,则是大规模数据相关分析中必须解决的关键问题.作者的研究成果[7]给出了一种可行的拆分与融合策略,文献[72]也指出随机拆分策略是可能的解决路径.当然,在设计拆分与融合策略时,如何确定样本子集规模、如何保持子集之间的信息传递、如何设计各子集结果的融合原理等都是具有挑战性的问题.

(4)增长性数据的相关分析.大数据中,数据呈现快速增长特征.更为重要的是,诸如电商精准推荐等典型增长性数据相关分析任务,迫切需要高效的在线相关分析技术.就增长性数据而言,可表现为样本规模的增长、维数规模的增长以及数据取值的动态更新.显然,对增长性数据相关分析而言,特别是对在线相关分析任务而言,每次对数据整体进行重新计算对于用户而言是难以接受的,更难以满足用户的实时性需求.我们认为,无论何种类型的数据增长,往往与原始数据集存在某种的关联模式,利用已有的关联模式设计具有递推关系的批增量算法是一种行之有效的计算策略.作者的研究成果[73-74]面向数据降维问题,围绕增长性数据开展了批增量算法研究,取得了显著成效.那么,面向大数据的相关分析任务,探测增长性数据与原始数据集的关联模式,进而发展具有递推关系的高效批增量算法,可为增长性数据相关分析尤其是在线相关分析提供有效的技术手段.

综合来看,尽管大数据相关分析的研究成果尚不丰富,但是,围绕非线性、高维性、海量性等复杂特征的现有进展已经为大数据相关分析提供了一个基本的研究框架,为更多有价值研究的不断涌现奠定了重要基础.然而,需要指出的是,在大数据相关分析的现有研究中仍然具有一个共性困难,即可计算性挑战.就大数据的可计算问题而言,以 MapReduce 为代表的非关系数据管理技术为大数据分析与处理提供了一种并行处理架构,并围绕频繁序列模式挖掘[75]、聚类[76]等数据挖掘任务开展了高效计算方法研究.但是,围绕大数据相关分析可计算性的研究仍然很少,从现有成果来看,仅仅进行了一些初步的探索.期刊《Big Data Research》于 2014 年推出的大数据可扩展计算(Scalable Computing for Big Data)专辑中的论文[77],基于相关图的极大团挖掘方法提出了一种高维相关子空间的搜索策略,避免了传统 Apriori 算法中具有较高计算耗时的逐层搜索模式,进而为大数据中多变量相关分析提供了一种快速计算方法.文献[78]则基于云计算架构,通过云运算将各端点云合并为中心云进而产生中心云滴,在此基础上,以中心云滴为大数据的不确定性复原小样本并针对其进行典型相关分析运算,进而提出了具有较高计算效率的大数据典型相关分析的云模型方法.当然,值得进一步强调的是,可计算性挑战作为大数据分析与挖掘中普遍存在的共性难题,可以预见,其必然受到更多的关注.Kleiner 等人在 ICML2012 上发表的研究成果"The big data Bootstrap"[79]借鉴现有的技术手段,运用 Bootstrap 方法给出了一种大数据的重采样策略来实现大数据的高效计算.进一步地,通过分析大数据的多层次/多粒度特性[80-81],基于粒计算理论与方法的高效算法研究也逐渐受到重视.文献[82]通过对数据空间与特征空间的粒化,运用集成学习技术开展了大规模数据的聚类分析研究;文献[83]则提出了一种利用决策树思想的大数据分解方法,进而在每个分解的数据粒上分别学习 SVM 分类器,极大提高了 SVM 的学习效率;作者的研究成果[84],基于信息粒构造了目标概念的正向近似,进而提出了一种有效的特征选择加速器,显著提高了计算性能.总体来看,上述成果为大数据分析的可计算提供了多角度的研究路径,进一步地,如何实现特定数据分析任务与MapReduce、粒化策略等大数据可计算手段的有机结合、如何平衡算法效率与求解结果的精度进而高效获得可行的满意近似解等问题则将是探索可计算性难题的新挑战,这些问题的有效解决也将为大数据相关分析提供强有力的技术支撑.

