第 1 章

大数据分析概述

大数据（Big Data）是信息技术领域的核心概念之一。在此背景下，人工智能、大模型、数据仓库、数据安全、数据分析与数据挖掘等关键技术，共同构成了实现数据价值的关键技术体系。随着相关技术的发展与应用，大数据分析技术的重要性日益凸显。

1.1 大数据分析背景

1. 大数据的狭隘定义

大数据是指无法在特定时间范围内用规范化手段进行捕获、处理和筛选的数据集合，是需要新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产。

2. 大数据的产生

“大数据”概念的雏形可见于未来学家阿尔文·托夫勒 1980 年的著作《第三次浪潮》。他在书中论断，信息爆炸将成为第三次浪潮的主旋律。2008 年，《自然》杂志的“大数据”专题使其获得科学界正式关注，并于 2009 年后迅速成为行业核心术语。这一趋势的数据基础，在很大程度上由 2004 年兴起的社交媒体所奠定，它将个体用户转换为实时数据源，直接驱动了数据总量的指数级增长。

3. 大数据的特征

- 容量（Volume）：数据集合的巨大规模，从太字节级别起步，是定义大数据的基础。

- 种类（Variety）：数据类型的多样性。
- 速度（Velocity）：获得数据的速度。
- 可变性（Variability）：数据流可能存在持续变化，包括周期性峰值等，这为数据处理流程带来了挑战。
- 真实性（Veracity）：数据的可靠性与质量，准确、干净的数据是获得有效分析结果的保证。
- 复杂性（Complexity）：数据来自多种异构来源，需要进行关联、清洗与整合，管理难度较大。
- 价值（Value）：合理运用大数据，以低成本创造高价值。

4. 大数据的结构

大数据包括结构化、半结构化和非结构化数据，非结构化数据越来越成为数据的主要部分。据互联网数据中心（IDC）的调查报告显示：企业中 80%的数据都是非结构化数据，这些数据每年都按指数增长 60%。

5. 大数据分析

大数据分析技术应用于 IT 管理，其核心逻辑在于：通过融合实时数据流与历史数据，构建有效的分析模型，从而实现对未来运行中断与性能问题的预测与防范。

6. 大数据分析的意义

现在的社会是一个高速发展的社会，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的产物。阿里巴巴创始人马云在演讲中就提到，“未来的时代将不是 IT 的时代，而是 DT 的时代”。DT 就是 Data Technology（数据技术），可以看出大数据对于阿里巴巴集团来说举足轻重。

有人把数据比喻为蕴藏能量的煤矿。煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，而露天煤矿、深山煤矿的挖掘成本又不一样。与此类似，大数据并不在于“大”，而在于“有用”。价值含量、挖掘成本比数量更重要。对于很多行业而言，如何利用好这些大规模的数据是赢得竞争的关键。

1.2 大数据分析的应用

未来将是一个“大数据”引领的智慧科技时代，随着社交网络的逐渐成熟，移动带宽的迅速提升，云计算、物联网应用更加丰富，更多的传感设备、移动终端接入网络，由此产生的数据及其增长速度将比历史上任何时期都要多、要快。

虽然大数据在不同领域有不同的应用，但是总的来说，大数据的应用主要体现在三个方面，分别是分析预测、决策制定和技术创新。同时，大数据在很大程度上推动了人工智能的发展。

1. 分析预测

分析预测是比较早的落地应用之一，同时能够比较直观地获得价值，所以当前大数据的场景

分析依然是比较重要的落地应用。分析预测涉及的行业非常多，比如舆情分析、流感预测、金融预测、销售分析等，随着传统行业信息化改造的推进，数据分析将是比较常见的大数据应用。

2. 决策制定

决策制定通常是大数据应用的重要目的，销售部门需要根据数据分析来制定产品的销售策略，设计部门需要根据数据分析来制定产品的设计策略，生产部门需要根据数据分析来优化生产流程，人事部门需要根据数据分析来衡量员工的工作价值从而制定考核策略，财务部门需要根据数据分析来制定财务策略，等等。通常来说，数据分析一个重要的目的就是制定相应的策略。

3. 技术创新

大数据应用能够全面促进企业创新，这不仅体现在技术创新上，还体现在管理创新上。通过数据能够挖掘出更多关于产品和市场的信息，这些信息会指导企业进行相应产品的设计，以满足市场的需求。同时在企业管理方面，以数据为驱动的管理方式能够极大地调动员工的能动性。

1.3　大数据分析算法

1. 大数据分析与数据分析的区别

大数据分析是指对无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的海量、高增长率和多样化的数据集合，采用新的处理模式以获得更强的决策力、洞察力和流程优化能力。

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息并形成结论，从而对数据进行详细研究和概括总结。

大数据分析的优势是能清楚地阐述数据采集和处理过程以及解读最终结果，同时提出模型的优化和改进之处，以利于提升大数据分析的商业价值。

大数据分析与数据分析的核心区别是处理的数据规模不同，由此导致两个方向的从业者的技能也不同。大数据分析与数据分析的根本区别是分析的思维与所用的工具的不同。

2. 机器学习和数据挖掘的联系与区别

从数据分析的角度来看，数据挖掘与机器学习（Machine Learning，ML）有很多相似之处，但不同之处也十分明显，例如，数据挖掘并没有机器学习探索人的学习机制这一科学发现任务。数据挖掘中的数据分析是针对海量数据进行的，从某种意义来说，机器学习的科学成分更重一些，而数据挖掘的技术成分更重一些。

机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。它专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，并重新组织已有的知识结构，使之不断改善自身的性能。

数据挖掘是从海量数据中获取有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程。数据挖掘用到了大量的机器学习领域的数据分析技术，以及数据库领域的数据管理技术。

机器学习不仅涉及对人类认知学习过程的探索，还涉及对数据的分析处理。实际上，机器学

习已经成为计算机数据分析技术的创新源头之一。由于几乎所有的学科都有数据分析任务，机器学习已经开始影响计算机科学的众多领域，甚至影响计算机科学之外的很多学科。机器学习是数据挖掘中的一种重要工具。然而，数据挖掘不仅要研究、拓展、应用一些机器学习方法，还要通过许多非机器学习技术解决数据仓储、大规模数据、数据噪声等实践问题。机器学习的涉及面很广，常用在数据挖掘上的方法是“从数据中学习”。但机器学习并不局限于数据挖掘，一些机器学习的子领域甚至与数据挖掘关系不大，如增强学习与自动控制等。

3. 统计学与机器学习的联系与区别

统计学和机器学习之间的界定一直很模糊。业界和学界曾一直认为机器学习只是为统计学披了一层光鲜的外衣。但事实上，统计学与机器学习存在区别，统计模型与机器学习也有所不同。机器学习和统计学的主要区别在于它们的目的：机器学习模型旨在实现尽可能准确的预测，而统计模型则旨在推断变量之间的关系。

首先，我们必须明白，统计学和统计建模是不一样的。统计学是对数据的数学研究，除非有数据，否则无法进行统计。统计模型是数据的模型，主要用于推断数据中不同变量间的关系，或预测未来值。通常情况下，两者相辅相成。机器学习通常会牺牲可解释性以获得强大的预测能力。例如，从线性回归到神经网络，尽管解释性变差，但是预测能力却大幅提高。

统计模型与机器学习在线性回归的应用上存在差异。尽管二者在回归分析中使用的方法相似，常被误认为属于同一类算法，但实际上并非如此。这种误解主要源于建模方法的相似性，但它们的目的不同。线性回归是一种统计方法，既可用于训练一个线性回归器，也可通过最小二乘法拟合一个统计回归模型。机器学习（此处特指监督学习）的目标是获得一个可重复用于预测的模型，通常不关注模型的可解释性，而更重视预测结果的准确性；统计建模则更侧重于探究变量之间的关系及其统计显著性，预测只是其附带功能。

4. 统计学与数据挖掘的联系与区别

统计学和数据挖掘有着共同的目标：发现数据中的结构。事实上，由于它们的目标相似，有人认为数据挖掘是统计学的分支。这种看法存在偏差，因为数据挖掘还应用了其他领域的思维、工具和算法，尤其是计算机科学技术，例如数据库技术和机器学习，而且数据挖掘关注的某些领域和统计学家关注的有很大差别。

5. 大数据分析的 10 种统计方法

数据分析师不完全是软件工程师，而应是编程、统计和批判性思维三者的结合体。统计学习是培养现代数据分析师的基础。下面分享 10 种统计方法，任何数据分析师都应该掌握，以更高效地处理大数据集。

1）线性回归

线性回归是一种通过拟合因变量和自变量之间的最佳线性关系来预测目标变量的方法。线性回归主要分为简单线性回归和多元线性回归。简单线性回归使用一个自变量，通过拟合一个最佳线性关系来预测因变量；而多元线性回归使用一个以上的自变量来预测因变量。

2）分类

分类是一种数据挖掘技术，用来将一个整体数据集分成几个类别，以便更准确地进行预测和分析。

3）重采样方法

重采样是从原始数据样本中反复抽样的方法，是一种非参数统计推断方法。重采样是在实际数据的基础上生成唯一的抽样分布。

4）子集选择

子集选择首先确定我们认为与响应有关的 P 个预测因子的一个子集，然后使用该子集的特征通过最小二乘法拟合模型。

5）特征缩减技术

通过对损失函数加入正则项，可在训练求解参数的过程中将影响较小的特征的系数衰减到 0，只保留重要的特征。

6）降维

降维是将估计的 P+1 个系数减少为 M+1 个系数，其中 M 可以将主成分回归描述为从一组大的变量中导出低维度特征集的方法。

7）非线性回归

非线性回归是回归分析的一种形式，在这种分析中，观测数据通过模型参数和因变量的非线性组合函数建模，数据用逐次逼近法进行拟合。

8）树形方法

树形方法可以用于回归和分类问题，涉及将预测空间分层或分段为一些简单的区域。由于分割预测空间的分裂规则可以用树形总结，因此这类方法也被称为决策树方法。

9）支持向量机

支持向量机（Support Vector Machine，SVM）是一种分类技术，简单地说，就是寻找一个超平面以最好地将两类点与最大边界区分开。

10）无监督学习

无监督学习就是在无类别信息的情况下寻找到好的特征。

1.4　大数据分析工具

1. 大数据分析前端展现

用于展现分析的前端开源工具有 JasperSoft、Pentaho、Spagobi、Openi、Birt 等。

用于展现分析的商用分析工具有 Style Intelligence、RapidMiner Radoop、Cognos、BO、Microsoft Power BI、Oracle、MicroStrategy、QlikView、Tableau 等。

国内大数据分析工具有 BDP、国云数据（大数据魔镜）、思迈特、FineBI 等。

2. 大数据分析数据仓库

大数据分析数据仓库有 Teradata AsterData、EMC GreenPlum、HP Vertica 等。

3. 大数据分析数据集市

大数据分析数据集市有 QlikView、Tableau、Style Intelligence 等。

4. 统计分析

统计分析法是通过对研究对象的规模、速度、范围、程度等数量关系的分析研究，认识和揭示事物间的相互关系、变化规律和发展趋势，借以实现对事物的正确解释和预测的一种研究方法。

