分类号_____　　密级_____
U D C _____　　编号　10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

（专业学位）

论文题目　基于非负矩阵分解的
函数型聚类算法研究及应用

研　究　生　姓　名：赵芳芳

指导教师姓名、职称：高海燕　教授

学 科、专 业 名 称：统计学　应用统计硕士

研　究　方　向：大数据分析

提　交　日　期：2024 年 6 月 3 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：赵芳芳 签字日期：2024.6.3

导师签名：高海燕 签字日期：2024年6月3日

导师(校外)签名：_____ 签字日期：_____

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：赵芳芳 签字日期：2024.6.3

导师签名：高海燕 签字日期：2024年6月3日

导师(校外)签名：_____ 签字日期：_____

Research and Application of Functional Clustering Algorithm Based on Non-Negative Matrix Factorization

Candidate : Fangfang Zhao

Supervisor: Haiyan Gao

摘 要

数据采集频率增加导致出现了具有“函数”性质的数据，函数型数据分析(Functional Data Analysis，FDA)应运而生，其中函数型聚类分析(Functional Clustering)成为探索函数型数据的重要工具。目前，对于函数型聚类分析的研究大多采用“曲线拟合+聚类”的两步法实现，存在提取最优类别信息效果不佳、计算成本高等问题，这会影响聚类效果的准确性和效率。为了解决这些问题，本文在非负矩阵分解(Non-negative Matrix Factorization，NMF)的框架下，采用函数型数据分析方法，侧重研究基于非负矩阵分解的函数型聚类一步法，主要研究内容包括以下三部分：

（1）为了有效利用数据的非线性和低维流形结构，提出了基于双随机图正则化矩阵分解的函数型聚类算法(Functional Clustering Algorithm Based on Bi-stochastic Graph Regularized Matrix Factorization，BSMFFC)。引入图正则化技术，构造最近邻图来模拟流形结构，结合双随机矩阵动态更新图，从而充分利用了数据的固有几何结构信息。同时，给出了模型的优化求解算法，并计算了时间复杂度。模拟实验结果验证该算法可行性，实例应用验证了该算法的实用性。

（2）针对含有噪声和异常值的情况，提出了基于鲁棒图正则化矩阵分解的函数型聚类算法(Functional Robust Manifold Nonnegative Matrix Factorization，FRMNMF)。利用$l_{2,1}$范数来定义损失函数，从而减弱数据中噪声或异常值的影响，并利用流形学习，保证了数据的局部不变性。给出更新算法，并证明了算法收敛性和计算复杂度。在合成数据和真实数据上的实验结果表明，该算法在函数型聚类任务中具有一定的鲁棒性。以城镇居民人均可支配收入数据应用为例，其聚类结果表明该算法的可行性、合理性及实际应用价值。

（3）针对多视角函数型数据出现缺失的情况，提出一种自加权不完整多视角函数型聚类算法(Adaptive Incomplete Multi-view Clustering for Functional dataset，AIMFC)。将多视角学习、非负矩阵分解以及矩阵填充进行融合，采用自加权方法为每个视角分配相应的权重。给出算法更新公式，借助模拟实验，验证了该算法的可行性。此外，对于针对北京市空气污染物小时浓度数据所得出的聚类结果而言，改进后的算法在缺失数据聚类问题上表现出了优异的效果。

关键词：函数型数据　聚类分析　非负矩阵分解　鲁棒性　缺失值　多视角学习

Abstract

The increasing frequency of Data collection has led to the emergence of "Functional" data, and Functional Data Analysis (FDA) has emerged, in which Functional Clustering has become a crucial instrument for exploring functional data. Currently, most researches on functional cluster analysis adopt the two-step method of "curve fitting + clustering", which has problems such as poor effect of extracting optimal category information and high computing cost, which will affect the accuracy and efficiency of clustering effect. In order to solve these problems, this paper adopts functional data analysis method under the framework of Non-negative Matrix Factorization (NMF), focusing on the study of functional clustering one-step method based on non-negative matrix factorization. The main research contents include the following three parts:

(1) In order to effectively utilize the nonlinear and low-dimensional manifold structure of data, a functional clustering algorithm based on the regularization matrix decomposition of double random graphs is proposed. (Functional Clustering Algorithm Based on Bi-stochastic Graph Regularized Matrix Factorization， BSMFFC). By introducing graph regularization technique, the nearest neighbor graph is constructed to simulate manifold structure, and the graph is dynamically updated with doubly stochastic matrix, thus making full use of the inherent geometric structure information of data. At the same time, the optimization algorithm of the model is given, and the time complexity is calculated. The simulation results verify the feasibility of the algorithm, and the practical application of the algorithm is verified.

(2) For the cases containing noise and outliers, a Functional Robust Manifold Nonnegative Matrix Factorization (FRMNMF) algorithm is

proposed based on the robust graph regularization matrix factorization. The $l_{2,1}$-norm is used to define the loss function to reduce the influence of noise or outliers in the data, and the manifold learning is used to ensure the local invariance of the data. The updated algorithm is given, and its convergence and computational complexity are proved. Experimental results on synthetic data and real data show that the proposed algorithm is robust in functional clustering tasks. Taking the per capita disposable income data of urban residents as an example, the clustering results show the feasibility, rationality and practical application value of the algorithm.