7 总 结

大数据相关分析作为探寻与发现事物内在规律的重要"导航"工具,其自然成为大数据分析与挖掘的关键科学问题.本文在对统计学中的经典相关分析理论进行归纳、总结的基础上,从大规模数据的通用性和均等性视角阐述了基于互信息的两个变量间非线性相关分析理论;从高维数据可行计算的角度分析了基于矩阵计算的相关系数;从非线性、高维性数据的复杂结构方面解析了基于距离的相关系数.

进一步地，从高维数据相关分析、多变量数据相关分析、大规模数据相关分析、增长性数据相关分析及其可计算性方面提出了未来的研究方向.

当然，大数据相关分析的研究尚处于起步阶段，可以预见，在未来的大数据研究中，具有快捷、清晰、高效探测事物内在关系、规律功能的大数据相关分析将涌现大量的重要研究成果. 本文针对大数据相关分析的综述研究希望能够为关注大数据相关分析理论与应用的研究者与实践领域专家提供借鉴.

参 考 文 献

[1] Big data. Nature, 2008, 455(7209): 1-136

[2] Dealing with data. Science, 2011, 331(6018): 649-729

[3] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. USA: Mckinsey Global Institute, White Paper, 2011

[4] World Economic Forum. Big data, big impact: New possibilities for international development. World Economic Forum, 2012

[5] Li Guo-Jie, Cheng Xue-Qi. Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657(in Chinese)
(李国杰，程学旗. 大数据研究:未来科技及经济社会发展的重大战略领域:大数据的研究现状与科学思考. 中国科学院院刊, 2012, 27(6): 647-657)

[6] Wang Shan, Wang Hui-Ju, Qin Xiong-Pai, Zhou Xuan. Architecting big data: Challenges, studies and forecast. Chinese Journal of Computers, 2011, 34(10): 1741-1752(in Chinese)
(王珊，王会举，覃雄派，周煊. 架构大数据：挑战、现状与展望. 计算机学报, 2011, 34(10): 1741-1752)

[7] Liang J Y, Wang F, Dang C Y, Qian Y H. An efficient rough feature selection algorithm with a multi-granulation view. International Journal of Approximate Reasoning, 2012, 53: 912-926

[8] Zhou Hang-Xing, Chen Song-Can. Ordinal discriminative canonical correlation analysis. Journal of Software, 2014, 25(9): 2018-2025(in Chinese)
(周航星，陈松灿. 有序判别典型相关分析. 软件学报, 2014, 25(9): 2018-2025)

[9] Huo Zheng, Meng Xiao-Feng. A survey of trajectory privacy-preserving techniques. Chinese Journal of Computers, 2011, 34(10): 1820-1830(in Chinese)
(霍峥，孟小峰. 轨迹隐私保护技术研究. 计算机学报, 2011, 34(10): 1820-1830)

[10] Meng Xiao-Feng, Gao Hong. Introduction of big data subject. Journal of Software, 2014, 25(4): 691-692(in Chinese)
(孟小峰，高宏. 大数据专题前言. 软件学报, 2014, 25(4): 691-692)

[11] Chen En-Hong, Yu Jian. Introduction of big data analytics. Journal of Software, 2014, 25(9): 1887-1888(in Chinese)
(陈恩红，于剑. 大数据分析专刊前言. 软件学报, 2014, 25(9): 1887-1888)

[12] CCF Task Force on Big Data. China Big Data technology and industrial development white paper. China Computer Federation, 2013(in Chinese)
(中国计算机学会大数据专家委员会. 中国大数据技术与产业发展白皮书. 中国计算机学会, 2013)

[13] CCF Task Force on Big Data, The Big Data Industrial Alliance of China National Zhongguancun Science Park. China Big Data Technology and Industry Development Report. Beijing: China Machine Press, 2014(in Chinese)
(中国计算机学会大数据专家委员会，中关村大数据产业联盟. 中国大数据技术与产业发展报告(2014). 北京：机械工业出版社, 2014)

[14] Zhao Guo-Dong, Yi Huan-Huan, Mi Wan-Jun, E Wei-Nan. Big Data Era Historical Opportunity: Industrial Transformation and Data Science. Beijing: Tsinghua Press, 2013(in Chinese)
(赵国栋，易欢欢，糜万军，鄂维南. 大数据时代的历史机遇：产业变革与数据科学. 北京：清华出版社, 2013)