5. 可视化辅助工具

数据可视化技术的基本思想是将数据库中的每个数据项表示为单个图元元素，大量数据项构成数据图像，并将数据的各属性值以多维形式呈现，从而支持从不同维度观察数据，实现更深入的分析。一旦原始数据以图像形式展现，以此进行决策就变得更加容易。为了满足并超越客户的期望，大数据可视化工具应该具备以下特征：

- 能够处理不同种类的传入数据。
- 能够应用不同种类的过滤器来调整结果。
- 能够在分析过程中与数据集进行交互。
- 能够连接其他软件来接收输入数据，或为其他软件提供输入数据。
- 能够为用户提供协作功能。

下面介绍目前比较实用且流行的 4 种大数据可视化工具，它们提供了上述所有或者部分特征。

1）Jupyter：大数据可视化的一站式平台

Jupyter是一个开源项目，通过十多种编程语言实现大数据分析、可视化和软件开发的实时协作。Jupyter的界面包含代码输入窗口，通过运行输入的代码并基于所选择的可视化技术提供视觉可读的图像。

2）Tableau：AI、大数据和机器学习应用可视化的最佳解决方案

Tableau 是大数据可视化的市场领导者之一，在为大数据操作、深度学习算法和多种类型的 AI 应用程序提供交互式数据可视化方面尤为高效。

3）Google Chart：Google 支持免费且强大的整合功能

Google Chart（谷歌图表）是大数据可视化的最佳解决方案之一，它是完全免费的，并得到了 Google 的大力技术支持。

4）D3.js：以任何用户需要的方式直观地显示大数据

D3.js 代表 Data Driven Document，是一个用于实时交互式大数据可视化的 JavaScript 库。由于 D3.js 并非开箱即用的工具，用户在使用它处理数据之前需要具备扎实的 JavaScript 基础，并且要以一种能被其他人理解的形式呈现结果。除此以外，该库将数据以 SVG 和 HTML 5 格式呈现，所以像 IE 7 和 IE 8 这样的旧式浏览器不能使用 D3.js 的功能。

6. 大数据处理框架

近年来，全球数据呈几何级增长，数据的存储与计算已成为世界级难题。分布式文件系统旨在解决大数据存储问题。下面介绍一些分布式计算框架。

1）Hadoop 框架

Hadoop 是目前全球应用较广泛的大数据工具之一，凭借极高的容错性和极低的硬件价格，在大数据领域占据重要地位。Hadoop 是第一个在开源社区引发高度关注的批处理框架，它提出的 Map 和 Reduce 计算模式简洁而优雅。如今，Hadoop 已经发展为一个庞大的生态圈，实现了大量算法和组件。由于 Hadoop 的计算任务需要在集群的多个节点上多次读写数据，因此在速度上会稍显劣势，但其吞吐量是其他框架所不能匹敌的。

2）Storm 框架

与 Hadoop 的批处理模式不同，Storm 采用的是流计算框架，由 Twitter 开源并托管在 GitHub 上。与 Hadoop 类似的是，Storm 也提出了两个计算角色，分别为 Spout 和 Bolt。

3）Samza 框架

Samza 是一种流计算框架，目前只支持 JVM 语言，在灵活度上略显不足，并且必须和 Kafka 共同使用。相应地，它也继承了 Kafka 的低延时、分区、避免回压等优势。

4）Spark 框架

Spark 是 Hadoop 和 Storm 框架的集合体，是一种混合式的计算框架。它既内置实时流处理能力，又可与 Hadoop 集成，替代其中的 MapReduce；同时，Spark 也能独立部署集群，但仍需依赖 HDFS 等分布式存储系统。Spark 的优势在于其运算速度：与 Storm 类似，Spark 基于内存计算，并在内存不足时可利用磁盘继续运算。实验表明，Spark 的速度可达 Hadoop 的数十倍甚至百倍，且总体成本可能更低。然而，Spark 目前尚未像 Hadoop 那样支持上万节点规模的集群，因此现阶段将 Spark 与 Hadoop 结合使用更为合适。

7. 数据库

数据库可视为电子化的文件柜——存储电子文件的场所，用户可以对文件中的数据进行新增、查询、更新、删除等操作。

8. 数据仓库/商业智能

数据仓库（Data Warehouse，DW 或 DWH）是为企业各级决策制定过程提供全面数据支持的战略性集合。它是一个为分析性报告和决策支持而构建的单一数据存储。数据仓库可为需要业务智能的企业提供业务流程改进指导，并支持对时间、成本、质量和控制等方面的监控。

商业智能（Business Intelligence，BI）又称商业智慧或商务智能，使用现代数据仓库技术、线上分析处理技术、数据挖掘和数据展现技术进行数据分析，以实现商业价值。

伴随数据库技术的提高和数据处理技术的发展以及各行业业务自动化的实现，商业领域产生了大量的业务数据，想要从这些海量数据中提取出真正有价值的信息，将数据转化为知识，以支持商业决策，需要用到能提取和存储有用信息，并支持决策的数据仓库、联机分析处理（On-Line Analysis Processing，OLAP）以及数据挖掘（Data Mining，DM）等技术。因此，从技术层面来讲，

商业智能不是什么新技术，它是数据仓库、联机分析处理和数据挖掘等技术的综合运用。

9. 数据挖掘

数据挖掘，又译为资料探勘、数据采矿，是数据库知识发现中的一个步骤。它通常指从海量数据中，借助统计分析、在线分析处理、情报检索、机器学习、专家系统（基于经验法则与算法搜索数据中隐藏信息的方法）以及模式识别等多种技术手段，挖掘和提炼有价值知识的过程，且该领域与计算机科学有着紧密的关联。

10. 编程语言

做好大数据分析不能缺少编程语言基础，掌握 Python、R、Ruby、Java 等编程知识是必不可少的。

1.5 本章小结

大数据技术经过多年的发展已经趋于成熟，逐渐成为一条较为清晰的产业链，包括数据的采集、整理、分析、呈现等，不同的环节都有众多的参与者。随着大数据逐渐落地到广大的传统行业，大数据的应用场景会得到进一步的拓展，大数据的价值也将逐渐提升。本章简要介绍了大数据分析的背景知识、场景应用、分析算法和大数据分析的必备技能与工具，为后续内容奠定了基础。

第2章

数据特征算法分析

大数据分析挖掘是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。数据和特征决定了大数据分析的模型构建，模型和算法是推进大数据分析的工具和手段，特征工程的目的是最大限度地从原始数据中提取特征以供算法和模型使用。

2.1 数据分布性分析

统计数据的分布特征可以从两方面进行描述：一是数据分布的集中趋势；二是数据分布的离散程度。集中趋势和离散程度是数据分布特征对立且统一的两方面。本节通过介绍平均指标和变异指标这两种统计指标的概念及其计算来讨论数据集中趋势和离散程度这两方面的特征。

2.1.1 数据分布特征集中趋势的测定

集中趋势是指一组数据向某中心值靠拢的倾向，集中趋势的测度实际上是对数据一般水平代表值或中心值的测度。不同类型的数据用不同的集中趋势测度值，低层次数据的集中趋势测度值适用于高层次的测量数据；反过来，高层次数据的集中趋势测度值并不适用于低层次的测量数据。选用哪一个测度值来反映数据的集中趋势，要根据所掌握的数据的类型来确定。

通常用平均指标作为集中趋势测度指标。下面将重点介绍众数、中位数两个位置平均数和算术平均数（Arithmetic Mean）、调和平均数（Harmonic Mean）及几何平均数（Geometric Mean）3个数值型平均数。

1. 众数

众数是指一组数据中出现次数最多的变量值，用 M_0 表示。从变量分布的角度看，众数是具有明显集中趋势的数值，一组数据分布的最高峰点所对应的变量值即为众数。当然，如果数据的分布没有明显的集中趋势或最高峰点，众数就可以不存在；如果有多个高峰点，就有多个众数。

1）定类数据和定序数据众数的确定

在使用定类数据与定序数据计算众数时，只需找出出现次数最多的组所对应的变量值即可。

2）未分组数据或单变量值分组数据众数的确定

在使用未分组数据或单变量值分组数据计算众数时，只需找出出现次数最多的变量值即可。

3）组距分组数据众数的确定

对于组距分组数据来说，众数的数值与其相邻两组的频数分布有一定的关系，这种关系可作如下理解：

设众数组的频数为 f_m，众数组前一组的频数为 f_{-1}，众数组后一组的频数为 f_{+1}。当众数组相邻两组的频数相等时，即 $f_{-1}=f_{+1}$，众数组的组中值即为众数；当众数组的前一组的频数多于众数组后一组的频数时，即 $f_{-1}>f_{+1}$，众数会靠向其前一组，众数小于其组中值；当众数组后一组的频数多于众数组前一组的频数时，即 $f_{-1}<f_{+1}$，众数会靠向其后一组，众数大于其组中值。基于这种思路，借助几何图形导出的分组数据众数的计算公式如下：

$$\begin{aligned} M_0 &\doteq L+\frac{f_m-f_{-1}}{(f_m-f_{-1})+(f_m-f_{+1})}\times i \\ M_0 &\doteq U-\frac{f_m-f_{+1}}{(f_m-f_{-1})+(f_m-f_{+1})}\times i \end{aligned} \tag{2.1}$$

其中，L 表示众数所在组的下限，U 表示众数所在组的上限，i 表示众数所在组的组距。

上述下限和上限公式是假定数据分布具有明显的集中趋势，且众数组的频数在该组内是均匀分布的；若这些假定不成立，则众数的代表性会很差。从众数的计算公式可以看出，众数是根据众数组及相邻组的频率分布信息来确定数据中心点位置的，因此众数是一个位置代表值，它不受数据中极端值的影响。

2. 中位数

中位数是将总体各单位标志值按大小顺序排列后，处于中间位置的那个数值。各变量值与中位数的离差绝对值之和最小，即：

$$\sum_{i=1}^{n}\left|X_i-M_e\right|=\min \tag{2.2}$$

1）定序数据中位数的确定

确定定序数据中位数的关键是确定中间位置，中间位置所对应的变量值即为中位数。

①未分组原始资料中间位置的确定：

$$\begin{cases} \text{中位数位置}=\dfrac{N+1}{2} & N\text{为奇数} \\ \text{中位数位置}=\dfrac{N}{2} & N\text{为偶数} \end{cases} \tag{2.3}$$

②分组数据中间位置的确定：

$$\text{中位数位置}=\frac{\sum f}{2} \tag{2.4}$$

2）数值型数据中位数的确定

$$\text{数值型数据资料}=\begin{cases} \text{未分组资料} \\ \text{分组资料}\begin{cases} \text{单变量值分组资料} \\ \text{组距分组资料} \end{cases} \end{cases}$$