(3) Aiming at the absence of Multi-view Functional data, an Adaptive Incomplete multi-view Clustering for Functional dataset (AIMFC) is proposed. Multi-perspective learning, non-negative matrix decomposition and matrix filling are integrated, and the self-weighting method is used to assign the corresponding weight to each perspective. The algorithm updating formula is given, and the feasibility of the algorithm is verified by simulation experiment. In addition, for the clustering results obtained from the hourly concentration data of air pollutants in Beijing, the improved algorithm shows excellent results in the missing data clustering problem.

Keywords: Functional Data; Clustering; Non-negative Matrix Factorization; Robustness; Missing Data; Multi-view Learning

目 录

1 绪论

1.1 研究背景及意义

随着物联网和大数据监测技术的高速发展，数据采集频率不断提高导致出现了一种具有无穷维、连续特征的数据，比如交通实时监测的车流量数据、医学领域的核磁共振数据、可穿戴设备的监测数据等，这些具有函数性特征的数据被称为“函数型数据”。而将这些数据看成整体进行统计分析，则称为函数型数据分析，其分析和处理的对象是函数或曲线。

传统聚类分析方法在函数型数据上难以直接运用，为了有效的对函数型数据进行聚类，函数型聚类分析方法应运而生。从实际角度上，观察到的函数型数据一般为离散高频数据，函数型聚类分析需要将数据拟合成光滑曲线，然后对曲线进行聚类，将曲线生成过程和聚类过程并联进行的函数型聚类一步法能够最有效地提取最优类别信息，有助于提升聚类效果。非负矩阵分解要求分解后的分量均为非负值，不仅能够降低数据的维度，还能保持分解后数据的非负性，常用于对数据的聚类。因此，依据非负矩阵分解本身具有的降维作用和聚类特性，本文围绕函数型聚类方法展开研究，从单视角和多视角两个方面出发，提出几种基于非负矩阵分解的函数型聚类模型和算法，旨在提高聚类性能，用以克服现有函数型数据聚类方法的局限性。

本文具有一定的理论价值和实际意义，在高频采集函数型数据进行聚类的过程中，函数型聚类方法相较于传统聚类方法具备明显的优势。函数型聚类方法通过对曲线本身的分析，能够更深入地挖掘曲线内部的动态变化特征，并将这些变化信息纳入到聚类相似性的考虑范畴中。这样的方法更加有助于我们全面了解函数型数据的潜在规律和特征，为进一步相关研究提供了更准确、全面的基础。相对于传统聚类方法，函数型聚类方法在高频采集函数型数据的聚类分析中具有更多的优势。引入非负矩阵分解框架的函数型聚类方法将拟合、降维、聚类融在同一个距离损失函数之中，避免了惩罚的方式实现要素并联，计算成本更低。在此基础上，通过利用流形学习、自加权等技术、使算法具有优越的聚类性能，具有一定的理论意义。对于数据采集频率高的领域而言，函数型聚类方法在展现其实

际价值时具有显著优势。结合实际数据与聚类结果，判断聚类性能，对现实中标签未知数据的类别差异分析具有重要的现实意义，可以解决社会科学、经济等领域中的聚类问题。

1.2 国内外研究现状

1.2.1 函数型聚类算法

函数型数据聚类是函数型数据分析领域的一大重要研究方向。现有的文献中提出的函数型聚类方法主要分为两大类：第一类是原始数据法(Bouveyron 和 Brunet, 2014)[1]，该方法属于高维数据分析方法；第二类是从函数出发的投影法(Jacques 和 Preda, 2014)[2]，以无限维函数为基，进行数据拟合，寻找包含适当类别信息的子空间并且进行聚类。按照子空间搜索方式的差别，投影法区分为滤波法和自适应法(Chiou 和 Li, 2007[3]; 王德青等人, 2018[4])。

从实际角度来看，分析函数型数据时必须要进行将离散点拟合成曲线的步骤，然后对拟合的曲线进行聚类分析。因此，依据拟合过程是否独立于聚类过程，函数型聚类算法又可以分为一步法和两步法。两步法是曲线拟合过程和聚类过程分别进行的方法，在两步法中，既有基于模型的方法，如曲线密度近似表述基础上的高斯混合模型(Jacques 和 Preda, 2013)[5]，也有基于距离的方法，如非参数拟合基础上的系数 K-means 聚类(Abraham 和 Cornillon, 2003)[6]、加权距离 K-means 聚类(黄恒君, 2013)[7]、考虑组间差异的系数聚类(许腾腾等,2019)[8]等。两步法还包含了子空间搜寻与聚类同时开展的自适应方法，如函数型子空间分割聚类算法(Yamamoto 和 Hwang, 2017)[9]、函数型因子 K-means 算法(Yamamoto 和 Terada, 2014)[10]等。一步法是曲线拟合过程和聚类过程同时进行优化的方法。目前，关于一步法主要是基于模型的算法，在混合线性模型下开展，在拟合过程中，将类别信息设置为随机变量，继续对随机变量的分布进行分析，如 P 样条的线性混合聚类(Coffey 等, 2014)[11]、贝叶斯混合效应模型聚类(Giacofci 等, 2013)[12]等。