[15] Galton F. Co-relations and their measurement, chiefly from anthropometric data. Proceedings of the Royal Society of London, 1888, 45: 135-145

[16] Mayer-Schonberger V, CuKier K. Big Data: A Revolution That Will Transform How We Live, Work and Think. New York: Eamon Dolan/Houghton Mifflin Harcourt, 2013

[17] Linden G, Smith B, York J. Amazon. com Recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 2003, 7(1): 76-80

[18] Freyer D A, Hsieh Y H, Levin S R, et al. Google flu trends: Correlation with emergency department influenza rates and crowding metrics. Clinical Infectious Diseases, 2012, 54(4): 463-469

[19] Lü L Y, Medo M, Yeung C H, et al. Recommender systems. Physics Reports, 2012, 519: 1-49

[20] Acquisti A, Gross R. Predicting social security numbers from public data. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(27): 10975-10980

[21] Fan J Q, Liu H. Statistical analysis of big data on pharmacogenomics. Advanced Drug Delivery Reviews, 2013, 65(7): 987-1000

[22] Lu X, Bengtsson L, Holme P. Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(29): 11576-11581

[23] Fang Z Y, Fan X W, Chen G. A study on specialist or special disease clinics based on big data. Frontiers of Medicine, 2014, 8(3): 376-381

[24] Pearson K. Mathematical contributions to the theory of evolution (Ⅲ): Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 1895, 187: 253-318

[25] Martínez-Gómez E, Richards M T, Richards D St P. Distance correlation methods for discovering associations in large astrophysical databases. The Astrophysical Journal, 2014, 781(1): 39-50

[26] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets. Science, 2011, 334: 1518-1524

[27] Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. The Annals of Statistics, 2007, 35(6): 2769-2794

[28] Aspremont A, Ghaoui L E, Jordan M I, Lanckriet G R G. A direct formulation for sparse PCA using semidefinite programming. Society for Industrial and Applied Mathematics, 2007, 49(3): 434-448

[29] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288

[30] Upton G, Cook L. A Dictionary of Statistics. Oxford: Oxford University Press, 2008

[31] Goodmam L A, Kruskal W H. Measure of association for cross classifications. Journal of the American Statistical Association, 1954, 49(268): 732-764

[32] Goodmam L A, Kruskal W H. Measure of association for cross classifications, Ⅱ: Further discussion and references. Journal of the American Statistical Association, 1959, 54(285): 123-163

[33] Goodmam L A, Kruskal W H. Measure of association for cross classifications, IV: Simplification of asymptotic variances. Journal of the American Statistical Association, 1972, 67(338): 415-421

[34] Spearman C. The proof and measurement of association between two things. The American Journal of Psychology, 1904, 15(1): 72-101

[35] Kendall M G. A new measure of rank correlation. Biometrika, 1938, 30(1/2): 81-93

[36] Kowalski C J. On the effects of non-normality on the distribution of the sample product moment correlation coefficient. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1972, 21(1): 1-12

[37] Gnanadesikan R, Kettenring J R. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 1972, 28(1): 81-124

[38] Shoemaker L H, Hettmansperger T P. Robust estimates and tests for the on- and two-sample scale models. Biometrika, 1982, 69(1): 47-53

[39] Wilcox R R. Introduction to Robust Estimation and Hypothesis Testing. San Diego: Academic Press, 1997

[40] Rodgers J L, Nicewander W A. Thirteen ways to look at the correlation coefficient. The American Statistician, 1988, 42(1): 59-66

[41] Hotelling H. Relations between two sets of variates. Biometrika, 1936, 28(3/4): 321-377

[42] Sun Ting-Kai, Chen Song-Can. A survey on canonical correlation analysis//Zhou Zhi-Hua, Wang Jue eds. Machine Learning and Application. Beijing: Tsinghua University Press, 2007: 85-108(in Chinese)
(**孙廷凯，陈松灿. 典型相关分析研究进展//周志华，王珏编. 机器学习及其应用. 北京：清华大学出版社**，2007：85-108)

[43] Melzer T, Reiter M, Bischof H. Appearance models based on kernel canonical correlation analysis. Pattern Recognition, 2003, 36: 1961-1971