①未分组资料：

首先必须将标志值按大小排序。设排序的结果为：$x_1 \leqslant x_2 \leqslant x_3 \leqslant \cdots \leqslant x_n$，则

$$M_e=\begin{cases} X_{\left(\frac{N+1}{2}\right)} & \text{当}N\text{为奇数时} \\ \dfrac{1}{2}\left(X_{\frac{N}{2}}+X_{\frac{N}{2}+1}\right) & \text{当}N\text{为偶数时} \end{cases} \tag{2.5}$$

②单变量分组资料：

$$M_e=\begin{cases} X_{\left(\frac{\sum f+1}{2}\right)} & \Sigma f\text{为奇数时} \\ X_{\left(\frac{\sum f}{2}\right)} & \Sigma\text{为偶数时} \end{cases} \tag{2.6}$$

③组距分组资料：

根据上面的位置公式确定中位数所在的组，假定中位数组内的各单位是均匀分布的，则可利用下面的公式计算中位数的近似值：

$$M_e=L+\frac{\frac{\sum f}{2}-S'_{m-1}}{f_m}\cdot i$$

$$M_e=U-\frac{\frac{\sum f}{2}-S'_{m+1}}{f_m}\cdot i \tag{2.7}$$

其中，S_{m-1} 是到中位数组前面一组的向上累计频数，S'_{m+1} 则是到中位数组后面一组的向下累计频数，f_m 为中位数组的频数，i 为中位数组的组距。

3. 算术平均数

算术平均数也称为均值（Mean），是全部数据算术平均的结果。算术平均法是计算平均指标最基本、最常用的方法。算术平均数在统计学中具有重要的地位，是集中趋势的主要测度值，通常用 $\overline{X}$ 表示。根据所掌握数据形式的不同，算术平均数有简单算术平均数（Simple Arithmetic Mean）和加权算术平均数（Weighted Arithmetic Mean）两种。

1）简单算术平均数

未经分组整理的原始数据，其算术平均数的计算就是直接将一组数据的各个数值相加再除以数值个数。设总体数据为 $X_1, X_2, \cdots, X_n$，样本数据为 $x_1, x_2, \cdots, x_n$，则统计总体均值 $\overline{X}$ 和样本均值 $\overline{x}$ 的计算公式为：

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.8}$$

2）加权算术平均数

根据分组整理的数据在计算算术平均数时，就要以各组变量值出现的次数或频数为权数进行计算。设原始数据（总体或样本数据）被分成 K 或 k 组，各组的变量值为 $X_1, X_2, \cdots, X_K$ 或 $x_1, x_2, \cdots, x_k$，各组变量值的次数或频数分别为 $F_1, F_2, \cdots, F_K$ 或 $f_1, f_2, \cdots, f_k$，则总体或样本的加权算术平均数为：

$$\overline{X} \doteq \frac{X_1F_1 + X_2F_2 + \cdots + X_KF_K}{F_1 + F_2 + \cdots + F_K} = \frac{\sum_{i=1}^{K} X_iF_i}{\sum_{i=1}^{K} F_i}$$

$$\overline{x} \doteq \frac{x_1f_1 + x_2f_2 + \cdots + x_kf_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} x_if_i}{\sum_{i=1}^{k} f_i} \tag{2.9}$$

公式（2.9）中使用各组的组中值代表各组的实际数据，使用代表值时假定各组数据在组中均匀分布，但实际情况与这一假定会有一定的偏差，使得利用分组资料计算的平均数与实际平均值会产生误差，它是实际平均值的近似值。

加权算术平均数的大小不仅受各组变量值 x_i 大小的影响，而且受各组变量值出现的频数（权数 f_i）大小的影响。如果某一组的权数大，说明该组的数据较多，那么该组数据的大小对算术平均数的影响就越大；反之，则越小。实际上，我们将上式变形为下面公式（2.10）的形式，就能更清楚地看出这一点。

$$\bar{x}=\frac{\sum_{i=1}^{K}x_i f_i}{\sum_{i=1}^{K}f_i}=\sum_{i=1}^{K}x_i\frac{f_i}{\sum_{i=1}^{K}f_i} \tag{2.10}$$

由公式（2.10）可以清楚地看出，加权算术平均数受各组变量值（x_i）和各组权数（频率 $f_i/\Sigma f_i$）大小的影响。频率越大，相应的变量值计入平均数的份额也越大，对平均数的影响就越大；频率越小，相应的变量值计入平均数的份额也越小，对平均数的影响就越小。这就是权数权衡轻重作用的实质。

算术平均数在统计学中具有重要的地位，它是进行统计分析和统计推断的基础。从统计思想上看，算术平均数是一组数据的重心所在，它是消除了一些随机因素影响后或者数据误差相互抵消后的必然结果。

算术平均数具有以下重要的数学性质，这些数学性质在实际中有着广泛的应用，同时也体现了算术平均数的统计思想。

（1）各变量值与其算术平均数的离差之和等于零，即：

$$\sum_{i=1}^{n}(x_i-\bar{x})=0 \tag{2.11}$$

（2）各变量值与其算术平均数的离差平方和最小，即：

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\min \text{ 或 } \sum_{i=1}^{k}(x_i-\bar{x})^2 f_i=\min \tag{2.12}$$

4. 调和平均数

在实际工作中，经常会遇到只有各组变量值和各组标志总量而缺少总体单位数的情况，这时就要用调和平均数法计算平均指标。调和平均数是各个变量值倒数的算术平均数的倒数，习惯上用 H 表示。计算公式如下：

$$H=\frac{m_1+m_2+\cdots+m_k}{\frac{m_1}{x_1}+\frac{m_2}{x_2}+\cdots+\frac{m_k}{x_k}}=\frac{\sum_{i=1}^{K}m_i}{\sum_{i=1}^{K}\frac{m_i}{x_i}} \tag{2.13}$$

调和平均数和算术平均数本质上是一致的，唯一的区别是计算时使用了不同的数据。在实际应用时可掌握这样的原则：当计算算术平均数其分子资料未知时，就采用加权算术平均数计算；当分母资料未知时，就采用加权调和平均数计算。

$$H=\frac{\sum_{i=1}^{K}m_i}{\sum_{i=1}^{K}\frac{m_i}{x_i}}=\frac{\sum_{i=1}^{K}x_i f_i}{\sum_{i=1}^{K}\frac{x_i f_i}{x_i}}=\frac{\sum_{i=1}^{K}x_i f_i}{\sum_{i=1}^{K}f_i}=\bar{x} \tag{2.14}$$

5. 几何平均数

几何平均数是适应于特殊数据的一种平均数，在实际生活中，通常用来计算平均比率和平均速度。当所掌握的变量值本身是比率的形式，而且各比率的乘积等于总的比率时，就应采用几何平均法计算平均比率。公式如下：

$$G_M = \sqrt[N]{X_1 \times X_2 \times \cdots \times X_N} = \sqrt[N]{\prod_{i=1}^{N} X_i} \tag{2.15}$$

也可以将其看作算术平均数的一种变形：

$$\log G_M = \frac{1}{N}(\log X_1 + \log X_2 + \cdots + \log X_N) = \frac{\sum_{i=1}^{N} \log X_i}{N} \tag{2.16}$$

6. 众数、中位数与算术平均数的关系

算术平均数与众数、中位数的关系取决于频数分布的状况。它们的关系如下：

（1）当数据具有单一众数且频数分布对称时，算术平均数与众数、中位数三者完全相等，即 $M_0=M_e=\overline{X}$ 。

（2）当频数分布呈现右偏态时，说明数据存在最大值，必然拉动算术平均数靠向极大值一方，则三者之间的关系为 $\overline{X} > M_e > M_0$。

（3）当频数分布呈现左偏态时，说明数据存在最小值，必然拉动算术平均数靠向极小值一方，而众数和中位数由于是位置平均数，不受极值的影响，因此三者之间的关系为 $\overline{X} < M_e < M_0$。

当频数分布出现偏态时，极端值对算术平均数产生很大的影响，而对众数、中位数没有影响，此时用众数、中位数作为一组数据的中心值比算术平均数有较高的代表性。算术平均数与众数、中位数从数值关系上看，当频数分布的偏斜程度不是很大时，无论是左偏还是右偏，众数与中位数的距离都约为算术平均数与中位数的距离的两倍，即：

$$|M_e - M_0| = 2|\overline{X} - M_e|$$

$$M_0 = \bar{X} - 3(\bar{X} - M_e) = 3M_e - 2\bar{X} \tag{2.17}$$

2.1.2 数据分布特征离散程度的测定

离散程度是描述数据分布的另一个重要特征，反映各变量值远离其中心值的程度，因此也称为离中趋势，从另一个侧面说明了集中趋势测度值的代表程度。不同类型的数据有不同的离散程度测度值。描述数据离散程度的测度值主要有异众比率、极差、四分位差、平均差（Mean Deviation）、方差（Variance）、标准差（Standard Deviation）、离散系数等，这些指标又称为变异指标。

1. 异众比率

异众比率的作用是衡量众数对一组数据的代表性程度。异众比率越大，说明非众数组的频数占总频数的比重越大，众数的代表性就越差；反之，异众比率越小，众数的代表性就越好。异众比

率主要用于测度定类数据、定序数据的离散程度。

$$V_r = \frac{\sum F_i - F_m}{\sum F_i} = 1 - \frac{F_m}{\sum F_i} \tag{2.18}$$

其中，$\sum F_i$ 为变量值的总频数，F_m 为众数组的频数。

2. 极差

极差是一组数据的最大值与最小值之差，是最简单的离散程度测度值。极差的测度如下：

1）未分组数据

$$R = \max(X_i) - \min(X_i) \tag{2.19}$$

2）组距分组数据

R=最高组上限–最低组下限

3. 四分位差

中位数从中间点将全部数据等分为两部分。与中位数类似的还有四分位数、八分位数、十分位数和百分位数等，它们分别是用 3 个点、7 个点、9 个点和 99 个点将数据四等分、八等分、十等分和 100 等分后各分位点上的值。这里只介绍四分位数的计算，其他分位数与之类似。

一组数据排序后处于 25%和 75%位置上的值称为四分位数，也称四分位点。四分位数通过 3 个点将全部数据等分为 4 部分，其中每部分包含 25%的数据。很显然，中间的分位数就是中位数，因此通常所说的四分位数是指处在 25%位置上的数值（下四分位数）和处在 75%位置上的数值（上四分位数）。与中位数的计算方法类似，根据未分组数据计算四分位数时，首先对数据进行排序，然后确定四分位数所在的位置。

1）四分位数

设下四分位数为 Q_L，上四分位数为 Q_U：

①未分组数据：

$$Q_L = X_{\frac{n+1}{4}} \qquad Q_U = X_{\frac{3(n+1)}{4}} \tag{2.20}$$

当四分位数不在某一个位置上时，可根据四分位数的位置按比例分摊四分位数两侧的差值。

②单变量值分组数据：

$$Q_L = X_{\frac{\sum f}{4}} \qquad Q_U = X_{\frac{3\sum f}{4}} \tag{2.21}$$

③组距分组数据：

$$Q_L = L + \frac{\frac{\sum f}{4} - S_L}{f_L} \cdot i \qquad Q_U = U + \frac{\frac{3\sum f}{4} - S_U}{f_U} \cdot i \tag{2.22}$$