目前针对函数型聚类一步法的研究相对较少，主要采用了基于模型的方法，这些方法假设条件较为严格，而基于距离的方法则较为罕见。但事实上，函数型聚类一步法能够更好地权衡曲线拟合与聚类效果，既能提取重要的类别信息，又

能保留类别信息，因而理论上来说比两步法更为优越。黄恒君(2019)[13]将函数型聚类中的拟合过程与聚类过程结合，形成了基于距离的函数型聚类一步法，但是采用了惩罚的方式，目标函数较复杂，求解存在困难。而高海燕等(2020)[14]尝试将曲线拟合、降维和聚类纳入同一损失函数中，避免惩罚的方式实现要素并联，提出了基于非负矩阵分解的函数型聚类一步法(FNMF)，提高了现有函数型聚类分析方法的性能，推导了一种更加有效的优化算法来求解，为处理非负函数型数据提供了新的思路。

1.2.2 非负矩阵分解算法

非负矩阵分解(Nonnegative Matrix Factorization, NMF)是应用最为广泛的聚类方法之一，最初在开创性著作上作为一种矩阵分解技术(Lee 和 Seung, 1999)[15]被提出。NMF 将矩阵分解为两个矩阵的乘积，原始矩阵和分解矩阵均非负，因此 NMF 算法只适用于非负数据。后来，Ding 等(2005)[16]发现了 NMF 和 K-means 之间的联系，并进一步证明了 NMF 可以作为一种聚类方法。NMF 的非负约束，不仅有利于获得分解结果的可解释性，而且具有易于实现、占用存储空间小等诸多优势。NMF 的优势，使之获得了广泛的应用，比如：人脸识别(Li 和 Hou, 2001)[17]、文档分析(Xiong 和 Zang, 2014)[18]、图像标注(Tao 和 Li, 2017)[19]等。尽管 NMF 已经取得了良好的表现，但其仍然存在一定的局限性。

关于非负矩阵分解算法的研究趋于成熟，Ding(2010)[20]提出半非负矩阵分解，即 Semi-NMF，允许数据矩阵和基矩阵非负，更加适用于混合数据；Kong(2011)[21]提出了一种鲁棒的$l_{2,1}$-NMF，用$l_{2,1}$范数来代替 Frobenius 范数来测量重构误差，可以更好地应用于含噪声和离群点的数据；Deng(2011)[22]提出基于图正则化的 NMF 模型，即 GNMF，将图拉普拉斯引入到了 NMF 框架中，它包含基本的流形结构，利用了几何结构并考虑局部不变性；基于 Kong 等人和 Deng 等人的工作，Huang 等人(2014)[23]基于范数提出了一种鲁棒流形非负矩阵分解方法(RMNMF)，并在同一聚类框架下集成了 NMF 和谱聚类；Huang 等人(2017)[24]提出了带有自适应邻居(NMFAN)的 NMF 聚类方法，它根据每个数据点的局部连通性选择自适应的最优邻居来学习数据图，提高了图的质量，优化了聚类性能；Wang 等人(2021)[25]提出了一个稳健的双随机图正则化矩阵分解框架(RBSMF)用

于数据聚类，通过自适应学习相似图，图的更新和矩阵分解同时进行，使得学习的图更适合聚类。董文婷等(2023)[26]构建近邻图、最大熵图描述数据的局部结构和非局部结构，并使用$l_{2,1}$范数代价函数，提出一种鲁棒结构正则化非负矩阵分解方法；Liu(2012)[27]提出了一种新的半监督矩阵分解方法，将部分标签信息作为附加约束，充分利用了数据的标签信息。Yi 等(2022)[28]引入基于软标签矩阵的回归项，构建软标记矩阵与低维特征之间的关系，提出了一种半监督判别 NMF 方法。陈君航等(2023)[29]提出一种基于正交约束的广义可分离非负矩阵分解算法，限制迭代过程中关于行和列的迭代矩阵,确保得到行和列的特征,并获取更加精确的分解结果。侯兴荣等(2023)[30]采用一种新的数据局部相似性学习方法，同时学习数据的全局结构信息，从而能挖掘数据类内相似和类间相离的性质，提出了一种基于数据局部相似性学习的鲁棒非负矩阵分解算法。高海燕等(2023)[31]提出一种对称非负矩阵分解聚类算法，利用$l_{2,1}$范数作为损失函数缓解了噪声和异常值的影响，提高了算法的鲁棒性。