[44] Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance, 2003, 10(5): 603-621

[45] Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology, 2005, 4(1): Article 32

[46] Cruz-Cano R, Lee M L T. Fast regularized canonical correlation analysis. Computational Statistics and Data Analysis, 2014, 70: 88-100

[47] Yin X. Canonical correlation analysis based on information theory. Journal of Multivariate Analysis, 2004, 91(2): 161-176

[48] Sun Ping, Xu Zong-Ben, Shen Jian-Zhong. Nonlinear canonical correlation analysis for discrimination based on kernel methods. Chinese Journal of Computers, 2004, 27(6): 789-795(in Chinese)
(**孙平，徐宗本，申建中. 基于核化原理的非线性典型相关判别分析. 计算机学报**，2004，27(6)：789-795)

[49] Jia C, Wei S H. Convergence rate of kernel canonical correlation analysis. Science China: Mathematics, 2011, 54(10): 2161-2170

[50] Fan J Q, Han F, Liu H. Challenges of Big Data analysis. National Science Review, 2014, 1(2): 293-314

[51] Sapsford R, Jupp V. Data Collection and Analysis. London: SAGE in Association with the Open University, 2006

[52] Shannon C E. A mathematical theory of communication. The Bell System Technical Journal, 1948, 27(4): 379-423, 623-656

[53] Escoufier Y. Le traitement des variables vectorielles. Biometrics, 1973, 29(4): 751-760

[54] Smilde A K, Kiers H A L, Bijlsma S, et al. Matrix correlations for high-dimensional data: The modified RV-coefficient. Bionformatics, 2009, 25(3): 401-405

[55] Mayer C D, Lorent J, Horgan G W. Exploratory analysis of multiple omics datasets using the adjusted RV coefficient. Statistical Applications in Genetics and Molecular Biology, 2011, 10(1): 1-27

[56] Robert P, Escoufier Y. A unifying tool for linear multivariate statistical methods: The RV-coefficient. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1976, 25(3): 257-265

[57] Geman S. A limit theorem for the norm of random matrices. The Annals of Probability, 1980, 8(2): 252-261

[58] Yin Y Q, Bai Z D, Krishnaiah P R. On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. Probability Theory Related Fields, 1988, 78(4): 509-521

[59] Bickel P J, Levina E. Regularized estimation of large covariance matrices. The Annals of Statistics, 2008, 36(1): 199-227

[60] Cai T T, Zhang C H, Zhou H H. Optimal rates of covariance matrix estimation. The Annals of Statistics, 2010, 38(4): 2118-2144

[61] Huang J Z, Liu N P, Pourahmadi M, Liu L X. Covariance selection and estimation via penalised normal likelihood. Biometrika, 2006, 93(1): 85-98

[62] Levina E, Rothman A J, Zhu J. Sparse estimation of large covariance matrices via a nested lasso penalty. The Annals of Applied Statistics, 2008, 2(1): 245-263

[63] Bickel P, Levina E. Covariance regularization by thresholding. The Annals of Statistics, 2008, 36(6): 2577-2604

[64] Fan J Q, Liao Y, Mincheva M. Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2013, 75(4): 603-680

[65] Fan J Q, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 2001, 96(456): 1348-1360

[66] Zhang C H. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 2010, 38(2): 894-942

[67] Székely G J, Rizzo M L. The distance correlation t-test of independence in high dimension. Journal of Multivariate Analysis, 2013, 117: 193-213

[68] Sriperumbudur B K, Fukumizu K, Lanckriet G R G. Universality, characteristic kernels and RKHS embedding of measures. Journal of Machine Learning Research, 2011, 12: 2389-2410

[69] Racherla P N, Shindell D T, Faluvegi G S. The added value to global model projections of climate change by dynamical downscaling: A case study over the continental U. S. using the GISS-ModelE2 and WRF models. Journal of Geophysical Research, 2012, 117(D20118): 1-8

[70] Gromenko O, Kokoszka P, Zhu L, Sojka J. Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. The Annals of Applied Statistics, 2012, 6(2): 669-696