2）四分位差

四分位数是离散程度的测度值之一，是上四分位数与下四分位数之差，又称为四分位差，亦称为内距或四分间距（Inter-Quartile Range），用 Q_d 表示。四分位差的计算公式为：

$$Q_d = Q_U - Q_L \tag{2.23}$$

4. 平均差

平均差是离散程度的测度值之一，是各变量值与其算术平均数离差绝对值的平均数，用 M_d 表示。平均差能全面反映一组数据的离散程度，但该方法数学性质较差，实际中应用较少。

1）简单平均法

对于未分组资料采用简单平均法。其计算公式为：

$$M_D = \frac{\sum_{i=1}^{N}\left|X_i - \bar{X}\right|}{N} \tag{2.24}$$

2）加权平均法

在资料分组的情况下，应采用加权平均法。其计算公式为：

$$M_D \doteq \frac{\sum_{i=1}^{K}\left|X_i - \bar{X}\right|F_i}{\sum_{i=1}^{K}F_i} \tag{2.25}$$

5. 方差和标准差

方差和标准差同平均差一样，也是根据全部数据计算的，反映每个数据与其算术平均数相比平均相差的数值，因此能够准确地反映数据的差异程度。它们与平均差的不同之处在于计算时的处理方法不同，平均差是取离差的绝对值来消除正负号，而方差、标准差是取离差的平方来消除正负号，这更便于数学上的处理。因此，方差、标准差是实际中应用广泛的离中程度度量值。

1）总体的方差和标准差

①设总体的方差为 σ^2，标准差为 σ，对于未分组整理的原始资料，方差和标准差的计算公式分别为：

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}} \tag{2.26}$$

②对于分组数据，方差和标准差的计算公式分别为：

$$\sigma^2 \doteq \frac{\sum_{i=1}^{K}(X_i - \bar{X})^2 F_i}{\sum_{i=1}^{K}F_i} \qquad \sigma \doteq \sqrt{\frac{\sum_{i=1}^{K}(X_i - \bar{X})^2 F_i}{\sum_{i=1}^{K}F_i}} \tag{2.27}$$

2）样本的方差和标准差

样本的方差、标准差与总体的方差、标准差在计算上有所差别。总体的方差和标准差在对各个离差平方平均时是除以数据个数或总频数，而样本的方差和标准差在对各个离差平方平均时是用样本数据个数或总频数减 1（自由度）去除总离差平方和。

①设样本的方差为 S^2，标准差为 S，对于未分组整理的原始资料，方差和标准差的计算公式为：

$$S_{n-1}^2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}\qquad S_{n-1}=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}} \tag{2.28}$$

②对于分组数据，方差和标准差的计算公式为：

$$S_{n-1}^2\doteq\frac{\sum_{i=1}^{k}(x_i-\overline{x})^2 f_i}{\sum_{i=1}^{k}f_i-1}\qquad S_{n-1}\doteq\sqrt{\frac{\sum_{i=1}^{k}(x_i-\overline{x})^2 f_i}{\sum_{i=1}^{k}f_i-1}} \tag{2.29}$$

当 n 很大时，样本方差 S^2 与总体方差 σ^2 的计算结果相差很小，这时样本方差也可以用总体方差的公式来计算。

6. 相对离散程度：离散系数

前面介绍的平均差、方差和标准差都是反映一组数值变异程度的绝对值，其数值的大小不仅取决于数值的变异程度，还与变量值水平的高低、计量单位的不同有关。因此，不宜直接利用上述变异指标对不同水平、不同计量单位的现象进行比较，应当先进行无量纲化处理，即将上述反映数据的绝对差异程度的变异指标转化为反映相对差异程度的指标，再进行对比。离散系数通常用 V 表示，常用的离散系数为标准差系数，测度了数据的相对离散程度，用于对不同组别数据离散程度进行比较，计算公式为：

$$V_\sigma=\frac{\sigma}{\overline{X}}\quad 或\quad V_s=\frac{S}{\overline{x}} \tag{2.30}$$

2.1.3　数据分布特征偏态与峰度的测定

偏态和峰度是对数据分布特征的描述。偏态是对数据分布的偏移方向和程度所做的描述，峰度是对数据分布的扁平程度所做的描述。

1. 动差法

动差又称矩，原是物理学上用以表示力与力臂对重心关系的术语，它和统计学中变量与权数对平均数的关系在性质上类似，所以统计学也用动差来说明频数分布的性质。

一般来说，取变量的 a 值为中点，所有变量值与 a 之差的 K 次方的平均数称为变量 X 关于 a 的 K 阶动差。用公式表示为：

$$\frac{\sum(X-a)^K}{N} \tag{2.31}$$

当 a=0 时，即变量以原点为中心，公式（2.31）称为 K 阶原点动差，用大写英文字母 M 表示。

一阶原点动差：$$M_1=\frac{\sum X}{N} \tag{2.32}$$

二阶原点动差：$$M_2=\frac{\sum X^2}{N} \tag{2.33}$$

三阶原点动差：$$M_3=\frac{\sum X^3}{N} \tag{2.34}$$

当 $a=\bar{X}$ 时，即变量以算术平均数为中心，公式（2.31）称为 K 阶中心动差，用小写英文字母 m 表示。

一阶中心动差：$$m_1=\frac{\sum(X-\bar{X})}{N}=0 \tag{2.35}$$

二阶中心动差：$$m_2=\frac{\sum(X-\bar{X})^2}{N}=\sigma^2 \tag{2.36}$$

三阶中心动差：$$m_3=\frac{\sum(X-\bar{X})^3}{N} \tag{2.37}$$

2. 偏态及其测度

偏态是对分布偏斜方向及程度的度量。从前面的内容中我们已经知道，频数分布既有对称的，也有不对称的（偏态的）。在偏态的分布中，又有两种不同的形态，即左偏和右偏。我们可以利用众数、中位数和算术平均数之间的关系判断分布是左偏还是右偏，但要度量分布偏斜的程度就需要计算偏态系数了。

采用动差法计算偏态系数时，是用变量的三阶中心动差 m_3 与 σ^3 进行对比，计算公式为：

$$\alpha=\frac{m_3}{\sigma^3} \tag{2.38}$$

当分布对称时，变量的三阶中心动差 m_3 由于离差三次方后正负相互抵消而取得 0 值，因此 a=0；当分布不对称时，正负离差不能抵消，就形成正的或负的三阶中心动差 m_3。当 m_3 为正值时，表示正偏离差值比负偏离差值大，可以判断为正偏或右偏；反之，当 m_3 为负值时，表示负偏离差值比正偏离差值大，可以判断为负偏或左偏。$|m_3|$越大，表示偏斜的程度越大。由于三阶中心动差 m_3 含有计量单位，为消除计量单位的影响，就用 m_3 除以 σ^3，使其转化为相对数。同样地，a 的绝对值越大，表示偏斜的程度就越大。

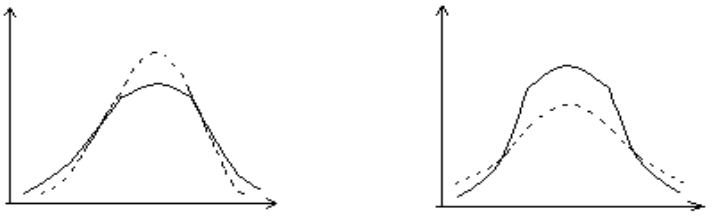
3. 峰度及其测度

峰度是用来衡量分布的集中程度或分布曲线的尖峭程度的指标。计算公式如下：

$$\alpha_4 = \frac{m_4}{\sigma_4} = \frac{\sum (X - \bar{X})^4 F_i}{\sigma^4 \cdot \sum F_i} \tag{2.39}$$

分布曲线的尖峭程度与偶数阶中心动差的数值大小有直接的关系，m_2 是方差，于是就以四阶中心动差 m_4 来度量分布曲线的尖峭程度。m_4 是一个绝对数，含有计量单位，为消除计量单位的影响，将 m_4 除以 σ^4，就得到无量纲的相对数。

衡量分布的集中程度或分布曲线的尖峭程度往往是以正态分布的峰度作为比较标准的。在正态分布条件下，$m^4/\sigma^4=3$，将各种不同分布的尖峭程度与正态分布进行比较。当峰度 $a_4>3$ 时，表示分布的形状比正态分布更瘦更高，这意味着该分布比正态分布更集中在平均数周围，这样的分布称为尖峰分布，如图 2.1（a）所示；当 $a_4=3$ 时，该分布为正态分布；当 $a_4<3$ 时，表示分布比正态分布更扁平，意味着该分布比正态分布更分散，这样的分布称为扁平分布，如图 2.1（b）所示。



（a）尖峰分布　　（b）扁平分布

图 2.1　尖峰与平峰分布示意图

2.2　数据相关性分析

数据相关性是指数据之间存在某种关系。在大数据时代，数据相关性分析因其可以快捷、高效地发现事物间的内在关联而受到广泛关注，它能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。数据相关性可以用时序分析、空间分析等方法进行分析。数据相关性分析面临着高维数据、多变量数据、大规模数据、增长性数据及其可计算方面的挑战。

2.2.1　数据相关关系

数据相关关系是指两个或两个以上变量的取值在某种意义下存在的规律，其目的在于探寻数据集里所隐藏的相关关系网。从统计学角度看，变量之间的关系大体可以分为两种类型：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，而相关关系的要求较为宽松，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：一个是相关分析，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，回归分析不仅刻画相关关系，更重要的是刻画因果关系。

1. 相关系数

对于不同测量尺度的变量，有不同的相关系数可用：

（1）Pearson 相关系数（Pearson Correlation Coefficient）：衡量两个等距尺度或等比尺度变量的相关性，是最常见的相关系数，也是学习统计学时接触的首个相关系数。

（2）净相关（Partial Correlation）：当模型中有多个自变量（或解释变量）时，去除掉其他自变量的影响，只衡量特定一个自变量与因变量之间的相关性。自变量和因变量皆为连续变量。

（3）相关比（Correlation Ratio）：衡量两个连续变量的相关性。

（4）Gamma 相关系数：衡量两个次序尺度变量的相关性。

（5）Spearman 等级相关系数：衡量两个次序尺度变量的相关性。

（6）Kendall 等级相关系数（Kendall Tau Rank Correlation Coefficient）：衡量两个人为次序尺度变量（原始资料为等距尺度）的相关性。