1.2.3 多视角聚类算法

大数据时代下，多视角数据可能通过不同的源头采集或是用于不同任务的不同特征进行表示，因此多视角聚类方法获得国内外学者的广泛研究和关注。

现有的多视角聚类算法可分为四类：①协同学习(Bickel 和 Scheffer, 2004[32]; Kumar, 2011[33]; Ye 等，2016[34]; Nie 等，2020[35])②多核学习(Men 和 Margolin，2014[36]; Zhu 等, 2018[37]; Wang 等, 2019[38])③多视角图聚类(Wang 等, 2016[39]; Saha 等, 2013[40]; Nie 等, 2016[41]; Cai 等, 2018[42]，夏冬雪等，2020[43])④多视角子空间聚类(Chaudhuri 等，2009[44]; Guo 等，2013[45]; Fan 等，2015[46]; Wang 等,2015[47]，吴峰等，2022[48])。多视角子空间聚类在多视角聚类研究中发展长远，尤其是基于非负矩阵分解的多视角子空间聚类。随着非负矩阵分解的普及和发展，Liu 等(2013)[49]首次研究了基于 NMF 的多视角聚类，为后续基于 NMF 的多视角聚类提供了思路；Zhang 等(2014)[50]研究了多流形正则化的 NMF 多视角聚类，并提出两种不同的融合方法；刘正等(2016)[51]提出一种基于 NMF 的特征加权多视角聚类算法；宗林林等(2017)[52]对每个视角分别引入局部流形学习，提出多流形 NMF 多视角聚类算法；Huang 等(2018)[53]结合 NMF 和 K-means 提出一

种鲁棒的多视角聚类方法。连佳琪等(2022)[54]利用多视图数据的多样性和差异性来学习表征，通过主成分分析来对数据进行全局结构的判别式学习，提出一种结构正则化多视图非负矩阵分解算法；杜虹燕等(2023)[55]通过各视角亲和矩阵自适应学习提取共识的亲和矩阵进行图嵌入来提取多视角数据共识局部结构信息；Li等(2023)[56]引入$l_{2,1}$范数来测量因式分解误差，设计了一种基于信息熵的自适应图方法，提出一种新的鲁棒多视图聚类方法。

事实上，函数型数据也以多元的形式出现，为此，学者们提出了许多关于多元函数型数据的聚类方法。例如：Jacques 和 Preda(2014)[2]、Schmutz 等(2020)[57]将函数型主成分作为随机变量，对基于模型的多元函数型聚类方法开展研究；Ieva 等(2013)[58]提出了多元函数型 K-means 聚类方法；Yamamoto 和 Hwang(2017)[9]将子空间分割技术与函数型子空间聚类结合提出了一种多元函数型聚类方法。以上多元函数型聚类方法均为将多个一元函数型的数据融合后进行分析，可能不足以充分挖掘数据中变量间互补信息，以及变量内部共有的信息。

姚晓红等(2022)[59]在多元函数型数据的生成以及聚类特征提取的过程中，结合非负矩阵分解，提出了一种基于半非负矩阵分解的多元函数型聚类模型。基于非负矩阵分解的多视角聚类方法具备多重优势，既可以有效地捕捉不同视角之间的共性和互补关系，又可以挖掘出潜藏的聚类特征。因此，对采用基于非负矩阵分解技术实现的多元函数型聚类问题的研究，无论是从理论还是实践层面，都具有重要的意义。

从以上国内外相关文献研究现状梳理中，可以看出，虽然学者在不同角度对于函数型聚类分析方法进行了大量的研究，但聚类性能相对较好的函数型聚类一步法却尚未形成成熟的体系。高海燕等(2020)[14]将非负矩阵分解作为降维手段，引入到函数型数据分析中来，构建的同时进行函数型数据生成和数据聚类特征提取的聚类一步法为本文提供了研究思路。

首先，现有的函数型聚类忽略了局部数据结构，导致无法保持相邻点的数据相关性。因此本文在引入图拉普拉斯的前提下，通过自适应学习相似图来提高图聚类的性能，考虑了数据的几何结构；其次，实际应用中所观测到的函数型数据大多含有噪声和异常值，故实验结果会因此受到影响，导致聚类性能差，本文重新考虑损失函数，以此提高模型的鲁棒性；考虑到多视角函数型数据存在缺失的

情况，本文利用指示矩阵填充实例，并且为每个视图自动分配适当的权重，平衡了噪声和数据缺失对于聚类的影响。本文将遵循已有函数型聚类分析研究思路，在利用非负矩阵分解的函数型聚类分析基本框架下，通过流形学习、双随机矩阵、自加权策略等技术，克服了传统函数型聚类方法的不足，丰富函数型聚类的一步法在单视角和多视角的相关应用研究。

1.3 研究思路及内容安排

本文以函数型数据为研究对象，主要侧重研究在非负矩阵框架下的函数型聚类算法。在相关文献综述的基础上(第 1 章)，回顾了函数型聚类及非负矩阵分解方法的发展历程(第 2 章)，针对现有方法所存在的问题，开展基于非负矩阵分解的函数型聚类方法研究。首先，提出了基于双随机图正则化矩阵分解的函数型聚类算法(第 3 章)，该框架可以有效挖掘数据内部几何结构，提高聚类算法的性能；由于实际应用中的函数型数据会包含噪声，为保证算法的鲁棒性，进一步提出基于鲁棒图正则化矩阵分解的函数型聚类算法(第 4 章)；函数型数据中通常以多元数据形式出现，且由于监测技术等原因，数据中存在缺失，一般的框架难以直接使用和处理此种情况，为此，本文提出自加权不完整多视角函数型聚类算法(第 5 章)。主要研究内容及思路如图 1.1 所示：