[71] Zhong J, DiDonato N. Hatcher P G. Independent component analysis applied to diffusion-ordered spectroscopy: Separating nuclear magnetic resonance spectra of analytes in mixtures. Journal of Chemometrics, 2012, 26(5): 150-157

[72] Xu Zong-Ben. Some scientific issues in the research of big data. Science & Technology for Development, 2014, 10(1): 66-75(in Chinese)
(徐宗本. 大数据研究的若干科学问题. 科技促进发展, 2014, 10(1): 66-75)

[73] Wang F, Liang J Y, Qian Y H. Attribute reduction: A dimension incremental strategy. Knowledge-Based Systems, 2013, 39: 95-108

[74] Liang J Y, Wang F, Dang C Y, Qian Y H. A group Incremental approach to feature selection applying rough set technique. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 294-308

[75] Miliaraki I, Berbweich K, Gemull R, Zoupanos S. Mind the gap: Large-scale frequent sequence mining//Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York, USA, 2013: 797-808

[76] Ene A, Im S, Moseley B. Fast clustering using MapReduce//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2011: 681-689

[77] Nguyen H V, Müller E, Böm K. A near-linear time subspace search scheme for unsupervised selection of correlation features. Big Data Research, 2014, 1: 37-51

[78] Yang Jing, Li Wen-Ping, Zhang Jian-Pei. Canonical correlation analysis of big data based on could model. Journal on Communications, 2013, 34(10): 121-134(in Chinese)
(杨静, 李文平, 张建沛. 大数据典型相关分析的云模型方法. 通信学报, 2013, 34(10): 121-134)

[79] Kleiner A, Talwalkar A, Sarkar P, Jordan M I. The big data bootstrap//Proceedings of the 29th International Conference on Machine Learning. Edinburgh, UK, 2012: 1759-1766

[80] Friedman N. Inferring cellular networks using probabilistic graphical models. Science, 2004, 303(5659): 799-805

[81] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. Nature, 1998, 393(6684): 440-442

[82] Ye Y M, Wu Q Y, Huang Z X, et al. Stratified sampling for feature subspace selection in random forests for high dimensional data. Pattern Recognition, 2013, 46(3): 769-787

[83] Chang F, Guo C Y, Lin X R. Tree decomposition for large-scale problems. Journal of Machine Learning Research, 2010, 11: 2935-2972

[84] Qian Y H, Liang J Y, Pedrycz W, Dang C Y. Positive approximation: An accelerator for attribute reduction in rough set theory. Artificial Intelligence, 2010, 174(9-10): 597-618

LIANG Ji-Ye, born in 1962, Ph. D., professor. His research interests include granular computing theory, data mining and machine learning.

FENG Chen-Jiao, born in 1977, Ph. D., lecturer. Her research interests include data mining, statistics learning method and big data analysis.

SONG Peng, born in 1979, Ph. D., associate professor. His research interests include intelligent decision and data mining.

Background

As a key technology to explore the inherent relevance of complex things in a quick and efficient way, correlation analysis for big data has become one of the core scientific problems in the area of big data analysis and mining. This paper, based on reviewing traditional correlation analysis methods and correlation analysis for big data, studies the frontier research and challenges about correlation analysis for big data. This paper is supported by NSFC's key program named "Theory and approaches of granular computing for Big Data". Big data is large-scale, multi-mode and incremental. Based on these three characteristics, the program aims at exploring a multi-views granulation mechanism of big data, proposing multi-granulation pattern discovery algorithms of big data, constructing a cross-granulation reasoning mechanism of big data, and developing granular computing theories and methods of big data systematically. It is our wish that this program provides significant theoretical value for big data mining and new techniques for developing big data industry quickly.

As one part of the research of NSFC's Key Program, Granular Computing Theory and Methods for Big Data, which, given the big-scale, multi-mode and ever-growing big data, seeks to explore the multi-perspective mechanism, introduce a new computing method, construct a inference mechanism for big data so as to systematically develop the granular computing theory and method for big data. The study is expected to gain significant theoretical value in the area of big data and provide technological support for the robust development of China's Big Data industry.

High level study is carried out on the high-dimensional data's feature selection, clustering, classification and published in *Artificial Intelligence*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Knowledge and Data Engineering*, *Science in China* and *Chinese Journal of Computers*, etc.