（7）Kendall 和谐系数：衡量两个次序尺度变量的相关性。

（8）Phi 相关系数（Phi Coefficient）：衡量两个真正名目尺度的二分变量的相关性。

（9）列联相关系数（Contingency Coefficient）：衡量两个真正名目尺度变量的相关性。

（10）四分相关（Tetrachoric Correlation）：衡量两个人为名目尺度（原始资料为等距尺度）的二分变量的相关性。

（11）Kappa 一致性系数（Kappa Coefficient of Agreement）：衡量两个名目尺度变量的相关性。

（12）点二系列相关系数（Point-Biserial Correlation Coefficient）：X 变量是真正名目尺度二分变量。Y 变量是连续变量。

（13）二系列相关系数（Biserial Correlation Coefficient）：X 变量是人为名目尺度二分变量。Y 变量是连续变量。

2. 数据种类

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，成为高维数据相关分析中的关键难题。在进行高维特征空间的相关分析时，数据可能呈现出块分布现象，如医疗数据仓库、电子商务推荐系统等。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，从本质上来看，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，它面临的挑战在于：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

在现实的大数据相关分析中，往往面临多变量的情况。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的挑战。

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的相关性，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是大规模数据的相关分析中必须解决的关键问题。有关学者给出了一种可行的拆分与融合策略，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颇具挑战性的问题。

4）增长性数据的相关分析

在大数据中，数据呈现快速增长的特征。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，对于增长性数据相关分析，特别是在线相关分析而言，每次对数据整体进行重新计算对用户来说是难以接受的，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，都与原始数据集存在某种关联模式，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。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。

3. 相关关系的种类

现象之间的相互关系很复杂，它们涉及的变动因素不同，作用方向不同，表现出来的形态也不同。相关关系大体分为以下几种。

1）正相关与负相关

按相关关系的方向可分为正相关和负相关。当两个因素（或变量）的变动方向相同时，即自变量 x 的值增大（或减小），因变量 y 的值也相应地增大（或减小），这样的关系就是正相关。例如家庭消费支出随收入的增加而增加就属于正相关。如果两个因素（或变量）变动的方向相反，即自变量 x 的值增大（或减小），因变量 y 的值随之减小（或增大），就称为负相关。例如商品流通费用率随商品经营规模的增大而逐渐减小就属于负相关。

2）单相关与复相关

按自变量的多少可分为单相关和复相关。单相关是指两个变量之间的相关关系，即所研究的问题只涉及一个自变量和一个因变量，如职工的生活水平与工资之间的关系就是单相关。复相关是指 3 个或 3 个以上变量之间的相关关系，即所研究的问题涉及若干个自变量与一个因变量，如同时研究成本、市场供求状况、消费倾向对利润的影响，这几个因素之间的关系就是复相关。

3）线性相关与非线性相关

按相关关系的表现形态可分为线性相关与非线性相关。线性相关是指在两个变量之间，当自变量 x 的值发生变动时，因变量 y 的值发生大致均等的变动，在相关图的分布上，近似地表现为直线形式。比如，商品销售额与销售量为线性相关。非线性相关是指在两个变量之间，当自变量 x 的值发生变动时，因变量 y 的值发生不均等的变动，在相关图的分布上，表现为抛物线、双曲线、指数曲线等非直线形式。比如，从人的生命全过程来看，年龄与医疗费支出呈非线性相关。

4）完全相关、不完全相关与不相关

按相关程度可分为完全相关、不完全相关和不相关。完全相关是指两个变量之间具有完全确定的关系，即因变量 y 的值完全随自变量 x 的值的变动而变动，它在相关图上表现为所有的观察点都落在同一条直线上，这时相关关系就转化为函数关系。不相关是指两个变量之间不存在相关关系，即两个变量的变动彼此互不影响。自变量 x 的值变动时，因变量 y 的值不随之做相应变动。比如，家庭收入多少与孩子数量多少之间不存在相关关系。不完全相关是介于完全相关和不相关之间的一种相关关系。比如，农作物产量与播种面积之间的关系。不完全相关关系是统计研究的主要对象。

2.2.2 数据相关分析的主要内容

相关分析是指对客观现象的相互依存关系进行分析、研究，这种分析方法叫相关分析法。相关分析的目的在于研究相互关系的密切程度及其变化规律，以便做出判断，从而进行必要的预测和控制。下面介绍相关分析的主要内容。

1）确定现象之间有无相关关系

这是相关与回归分析的起点，只有存在相互依存关系，才有必要进行进一步的分析。

2）确定相关关系的密切程度和方向

确定相关关系的密切程度主要通过绘制相关图表和计算相关系数来实现。只有达到一定密切程度的相关关系，才可配合具有一定意义的回归方程。

3）确定相关关系的数学表达式

为确定现象之间在变化上的一般关系，我们必须使用函数关系的数学公式作为相关关系的数学表达式。如果现象之间表现为直线相关，就可采用拟合直线方程的方法；如果现象之间表现为曲线相关，就可采用拟合曲线方程的方法。

4）确定因变量估计值的误差程度

使用拟合直线或曲线的方法可以找到现象之间一般的变化关系，也就是自变量 x 变化时，因变量 y 将会发生多大的变化。根据得出的直线方程或曲线方程可以给出自变量的若干数值，求得因变量的若干个估计值。估计值与实际值是有出入的，确定因变量估计值误差大小的指标是估计标准误差。估计标准误差大，表明估计不太精确；估计标准误差小，表明估计较精确。

2.2.3 相关关系的测定

相关分析的主要方法有相关表、相关图和相关系数 3 种。下面详细介绍这 3 种方法。

1）相关表

在统计中，制作相关表或相关图可以直观地判断现象之间大致存在的相关关系的方向、形式和密切程度。

在对现象总体中两种变量进行相关分析，以研究其相互依存关系时，如果将实际调查取得的一系列成对变量值的资料顺序地排列在一张表格上，这张表格就是相关表。相关表是统计表的一种。

根据资料是否分组，相关表可以分为简单相关表和分组相关表。

（1）简单相关表：

简单相关表是资料未经分组的相关表，它是一种把自变量按从小到大的顺序并配合因变量一一对应、平行排列起来的统计表。

（2）分组相关表：

在大量观察的情况下，原始资料很多，运用简单相关表很难表示。这时就要将原始资料进行分组，然后编制相关表，这种相关表称为分组相关表。分组相关表包括单变量分组相关表和双变量分组相关表两种。

- 单变量分组相关表：当原始资料很多时，对自变量数值进行分组，而对应的因变量不分组，只计算其平均值。根据资料的具体情况，自变量既可以是单项式，也可以是组距式。
- 双变量分组相关表：对两种有关变量都进行分组，交叉排列，并列出两种变量各组间的共同次数，这种统计表称为双变量分组相关表。这种表格形似棋盘，故又称棋盘式相关表。

2）相关图

相关图又称散点图，是以直角坐标系的横轴代表自变量 x，纵轴代表因变量 y，将两个变量间相对应的值用坐标点的形式描绘出来，用来反映两个变量之间相关关系的图形。

相关图既可以按未经分组的原始资料来编制，也可以按分组的资料（包括按单变量分组相关表和双变量分组相关表）来编制。通过相关图可以发现，当 y 对 x 是函数关系时，所有的相关点都会分布在某一条线上；而在相关关系的情况下，由于其他因素的影响，这些点并非完全落在一条线上，但其分布会显示出某种趋势。因此，相关图能直观地反映现象之间相关的方向和密切程度。

3）相关系数

相关表和相关图大体说明变量之间有无关系，但它们的相关关系的紧密程度却无法表达，因此需运用数学解析方法构建一个恰当的数学模型来显示相关关系及其密切程度。如果要对现象之间的相关关系的紧密程度做出确切的数量说明，就需要计算相关系数。

接下来介绍相关系数的计算。相关系数是在直线相关条件下，说明两个现象之间关系密切程度的统计分析指标，记为 γ。

相关系数的计算公式为：

$$\gamma=\frac{\sigma_{xy}{}^{2}}{\sigma_x\sigma_y}=\frac{\frac{1}{n}\sum\left(x-\bar{x}\right)\sum\left(y-\bar{y}\right)}{\sqrt{\frac{1}{n}\sum\left(x-\bar{x}\right)^2}\sqrt{\frac{1}{n}\sum\left(y-\bar{y}\right)^2}} \tag{2.40}$$

式中，n 表示资料项数，$\bar{x}$ 表示 x 变量的算术平均数，$\bar{y}$ 表示 y 变量的算术平均数，σ_x 表示 x 变量的标准差，σ_y 表示 y 变量的标准差，σ_{xy} 表示 xy 变量的协方差。

在实际问题中，如果根据原始资料计算相关系数，可运用相关系数的简捷法计算，其计算公式为：

$$\gamma=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\sqrt{n\sum y^2-\left(\sum y\right)^2}} \tag{2.41}$$

4）相关系数的分析

明晰相关系数的性质是进行相关系数分析的前提。现将相关系数的性质总结如下：

（1）相关系数的数值范围在−1 和+1 之间，即−1≤γ≤1。

（2）计算结果，当γ>0 时，x 与 y 为正相关；当γ<0 时，x 与 y 为负相关。

（3）相关系数γ的绝对值越接近 1，表示相关关系越强；越接近于 0，表示相关关系越弱。若|γ|=1，则表示两个现象完全直线相关；若|γ|=0，则表示两个现象完全不相关（不是直线相关）。

（4）相关系数γ的绝对值在 0.3 以下表示无直线相关，0.3 以上表示有直线相关，0.3～0.5 表示低度直线相关，0.5～0.8 表示显著相关，0.8 以上表示高度相关。

2.3 数据聚类分析

所谓数据聚类，是指根据数据的内在性质将数据分成一些聚合类，每一聚合类中的元素尽可能具有相同的特性，不同聚合类之间的特性差别尽可能大。

聚类分析（Cluster Analysis）的目的是分析数据是否属于各个独立的分组，使一组中的成员彼此相似，而与其他组中的成员不同。聚类分析对一个数据对象的集合进行分析，它与分类分析不同的是，所划分的类是未知的，因此聚类分析也称为无指导或无监督的（Unsupervised）学习。聚类分析的一般方法是将数据对象分组为多个类或簇（Cluster），同一簇中的对象具有较高的相似度，而不同簇中的对象差异较大。由于聚类分析的上述特征，在许多应用中，对数据集进行聚类分析后，可将一个簇中的数据对象作为一个整体对待。

数据聚类是对静态数据进行分析的一门技术，在许多领域受到广泛应用，包括机器学习、数据挖掘、模式识别、图像分析以及生物信息等。

2.3.1 聚类分析定义

1. 聚类应用

随着信息技术的高速发展，数据库应用的规模、范围和深度不断扩大，导致积累了大量的数据，而这些激增的数据后面隐藏着许多重要的信息，因此人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。目前的数据库系统可以高效、方便地实现数据的录入、查询、统计等功能，但是无法发现数据中存在的各种关系和规则，更无法根据现有的数据去预测未来的发展趋势。而数据聚类分析正是解决这一问题的有效途径，它是数据挖掘的重要组成部分，用于发现数据库中未知的对象类，为数据挖掘提供有力的支持，是近年来的研究热点之一。

聚类分析是一个极富挑战性的研究领域，基于聚类分析方法的数据挖掘在实践中已取得了较好的效果。聚类分析也可以作为其他算法的预处理步骤：聚类可以作为一个独立的工具来获知数据