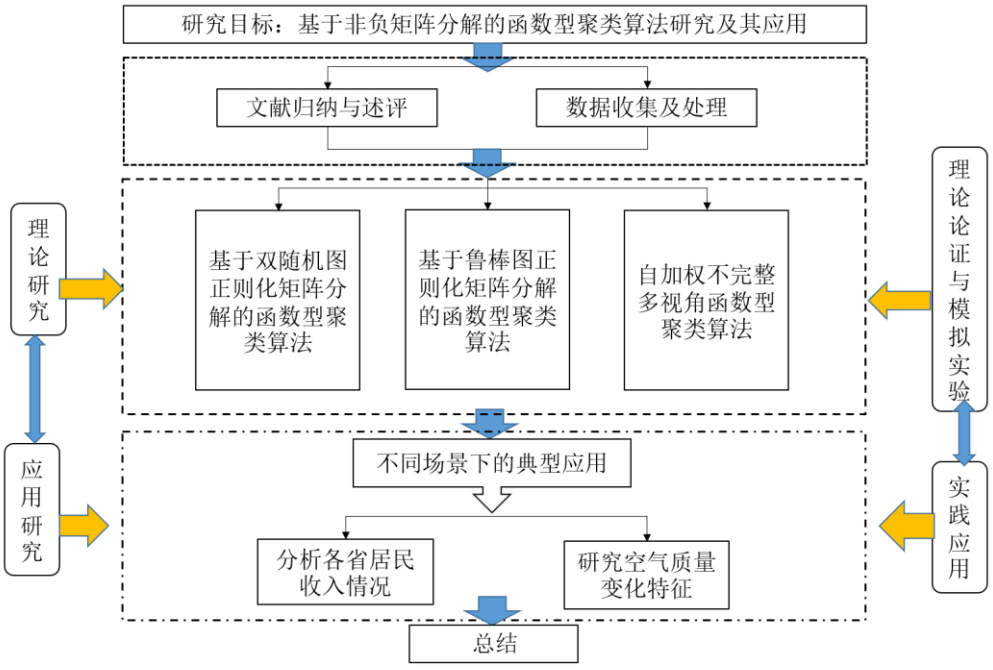


图 1.1 主要研究内容及思路

具体研究内容安排如下：

第一章，绪论。介绍了本文的研究背景及意义、研究现状、研究内容及创新点等。

第二章，预备知识。主要介绍了非负矩阵分解、基于非负矩阵分解的函数型聚类算法、图正则化矩阵分解和聚类评价指标。

第三章，基于双随机图正则化矩阵分解的函数型聚类算法(BSMFFC)。在已有方法的基础上，引入图拉普拉斯正则化项和双随机矩阵，构造最近邻图来模拟流形结构，构建可以挖掘数据几何结构且动态更新图学习的函数型聚类一步算法；其次，给出目标函数的更新公式，讨论了算法的时间复杂度；接着，通过随机模拟数据 I、Growth 成长数据和 TIMIT 语音数据进行模拟实验验证该方法的有效性；最后，通过对北京市空气污染物 NO2 的研究，证明该方法存在实际应用价值。

第四章，基于鲁棒图正则化矩阵分解的函数型聚类算法(FRMNMF)。利用$l_{2,1}$范数来衡量矩阵分解的误差，并结合流行学习项构建聚类算法模型；其次，给出了模型的优化求解算法和算法时间复杂度分析；接着，在随机模拟数据 II、Growth 数据、CanadianWeather 数据和 FatSpectrum 数据上进行模拟实验；最后，以城镇居民人均可支配收入数据为例进行聚类分析，证明算法具有实际应用性。

第五章，自加权不完整多视角函数型聚类算法(AIMFC)。通过构建不完备指示矩阵，定义每个视角的权重因子，增加正则化项，构建适用于不完备即缺失数据集的多视角聚类算法，并给出交替迭代更新规则。接着，在随机模拟数据 III 上进行模拟实验，验证算法可行性，最后，以北京市空气污染物小时浓度数据为对象，进行聚类分析，并采取不同变量作为不同视角进行分析，验证了所采用方法在实际应用中的价值。

第六章，结论与展望。对本文进行总结，并对后续工作进行展望。

1.4 创新点

本文主要开展基于非负矩阵分解的函数型聚类方法的相关研究工作，针对现有方法的一些不足之处，提出了三种改进的函数型聚类算法。本文主要的创新点有以下几个：

（1）构建一种基于双随机图正则化矩阵分解的函数型聚类算法

现有的函数型聚类方法多为两步法，即将数据拟合和聚类步骤分开进行，而本文在前期工作的基础上，利用非负矩阵分解的降维特性，结合了图拉普拉斯和双随机矩阵，提出的聚类方法适用于混合函数型数据，尊重数据空间的几何结构，提高了算法的聚类性能。

（2）构建可以处理噪声和异常值的鲁棒性函数型聚类算法

利用 Frobenius 范数来定义损失函数，由于每个数据点的误差以平方的形式进入目标函数，容易受到数据中噪声和异常值的影响，因此会降低算法性能。本文尝试用$l_{2,1}$范数来衡量矩阵分解的质量，降低噪声的影响，增强算法的鲁棒性。

（3）构建用于不完整多视角函数型数据的聚类算法

函数型数据通常以多元的形式出现，但是在数据分析中会出现某些数据点无法获得的情况造成数据缺失。针对这种情况，本文通过构建不完全指示矩阵，将缺少的实例填充到视图矩阵中；定义每个视角中的权重因子，自适应地为每个视角分配适当的权重，减少了噪声和缺乏实例的影响。

1.5 主要符号

为了更好地描述所提出的算法，在表 1.1 中总结了论文中的主要符号。

表 1.1 相关符号描述说明

符号	描述	符号	描述
$\boldsymbol{Y}$	原始数据矩阵	$\boldsymbol{\Lambda}$、$\boldsymbol{\Gamma}$	拉格朗日乘数
m	特征数	μ、ρ	ALM 罚系数
n	样本量	$\boldsymbol{Y}_v$	第v视角的原始离散矩阵
r	聚类数	n_v	视角数
$\boldsymbol{\Phi}$	曲线拟合过程中的基底矩阵	$\boldsymbol{O}_v$	第v视角的不完备指示矩阵
$\boldsymbol{U}$	非负矩阵分解的基矩阵	w_v	第v视角的权重因子
$\boldsymbol{V}$	系数矩阵	$\boldsymbol{V}^*$	视角间的共识度矩阵
$\boldsymbol{L}$	图拉普拉斯矩阵	$\|\cdot\|_{2,1}$	矩阵的$l_{2,1}$范数
λ、α、β	正则化参数	$\odot$	矩阵的哈达玛积

2 预备知识

2.1 非负矩阵分解

非负矩阵分解的概念由 Lee 等[15]于 1999 年首次提出，NMF 作为最流行的降维方法之一，因其良好的可解释性在聚类中得到广泛应用。设$\boldsymbol{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n]\in\mathbb{R}^{m\times n}$为数据矩阵，其中，$m$和$n$分别代表特征和样本的数量，每个样本为一个$m$维特征向量。NMF 的目标是找到两个非负矩阵$\boldsymbol{U}\in\mathbb{R}^{m\times k}$和$\boldsymbol{V}\in\mathbb{R}^{n\times k}$重构$\boldsymbol{X}$，即

$$\boldsymbol{X}\approx\boldsymbol{U}\boldsymbol{V}^T \tag{2.1}$$

其中，$k\ll\min\{m,n\}$，$\boldsymbol{U}$和$\boldsymbol{V}$分别解释为基矩阵和系数矩阵。为了获得$\boldsymbol{X}$的良好近似值，通常采用欧氏距离衡量重构误差，目标函数为

$$\min_{\boldsymbol{U}\geq 0,\boldsymbol{V}\geq 0}\left\|\boldsymbol{X}-\boldsymbol{U}\boldsymbol{V}^T\right\|_F^2 \tag{2.2}$$

其中，$\|\cdot\|_F$是 Frobenius 范数。对于任意矩阵$\boldsymbol{M}\in\mathbb{R}^{m\times n}$，$\boldsymbol{M}$的 Frobenius 范数定义为$\|\boldsymbol{M}\|_F=\sqrt{\sum_j^n\sum_i^m M_{i,j}^2}$。式(2.2)相对于$\boldsymbol{U}$或$\boldsymbol{V}$是凸的，利用乘法更新规则求解式(2.2)，更新公式为

$$\boldsymbol{U}_{ij}\leftarrow\boldsymbol{U}_{ij}\frac{(\boldsymbol{X}\boldsymbol{V})_{ij}}{(\boldsymbol{U}\boldsymbol{V}^T\boldsymbol{V})_{ij}},\quad \boldsymbol{V}_{ij}\leftarrow\boldsymbol{V}_{ij}\frac{(\boldsymbol{U}^T\boldsymbol{X})_{ij}}{(\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{V}^T)_{ij}} \tag{2.3}$$

2.2 基于非负矩阵分解的函数型聚类算法

(1) 曲线生成

定义一个独立同分布函数型数据样本$\boldsymbol{x}(t)=[x_1(t),x_2(t),\cdots,x_n(t)]^T$，$t\in\boldsymbol{T}=[a,b]$，其中$x_i(t)\ (i=1,2,\cdots,n)$是$L^2(\boldsymbol{T})$上的实值曲线。在既定空间中寻找一组基底函数，曲线可表示为$x_i(t)=\sum_{l=1}^p\alpha_{il}\phi_{il}(t)$，其中，$\boldsymbol{\phi}_i(t)=[\phi_{i1}(t),\phi_{i2}(t),\cdots,\phi_{ip}(t)]^T$为基函数，$\alpha_i(t)=[\alpha_{i1},\alpha_{i2},\cdots,\alpha_{ip}]^T$为系数列向量，如果每条曲线$x_i(t)$选取同样的基底函数$\boldsymbol{\phi}(t)=[\boldsymbol{\phi}_1(t),\boldsymbol{\phi}_2(t),\cdots,\boldsymbol{\phi}_p(t)]^T$进行拟合，则有：

$$\boldsymbol{x}(t)=\boldsymbol{A}^T\boldsymbol{\phi}(t) \tag{2.4}$$

其中，$\boldsymbol{A}=[\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_n]$是参数矩阵，$x(t)$中曲线之间的差异完全可由参数矩阵