的分布情况，使数据形成簇，其他算法再在生成的簇上进行处理。聚类算法既可作为特征和分类算法的预处理步骤，也可将聚类结果用于进一步的关联分析。迄今为止，人们提出了许多聚类算法，这些算法都试图解决大规模数据的聚类问题。聚类分析还成功地应用在模式识别、图像处理、计算机视觉、模糊控制等领域，并在这些领域中取得了长足的发展。

2. 数据聚类

所谓聚类，就是将一个数据单位的集合分割成几个称为簇或类别的子集，每个类中的数据都有相似性，它的划分依据就是“物以类聚”。数据聚类分析是根据事物本身的特性，研究对被聚类的对象进行类别划分的方法。聚类分析依据的原则是使同一聚簇中的对象具有尽可能高的相似性，而不同聚簇中的对象具有尽可能高的相异性。聚类分析主要解决的问题是如何在没有先验知识的前提下，实现满足这种要求的聚簇。聚类分析被称为无监督学习（Unsuper-Vised Study），主要体现在聚类学习的数据对象没有类别标记，需要由聚类学习算法自动计算。

2.3.2　聚类类型

经过持续了半个多世纪的深入研究，聚类技术已经成为常用的数据分析技术之一。各种算法的提出、发展、演化使得聚类算法家族不断壮大。下面就基于目前数据分析和数据挖掘业界主流的认知对聚类算法进行介绍。

1. 划分方法

给定具有 n 个对象的数据集，采用划分方法对数据集进行 k 个划分，每个划分（每个组）代表一个簇。$k \leqslant n$，并且每个簇至少包含一个对象，而且每个对象通常只能属于一个组。对于给定的 k 值，划分方法是：一般要做一个初始划分，然后采取迭代重新定位技术，通过让对象在不同组间移动来提高划分的准确度和精度。一个好的划分原则是：同一个簇中对象之间的相似性很高（或距离很近），而不同簇的对象之间相异度很高（或距离很远）。

（1）K-Means 算法：又叫 K 均值算法，这是目前最著名、使用最广泛的聚类算法。在给定一个数据集和需要划分的数目 k 后，该算法可以根据某个距离函数反复把数据划分到 k 个簇中，直到收敛为止。K-Means 算法用簇中对象的平均值来表示划分的每个簇，大致的步骤是：首先把随机抽取的 k 个数据点作为初始的聚类中心（种子中心），然后计算每个数据点到每个种子中心的距离，并把每个数据点分配到距离它最近的种子中心；一旦所有的数据点被分配完成，每个聚类中心（种子中心）按照本聚类（本簇）的现有数据点重新计算；这个过程不断重复，直到收敛，即满足某个终止条件，最常见的终止条件是误差平方和 SSE（指令集）局部最小。

（2）K-Medoids 算法：又叫 K 中心点算法，它与 K-Means 算法的划分过程相似，两者最大的区别是 K-Medoids 算法是用簇中最靠近中心点的一个真实的数据对象来代表该簇，而 K-Means 算法是用计算出来的簇中对象的平均值来代表该簇，这个平均值是虚拟的，并没有一个真实的数据对象具有这些平均值。

2. 层次方法

在给定 n 个对象的数据集后，可用层次方法（Hierarchical Methods）对数据集进行层次分解，

直到满足某种收敛条件为止。按照层次分解的不同形式，层次方法又可以分为凝聚层次聚类和分裂层次聚类。

（1）凝聚层次聚类：又叫自底向上方法，一开始将每个对象作为单独的一类，然后相继合并与其相近的对象或类，直到所有小的类别合并成一个类（即层次的最上面），或者达到一个收敛（即满足终止条件）。

（2）分裂层次聚类：又叫自顶向下方法，一开始将所有对象置于一个簇中，在迭代的每一步中，类会被分裂成更小的类，直到最终每个对象在一个单独的类中，或者达到一个收敛（即满足终止条件）。

3. 基于密度的方法

传统的聚类算法都是将对象之间的距离作为相似性的描述指标进行聚类划分，但是这些基于距离的方法只能发现球状类型的数据，而对于非球状类型的数据来说，只根据距离来描述和判断是不够的。鉴于此，人们提出了基于密度的方法（Density-Based Methods），其原理是只要邻近区域内的密度（对象的数量）超过了某个阈值，就继续聚类。换而言之，给定某个簇中的数据点（数据对象），在一定范围内必须包含一定数量的其他对象。该方法从数据对象的分布密度出发，把密度足够大的区域连接在一起，因此可以发现任意形状的类。该方法还可以过滤噪声数据（异常值）。基于密度的方法的典型算法包括 DBSCAN（Density-Based Spatial Clustering of Application with Noise）及其扩展算法 OPTICS（Ordering Points to Identify the Clustering Structure）。其中，DBSCAN 算法会根据一个密度阈值来控制簇的增长，将具有足够高密度的区域划分为类，并可在带有噪声的空间数据库里发现任意形状的聚类。尽管此算法优势明显，但其最大的缺点是需要用户确定输入参数，而且对输入参数十分敏感。

4. 基于网格的方法

基于网格的方法（Grid-Based Methods）将把对象空间量化为有限数目的单元，而这些单元则形成了网格结构，所有的聚类操作都在这个网格结构中进行。该算法的优点是处理速度快，其处理时间常常独立于数据对象的数目，只跟量化空间中每一维的单元数目有关。基于网格的方法的典型算法是 STING（Statistical Information Grid，统计信息网格）。该算法是一种基于网格的多分辨率聚类技术，将空间区域划分为不同分辨率级别的矩形单元，并形成一个层次结构，且高层的低分辨率单元会被划分为多个低一层次的较高分辨率单元。这种算法从最底层的网格开始，逐渐向上计算网格内数据的统计信息并存储。网格建立完成后，用类似 DBSCAN 的方法对网格进行聚类。

2.3.3 聚类应用

1. 数据聚类需要解决的问题

在聚类分析的研究中，有许多急待解决的问题，比如：

- 处理大数据量、具有复杂数据类型的数据集合时，聚类分析结果的精确性问题。
- 对高维数据的处理能力。
- 数据对象分布形状不规则时的处理能力。

- 处理噪声数据的能力,能够处理数据中包含的孤立点以及未知数据、空缺或者错误的数据。
- 对数据输入顺序的独立性，也就是对于任意的数据输入顺序产生相同的聚类结果。
- 减少对先决知识或参数的依赖性等问题。

这些问题的存在使得我们研究高准确率、低复杂度、I/O 开销小、适合高维数据、具有高度的可伸缩性的聚类方法迫在眉睫，这也是今后聚类方法研究的方向。

2. 数据聚类的应用

聚类分析可以作为一个独立的工具来获得数据的分布情况，通过观察每个簇的特点，集中对特定的某些簇进行进一步的分析，以获得需要的信息。聚类分析应用广泛，除了应用在数据挖掘、模式识别、图像处理、计算机视觉、模糊控制等领域外，它还应用在气象分析、食品检验、生物种群划分、市场细分、业绩评估等诸多方面。例如在商务上，聚类分析可以帮助市场分析人员从客户基本库中发现不同的客户群，并且用购买模式来刻画不同的客户群特征。聚类分析还可以应用在欺诈探测中，聚类中的孤立点就可能预示着欺诈行为的存在。聚类分析的发展过程也是聚类分析的应用过程，目前聚类分析在相关领域已经取得了丰硕的成果。

2.4　数据主成分分析

在实际问题中，我们经常会遇到研究多个变量的问题，而且在多数情况下，多个变量之间常常存在一定的相关性。由于变量个数较多，再加上变量之间存在相关性，势必增加了分析问题的复杂性。要想把多个变量综合为少数几个代表性变量，既能够代表原始变量的绝大多数信息，又互不相关，并且可以在新的综合变量的基础上进一步进行统计分析，就需要使用主成分分析。

2.4.1　主成分分析的原理及模型

1. 主成分分析的原理

主成分分析是采取一种数学降维的方法找出几个综合变量来代替原来众多的变量，使这些综合变量能尽可能地代表原来变量的信息量，而且彼此之间互不相关。这种把多个变量化为少数几个互相无关的综合变量的统计分析方法就叫作主成分分析或主分量分析。

主成分分析所要做的就是：设法将原来众多的具有一定相关性的变量重新组合为一组新的、相互无关的综合变量。通常，数学上的处理方法是将原来的变量进行线性组合，作为新的综合变量，但若这种组合不加以限制，则会有很多变量，应该如何选择呢？如果将选取的第一个线性组合（第一个综合变量）记为 F_1，自然希望它尽可能多地反映原来变量的信息，这里“信息”用方差来测量，即希望 $\mathrm{Var}(F_1)$尽可能大，表示 F_1 包含的信息尽可能多。因此，在所有的线性组合中，选取的 F_1 应该是方差最大的，故称 F_1 为第一主成分。如果第一主成分不足以代表原来 P 个变量的信息，再考虑选取 F_2（第二个线性组合），为了有效地反映原来的信息，F_1 已有的信息就不再出现在 F_2 中，用数学语言表达就是要求 $\mathrm{Cov}(F_1, F_2)=0$，称 F_2 为第二主成分；以此类推，可以构造出第三、第四……第 P 个主成分。

2. 主成分分析的数学模型

对于一个样本资料，观测 p 个变量 $x_1,x_2,\cdots,x_p$，n 个样品的数据资料矩阵为：

$$\boldsymbol{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}=(x_1,x_2,\cdots,x_p) \tag{2.42}$$

其中，$\boldsymbol{x}_j=\begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$，$j=1,2,\cdots,p$。

主成分分析就是将 p 个观测变量综合成 p 个新的变量（综合变量），即：

$$\begin{cases} F_1=a_{11}x_1+a_{12}x_2+\cdots+a_{1p}x_p \\ F_2=a_{21}x_1+a_{22}x_2+\cdots+a_{2p}x_p \\ \cdots \\ F_p=a_{p1}x_1+a_{p2}x_2+\cdots+a_{pp}x_p \end{cases} \tag{2.43}$$

简写为：

$$\begin{gathered} F_j=\alpha_{j1}x_1+\alpha_{j2}x_2+\cdots+\alpha_{jp}x_p \\ j=1,2,\cdots,p \end{gathered} \tag{2.44}$$

要求模型满足以下条件：

（1）F_i、F_j 互不相关（$i\neq j$，$i,j=1,2,\cdots,p$）。

（2）F_1 的方差大于 F_2 的方差大于 F_3 的方差。

（3）$a_{k1}{}^2+a_{k2}{}^2+\cdots+a_{kp}{}^2=1$，$k=1,2,\cdots p$。

于是，称 F_1 为第一主成分，F_2 为第二主成分，以此类推，有第 p 个主成分。主成分又叫主分量。这里 a_{ij} 称为主成分系数。

上述模型可用矩阵表示为：$\boldsymbol{F}=\boldsymbol{AX}$。

其中：

$$\boldsymbol{F}=\begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_p \end{pmatrix} \qquad \boldsymbol{X}=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \tag{2.45}$$

$$\boldsymbol{A}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}=\begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_p \end{pmatrix} \tag{2.46}$$