$\boldsymbol{A}$确定。

设$\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n]\in\mathbb{R}^{m\times n}$为数据矩阵，$\boldsymbol{y_i}=[y_{i1},y_{i2},\ldots,y_{im}]^T$，$y_{ij}$为函数的带有噪音的离散观测值，由模型$y_{ij}=x_i(t_{ij})+\epsilon_{ij}$生成。从而，通过最小化目标函数

$$\|\boldsymbol{Y}-\boldsymbol{\Phi A}\|_F^2 \tag{2.5}$$

可得$\boldsymbol{A}$的估计结果。其中，$\boldsymbol{\Phi}=[\boldsymbol{\phi}_1(t),\boldsymbol{\phi}_2(t),\ldots,\boldsymbol{\phi}_p(t)]\in\mathbb{R}^{m\times p}$为基函数$\boldsymbol{\phi}_i(t)$在$m$个采样点$t_1,t_2,\ldots,t_m$取值形成的的基矩阵。进而根据式(2.4)可得到曲线$\boldsymbol{x}(t)$的估计。

(2) 基于 NMF 的聚类过程

对函数曲线$\boldsymbol{x}(t)$聚类，则转化为对矩阵$\boldsymbol{A}$聚类，对于高维数据的聚类问题，则可以通过降维来获得更好的聚类效果。因此借助 NMF 的降维作用和聚类特性，基于 NMF 的函数型聚类算法(FNMF)的优化目标函数为

$$\min_{\boldsymbol{V}\geq 0,\,\boldsymbol{V'V}=\boldsymbol{I}}\|\boldsymbol{Y}-\boldsymbol{\Phi UV}^T\|_F^2+\alpha||\boldsymbol{D_dU}\|_F^2 \tag{2.6}$$

式(2.6)中，由于系数矩阵$\boldsymbol{A}$一般并无非负要求，故松弛 NMF 的非负性约束，对式(2.5)中的$\boldsymbol{A}$采用 Semi-NMF 表示为$\boldsymbol{A}=\boldsymbol{UV}^T$，且满足$\boldsymbol{V}>0$，$\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}$。$\alpha$为正则化参数，引入惩罚项$\|\boldsymbol{D}_d\boldsymbol{U}\|_F^2$以防止过拟合。

利用惩罚更新规则求解式(2.6)可得更新规则：

$$\boldsymbol{U}_{ij}\longleftarrow\boldsymbol{U}_{ij}\left(\left(\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\alpha\boldsymbol{D}_d^T\boldsymbol{D}_d\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{Y}\right)\boldsymbol{V}\right)_{ij} \tag{2.7}$$

$$\boldsymbol{V}_{ij}\longleftarrow\boldsymbol{V}_{ij}\sqrt{\frac{\left(\boldsymbol{Y}^T\boldsymbol{\Phi U}\right)_{ij}^{+}+\left(\boldsymbol{V\Lambda}^{-}\right)_{ij}}{\left(\boldsymbol{Y}^T\boldsymbol{\Phi U}\right)_{ij}^{-}+\left(\boldsymbol{V\Lambda}^{+}\right)_{ij}}} \tag{2.8}$$

其中，$\boldsymbol{\Lambda}=\boldsymbol{V}^T\boldsymbol{Y}^T\boldsymbol{\Phi U}$，$\boldsymbol{\Lambda}^{+}=\frac{1}{2}\left(\boldsymbol{\Lambda}+\boldsymbol{\Lambda}\right)$，$\boldsymbol{\Lambda}^{-}=\frac{1}{2}\left(|\boldsymbol{\Lambda}|-\boldsymbol{\Lambda}\right)$。

2.3 图正则化非负矩阵分解

标准 NMF 只利用了全局数据信息，而忽略了相邻数据的局部结构。而图约束是保持数据几何结构的有力工具，通过限制样本之间的连接关系来优化聚类过程。Cai 等[22]提出了一种图正则化 NMF（GNMF），通过 KNN 图构造一个邻接矩阵来编码 NMF 的低维流形。GNMF 的目标函数为

$$\min_{\boldsymbol{U}\geq 0,\boldsymbol{V}\geq 0}\left\|\boldsymbol{X}-\boldsymbol{UV}^T\right\|_F^2+\lambda\mathrm{tr}\left(\boldsymbol{V}^T\boldsymbol{LV}\right) \tag{2.8}$$