$\boldsymbol{A}$ 称为主成分系数矩阵。

2.4.2　数据主成分分析的几何解释

假设有 n 个样品，每个样品有两个变量，即在二维空间中讨论主成分的几何意义。设 n 个样品在二维空间中的分布大致为一个椭圆，如图 2.2 所示。

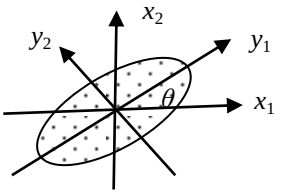


图 2.2　主成分几何解释图

将坐标系正交旋转一个角度 θ，使其在椭圆长轴方向取坐标 y_1，在椭圆短轴方向取坐标 y_2，旋转公式为：

$$\begin{cases} y_{1j} = x_{1j}\cos\theta + x_{2j}\sin\theta \\ y_{2j} = x_{1j}(-\sin\theta) + x_{2j}\cos\theta \end{cases}$$

$$j = 1, 2, \cdots, n \tag{2.47}$$

写成矩阵形式为：

$$\begin{aligned} \boldsymbol{Y} &= \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix} = \boldsymbol{U} \cdot \boldsymbol{X} \end{aligned} \tag{2.48}$$

其中，$\boldsymbol{U}$ 为坐标旋转变换矩阵，它是正交矩阵，即有 $\boldsymbol{U}'=\boldsymbol{U}^{-1}$，$\boldsymbol{U}\boldsymbol{U}'=\boldsymbol{I}$（$\boldsymbol{U}'$是 $\boldsymbol{U}$ 的转置矩阵，以下相同），即满足 $\sin^2\theta+\cos^2\theta=1$。

经过旋转变换后，得到如图 2.3 所示的新坐标。

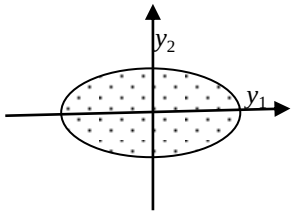


图 2.3　新坐标

新坐标 y_1-y_2 有如下性质：

（1）n 个点的坐标 y_1 和 y_2 的相关性几乎为零。

（2）二维平面上的 n 个点的方差大部分都归结在 y_1 轴上，而 y_2 轴上的方差较小。

y_1 和 y_2 称为原始变量 x_1 和 x_2 的综合变量。由于 n 个点在 y_1 轴上的方差最大，因此将二维空间的点用 y_1 轴上的一维综合变量来代替，损失的信息量最小，由此称 y_1 轴为第一主成分；y_2 轴与 y_1 轴正交，有较小的方差，称它为第二主成分。

2.4.3 数据主成分的导出

根据主成分分析的数学模型的定义，要进行主成分分析，就需要根据原始数据以及模型的 3 个条件，求出主成分系数，以便得到主成分模型。这就是导出主成分所要解决的问题。

（1）根据 2.4.1 节中主成分数学模型的条件（1），要求主成分之间互不相关，因此主成分之间的协方差矩阵应该是一个对角矩阵。即，对于主成分：

$$\boldsymbol{F}=\boldsymbol{AX} \tag{2.49}$$

其协方差矩阵应为：

$$\begin{aligned}\mathrm{Var}(\boldsymbol{F})&=\mathrm{Var}(\boldsymbol{AX})=(\boldsymbol{AX})\cdot(\boldsymbol{AX})'=\boldsymbol{AXX'A'}\\&=\boldsymbol{\Lambda}=\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{pmatrix}\end{aligned} \tag{2.50}$$

（2）设原始数据的协方差矩阵为 $\boldsymbol{V}$，若原始数据进行了标准化处理，则协方差矩阵 $\boldsymbol{V}$ 等于相关矩阵 $\boldsymbol{R}$，即有：

$$\boldsymbol{V}=\boldsymbol{R}=\boldsymbol{XX'} \tag{2.51}$$

（3）由 2.4.1 节中主成分数学模型条件（3）和正交矩阵的性质，若能够满足条件（3），则最好要求 $\boldsymbol{A}$ 为正交矩阵，即满足：

$$\boldsymbol{AA'}=\boldsymbol{I} \tag{2.52}$$

于是，将原始数据的协方差代入主成分的协方差矩阵公式得：

$$\begin{aligned}&\mathrm{Var}(F)=\boldsymbol{AXX'A'}=\boldsymbol{ARA'}=\boldsymbol{\Lambda}\\&\boldsymbol{ARA'}=\boldsymbol{\Lambda}\quad\boldsymbol{RA'}=\boldsymbol{A'\Lambda}\end{aligned} \tag{2.53}$$

展开上式得：

$$\begin{pmatrix}r_{11} & r_{12} & \cdots & r_{1p}\\ r_{21} & r_{22} & \cdots & r_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ r_{p1} & r_{p2} & \cdots & r_{pp}\end{pmatrix}\cdot\begin{pmatrix}a_{11} & a_{21} & \cdots & a_{p1}\\ a_{12} & a_{22} & \cdots & a_{p2}\\ \vdots & \vdots & \vdots & \vdots\\ a_{1p} & a_{2p} & \cdots & a_{pp}\end{pmatrix}=\begin{pmatrix}a_{11} & a_{21} & \cdots & a_{p1}\\ a_{12} & a_{22} & \cdots & a_{p2}\\ \vdots & \vdots & \vdots & \vdots\\ a_{1p} & a_{2p} & \cdots & a_{pp}\end{pmatrix}\cdot\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{pmatrix} \tag{2.54}$$

展开等式两边，根据矩阵相等的性质，这里只根据第一列得出的方程为：

$$\begin{cases} (r_{11}-\lambda_1)a_{11}+r_{12}a_{12}+\cdots+r_{1p}a_{1p}=0 \\ r_{21}a_{11}+(r_{22}-\lambda_1)a_{12}+\cdots+r_{2p}a_{1p}=0 \\ \cdots \\ r_{p1}a_{11}+r_{p2}a_{12}+\cdots+(r_{pp}-\lambda_1)a_{1p}=0 \end{cases} \tag{2.55}$$

为了得到该齐次方程的解，要求其系数矩阵行列式为 0，即：

$$\begin{vmatrix} r_{11}-\lambda_1 & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22}-\lambda_1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{1p} & r_{p2} & \cdots & r_{pp}-\lambda_1 \end{vmatrix}=0 \tag{2.56}$$

$$|\boldsymbol{R}-\boldsymbol{\lambda}_1\boldsymbol{I}|=0$$

显然，$\boldsymbol{\lambda}_1$是相关系数矩阵的特征值，$\boldsymbol{a}_1=(\boldsymbol{a}_{11},\boldsymbol{a}_{12},\cdots,\boldsymbol{a}_{1p})$是相应的特征向量。根据第二列、第三列等可以得到类似的方程，于是 $\boldsymbol{\lambda}_i$ 是特征方程 $|\boldsymbol{R}-\boldsymbol{\lambda}\boldsymbol{I}|=0$ 的特征根，$\boldsymbol{a}_j$ 是其特征向量的分量。

2.4.4　证明主成分的方差是依次递减的

设相关系数矩阵 $\boldsymbol{R}$ 的 p 个特征根为 $\boldsymbol{\lambda}_1\geqslant\boldsymbol{\lambda}_2\geqslant\cdots\geqslant\boldsymbol{\lambda}_p$，相应的特征向量为 $\boldsymbol{a}_j$。

$$\boldsymbol{A}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}=\begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_p \end{pmatrix} \tag{2.57}$$

相对于 F_1 的方差为：

$$\mathrm{Var}(F_1)=\boldsymbol{a}_1\boldsymbol{X}\boldsymbol{X}'\boldsymbol{a}_1'=\boldsymbol{a}_1\boldsymbol{R}\boldsymbol{a}_1'=\boldsymbol{\lambda}_1 \tag{2.58}$$

同样有 $\mathrm{Var}(F_i)=\boldsymbol{\lambda}_i$，即主成分的方差依次递减，并且协方差为：

$$Cov(\boldsymbol{a}_i'\boldsymbol{X}',\boldsymbol{a}_j\boldsymbol{X})=\boldsymbol{a}_i'\boldsymbol{R}\boldsymbol{a}_j$$

$$=\boldsymbol{a}_i'(\sum_{\alpha=1}^{p}\boldsymbol{\lambda}_\alpha\boldsymbol{a}_\alpha\boldsymbol{a}_\alpha')\boldsymbol{a}_j$$

$$=\sum_{\alpha=1}^{p}\boldsymbol{\lambda}_\alpha(\boldsymbol{a}_i'\boldsymbol{a}_\alpha)(\boldsymbol{a}_\alpha'\boldsymbol{a}_j)=0,\quad i\neq j \tag{2.59}$$

根据证明得知，主成分分析中的主成分协方差应该是对角矩阵，其对角线上的元素恰好是原始数据相关矩阵的特征值，而主成分系数矩阵 $\boldsymbol{A}$ 的元素则是原始数据相关矩阵特征值相应的特征向量。矩阵 $\boldsymbol{A}$ 是一个正交矩阵。

于是，变量$(x_1,x_2,\cdots,x_p)$经过变换后得到新的综合变量：

$$\begin{cases} F_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ F_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \cdots \\ F_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{cases} \tag{2.60}$$

新的随机变量彼此不相关，且方差依次递减。

2.4.5 数据主成分分析的计算

样本观测数据矩阵为：

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{2.61}$$

（1）对原始数据进行标准化处理：

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{\sqrt{\mathrm{var}(x_j)}} \quad (i=1, 2, \cdots, n; j=1, 2, \cdots, p) \tag{2.62}$$

其中：

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \mathrm{var}(x_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2 \tag{2.63}$$

$$j = 1, 2, \cdots, p$$

（2）计算样本相关系数矩阵：

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \tag{2.64}$$

为了方便，假定原始数据标准化后仍用 $\boldsymbol{X}$ 表示，则经标准化处理后的数据的相关系数为：

$$r_{ij} = \frac{1}{n-1}\sum_{t=1}^{n} x_{ti}x_{tj} \tag{2.65}$$

$$i, j = 1, 2, \cdots, p$$

（3）用雅可比方法求相关系数矩阵 $\boldsymbol{R}$ 的特征值（$\lambda_1, \lambda_2, \cdots, \lambda_p$）和相应的特征向量 $\boldsymbol{a}_i=(a_{i1},a_{i2},\cdots,a_{ip})$，$i=1,2,\cdots,p$。

（4）选择重要的主成分，并写出主成分表达式。

主成分分析可以得到 p 个主成分，但是，由于各个主成分的方差是递减的，包含的信息量也

是递减的，因此在实际分析时，一般不选取 p 个主成分，而是根据各个主成分累计贡献率的大小选取前 k 个主成分。这里贡献率是指某个主成分的方差占全部方差的比重，实际上就是某个特征值占全部特征值合计的比重，即：

$$\text{贡献率}=\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i} \tag{2.66}$$

贡献率越大，说明该主成分所包含的原始变量的信息越强。主成分个数 k 的选取主要根据主成分的累积贡献率来决定，即一般要求累计贡献率达到 85%以上，这样才能保证综合变量包括原始变量的绝大多数信息。

在实际应用中，选定重要主成分后，还需注意对其实际含义进行解释。主成分分析中的一个关键问题是如何赋予主成分新的意义并给出合理解释。通常，这种解释需结合主成分表达式中各变量的系数与定性分析进行。主成分是原始变量的线性组合，其系数有大有小、有正有负，有时多个变量的系数大小相当，因此不能简单地将某一主成分归结为某个原始变量的作用结果。一般而言，系数绝对值较大的变量对主成分的贡献更大；当多个变量的系数绝对值相近时，应认为该主成分综合反映了这些变量的共同信息。此时，需结合具体问题和专业知识，对这些变量共同体现的实际意义作出恰当解释，以实现深入分析的目的。

（5）计算主成分得分。

根据标准化的原始数据，将各个样品分别代入主成分表达式，就可以得到各主成分下的各个样品的新数据，即为主成分得分。具体形式如下：

$$\begin{pmatrix} F_{11} & F_{12} & \cdots & F_{1k} \\ F_{21} & F_{22} & \cdots & F_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{nk} \end{pmatrix} \tag{2.67}$$

（6）进一步的统计分析。

依据主成分得分的数据，可以进行进一步的统计分析。其中，常见的应用有主成分回归、变量子集合的选择、综合评价等。

2.5　数据动态性分析

动态数据是指观察或记录下来的一组按时间先后顺序排列的数据序列。

1. 数据特征

（1）构成：

- 时间。
- 反映现象在一定时间条件下的数量特征的指标值。

（2）表示：

- $x(t)$：时间 t 为自变量。
- 整数：离散的、等间距的。
- 非整数：连续的，实际分析时必须进行采样处理。
- 时间单位：秒、分、小时、日、周、月、年。

2. 动态数据分类——按照指标值的表现形式

（1）绝对数序列：

- 时期序列：可加性。
- 时点序列：不可加性。

（2）相对数/平均数序列。

3. 时间数据分类——按照时间的表现形式

- 连续。
- 离散。
- 时间序列中，时间必须是等间隔的。

4. 动态数据的特点

- 数据取值随时间变化。
- 在每一时刻取什么值，不可能完全准确地用历史值预报。
- 前后时刻（不一定是相邻时刻）的数值或数据点有一定的相关性。
- 整体存在某种趋势或周期性。

5. 动态数据的构成与分解

时间序列=趋势+周期+平稳随机成分+白噪声

6. 动态数据分析模型分类

（1）研究单变量或少数几个变量的变化：

- 随机过程：周期分析和时间序列分析。
- 灰色系统：关联分析，GM 模型。

（2）研究多变量的变化：

- 系统动力学建模。

7. 时间序列模型

- 研究一个或多个被解释变量随时间变化的规律的模型。
- 模型主要用于预测分析。
- 目的：精确预测未来的变化。
- 数据要求：序列平稳。

- 研究角度：
  - 时间域。
  - 频率域。
- 模型内容：
  - 周期分析。
  - 时间序列预测。

时间序列模型的表示：

$$x_t = f\left(x_{t-1}, x_{t-2}, \cdots\right) + \varepsilon_t \tag{2.68}$$

ε_t表示白噪声。

8. 动态系统模型

- 研究具有时变特点的多个因素之间的相互作用，以及这些作用与系统整体发展之间的关系。
- 模型主要用于模拟和情景分析。
- 研究重点：各种因素是如何相互作用影响系统总体发展的。

9. 模型表示

- 因果反馈逻辑图。
- 未来系统要素变化趋势图。

10. 建模步骤

01 分析数据的动态特征。
02 进行数据序列分解。
03 数据预处理。
04 模型构建。
05 模型确认。

11. 建模方法

（1）时间序列模型：

- 统计学方法：随机过程理论。
- 灰色系统方法。

（2）动态系统模型：

- 动态系统仿真方法。

12. 时间序列模型

（1）平稳随机过程：

如果一个随机过程的均值和方差在时间过程上是常数，并且在任何两个时期之间的协方差值仅依赖这两个时期间的距离和滞后，而不依赖计算这个协方差的实际时间，那么这个随机过程就被

称为平稳的随机过程。

- 严平稳：一种条件比较苛刻的平稳性定义。只有当序列所有的统计性质都不会随着时间的推移而发生变化时，该序列才能被认为是平稳的。
- 宽平稳：使用序列的特征统计量来定义的一种平稳性，认为序列的统计性质主要由它的低阶矩阵决定，所以只要保证序列低阶矩阵（二阶）平稳，就能保证序列的主要性质近似稳定。

（2）平稳序列的统计性质：

- 常数均值。
- 自协方差函数和自相关函数只依赖时间的平移长度，与时间的起止点无关。

（3）自相关函数：

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \tag{2.69}$$

其他的动态数据模型有线性模型法、非线性趋势等。

13. 时间序列建模

任何时间序列都可以看作一个平稳的过程，所看到的数据集可以看作该平稳过程的一个实现。主要方法有自回归 AR(p)、移动平均 MA(q)与自回归移动平均 ARMA(p,q)等。

1）自回归（AR）模型

时间序列可以表示成它的先前值和一个冲击值的函数：

$$x_t = \varnothing_1 x_{t-1} + \varnothing_{12} x_{t-2} + \cdots + \varnothing_p x_{t-p} + \varepsilon_t \tag{2.70}$$

2）滑动平均（MA）模型

序列值是现在和过去的误差或冲击值的线性组合：

$$x_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{2.71}$$

3）自回归滑动平均（ARMA）模型

序列值是现在和过去的误差或冲击值以及先前的序列值的线性组合：

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{2.72}$$

2.6 数据可视化

数据可视化是研究数据视觉表现形式的科学技术。这种视觉表现形式被定义为以某种概要方式提取的信息，包括相应信息单元的各种属性和变量。这是一个不断演化的概念，其边界持续扩展，主要指技术上较为高级的方法，这些方法利用图形、图像处理、计算机视觉和用户界面，通过对立

体、表面、属性及动画的表达与建模，实现对数据的可视化解释。相比立体建模等特定技术，数据可视化涵盖的技术方法更为广泛。为有效传达思想，美学形式与功能需齐头并进，通过直观呈现数据的关键方面与特征，实现对稀疏而复杂数据集的深入洞察。

数据可视化与信息图形、信息可视化、科学可视化以及统计图形密切相关。当前，在研究、教学和开发领域，数据可视化是一个极为活跃而又关键的方向。“数据可视化”这一术语实现了成熟的科学可视化领域与较年轻的信息可视化领域的融合。

数据可视化技术包含以下几个基本概念。

（1）数据空间：是由 n 维属性和 m 个元素组成的数据集所构成的多维信息空间。

（2）数据开发：是指利用一定的算法和工具对数据进行定量的推演和计算。

（3）数据分析：是指对多维数据进行切片、块、旋转等操作来剖析数据，从而能够多角度多侧面观察数据。

（4）数据可视化：是指将大型数据集中的数据以图形图像形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程。

数据可视化已发展出多种方法，根据其可视化原理的不同，可分为基于几何的技术、面向像素的技术、基于图标的技术、基于层次的技术、基于图像的技术和分布式技术等。数据可视化的适用范围有多种划分方式，其中常见的关注焦点是信息的呈现。数据可视化的两个主要组成部分是统计图形和主题图。

1. 数据的特性

要理解数据可视化，先要理解数据，再去掌握可视化的方法，这样才能实现高效的数据可视化。数据具有以下特性：

（1）量性：数据是可以计量的，所有的值都是数字。

（2）离散性：数字类数据可能在有限的范围内取值。

（3）持续性：数据可以测量，且在有限范围内。

（4）范围性：数据可以根据编组和类别来分类。

可视化的意义是帮助人更好地分析数据，也就是说，这是一种高效的手段，并不是数据分析的必要条件。如果我们采用了可视化方案，就意味着机器并不能精确地分析。当然，要明确可视化不能直接带来结果，它需要人来介入分析。

2. 数据可视化方法及工具

具有代表性的图形化数据的可视化方法如下：

- 柱形图。
- 散点图。
- 地图。
- 面积图。
- 漏斗图。
- 仪表盘。

- 饼图。
- 折线图。
- 矩形树图。

编程语言类数据可视化工具如下：

- R

R 经常被称为“为统计人员开发的一种语言”。如果需要深奥的统计模型用于计算，可以在 CRAN（Comprehensive R Archive Network，综合 R 档案网络）找到。说起用于分析和标绘，没有什么比得过 ggplot2。而如果想利用比机器提供的还强大的功能，那么可以使用 SparkR 绑定，在 R 上运行 Spark。

- Scala

Scala 在 JVM 上运行，成功地结合了函数范式和面向对象范式，目前它在金融界和需要处理海量数据的企业中取得了巨大进展，常常采用大规模分布式方式（比如 Twitter 和 LinkedIn）来处理。它还是驱动 Spark 和 Kafka 的语言之一。

- Python

Python 在学术界一直广受欢迎，尤其在自然语言处理（NLP）等领域。因此，若你有一个需要自然语言处理的项目，将面临众多选择，包括经典的 NLTK、基于 Gensim 的主题建模，以及快速而准确的 spaCy。在神经网络方面，Python 同样表现出色，拥有 Theano 和 TensorFlow；此外，还有面向机器学习的 Scikit-learn，以及面向数据分析的 NumPy 和 Pandas。

- Java

Java 非常适合大数据项目。Hadoop MapReduce 和 HDFS 均使用 Java 编写。此外，Storm、Kafka 和 Spark 也均可在 JVM 上运行（分别使用 Clojure 和 Scala），这使得 Java 成为这些项目中的“一等公民”。另外，像 Google Cloud Dataflow（现为 Apache Beam）等新技术，现在仍仅支持 Java。

在大数据时代，可视化图表工具无法“单独作战”。众所周知，大数据的价值在于数据挖掘，而数据可视化通常与数据分析功能紧密结合。数据分析又依赖数据接入与整合、数据处理、ETL（Extraction-Transformation-Loading，抽取-转换-加载）等能力，最终发展为一站式的大数据分析平台。

2.7 本章小结

数据和特征决定了机器学习的上限，而模型和算法只是逼近这一上限的手段。机器学习在数据分析中的目的之一是直观地展现数据，例如将花费数小时甚至更长时间才能归纳的数据转换成一眼就能读懂的指标；通过加、减、乘、除等各类公式计算出的两组数据的差异，可在图中通过元素的颜色、长度、大小等形成对比。

本章从数据分布性、数据相关性、数据聚类性、数据成分、数据动态性及数据可视化等方面介绍了机器学习的数据特征，便于读者对数据特征有一个清晰的认识。