其中，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$是拉普拉斯矩阵，$\boldsymbol{W}$为相似矩阵，度矩阵$\boldsymbol{D}$是对角阵，以$\boldsymbol{D}_{ii}=\sum_j^n\boldsymbol{W}_{ij}$为元素。Cai 等也给出了 GNMF 算法的更新规则：

$$\boldsymbol{U}_{ij} \longleftarrow \boldsymbol{U}_{ij}\frac{(\boldsymbol{X}\boldsymbol{V}^T)_ij}{(\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}^T)_{ij}} \tag{2.9}$$

$$\boldsymbol{V}_{ij} \longleftarrow \boldsymbol{V}_{ij}\frac{(\boldsymbol{U}^T\boldsymbol{X}+\lambda\boldsymbol{V}\boldsymbol{W})_ij}{(\boldsymbol{U}^T\boldsymbol{V}\boldsymbol{V}+\lambda\boldsymbol{V}\boldsymbol{D})_{ij}} \tag{2.10}$$

进一步，Peng 等开发了一种新的 GNMF 方法，将流形学习和特征学习集成到 NMF 中。一旦相似图构建，该图将在矩阵分解过程中保持不变，这对于 NMF 来说可能不是最优的，并且会限制聚类任务的性能。为此，Jia 等(2021)[60]提出了一种监督对称 NMF 方法(SymNMF)，可以自适应学习图并直接生成聚类结果。Huang 等(2017)[24]开发了一种具有自适应邻居的 NMF(NMFAN)，它学习了 NMF 的自适应图，并更好地利用了数据的流形结构。Sheng 等(2019)[61]提出了自适应局部学习正则化 NMF 聚类算法(ALLRNMF)，从矩阵分解的角度考虑了判别信息和数据流形。

2.4 聚类评价指标

为了评估模型的聚类性能，本文主要选取聚类精度(ACC)、归一化互信息(NMI)、纯度(PUR)、兰德指数(RI)这 4 个常用指标对类效果进行评估。接下来将分别介绍这 4 个聚类评估指标。

(1) 聚类精度 ACC

聚类精度 ACC 不仅可以发现聚类结果与真实类之间的一对一关系，还可以从对应的类中获取每个聚类所包含的数据点。具体来说，它的定义如下

$$ACC=\frac{\sum_{i=1}^{n}\delta\left(\operatorname{map}\left(r_i\right),l_i\right)}{n} \tag{2.11}$$

其中r_i是样本$\boldsymbol{x}_i$的聚类标签，l_i表示$\boldsymbol{x}_i$的真实标签。n是所有数据样本的数量。$map(r_i)$是最佳匹配函数，它可以置换所有的聚类结果，以最佳地将聚类标签映射到真实标签。$\delta(a,b)$是指标函数，如果$a=b$，则等于 1，否则等于 0。

(2) 归一化互信息 NMI

在聚类分析中，归一化互信息 NMI 是一种常用的度量聚类质量的指标。NMI 的定义如下:

$$NMI=\frac{\sum_{i=1}^{c}\sum_{j=1}^{c}n_{i,j}\log\frac{n_{i,j}}{n_i\hat{n}_j}}{\sqrt{\left(\sum_{i=1}^{c}n_i\log\frac{n_i}{n}\right)\left(\sum_{j=1}^{c}\hat{n}_j\log\frac{\hat{n}_j}{n}\right)}} \tag{2.12}$$

其中c表示类别数，n_i是属于聚类C_i $(1 \leq i \leq c)$的样本数。$\hat{n}_j$表示$L_j(1 \leq j \leq c)$类中包含的样本数，$n_{i,j}$表示聚类C_i和L_j类之间的重叠样本数。

(3) 纯度 PUR

聚类结果的纯度 PUR 常用于测量聚类中的样本分布情况。PUR 通过所有单个聚类纯度值的加权和来计算。纯度值通过以下方法获得

$$PUR = \sum_{i=1}^{c} \frac{n_i}{n} P(S_i), \quad P(S_i) = \frac{1}{n_i} \max_j n_i^j \tag{2.13}$$

其中 Si 具有大小n_i的特定簇，而n_i^j是分配给第j个簇的第i类样本的数量。

(4) 兰德指数 RI

兰德指数(Rand Index)可作为一种衡量聚类算法性能的工具，大致上体现了聚类算法将数据点分配到聚类中的准确度。具体地，兰德指数 RI 的计算公式如下所示：

$$\mathrm{RI} = \frac{a+b}{C_N^2} \tag{2.14}$$

其中a代表实际聚类结果中同处于一个类别的元素对数，b代表实际聚类结果中属于不同类别的元素对数，C_N^2是一个组合数。

3 基于双随机图正则化矩阵分解的函数型聚类算法

为提高聚类性能，本章提出了一种基于双随机图正则化矩阵分解的函数形聚类算法。其中，通过引入图拉普拉斯正则化项，将流形学习和函数型聚类方法相结合，并利用双随机矩阵自动更新图学习项，更加有效的利用了数据的内部几何结构。此外，还给出了模型的求解方式，并对其时间复杂度进行了计算。最后，本文在模拟数据集 I、Growth 和 TIMIT 语音数据集上对算法进行了聚类性能验证，同时在北京市 NO2 日均浓度数据上进行了实例验证和分析，证明该算法具有较卓越的空间分布识别能力，具有实用价值。

3.1 问题概述

聚能性能与数据内部相似性高度相关，因此，现有的函数型聚类方法存在一定的局限性：只考虑了降低数据的维度，没有考虑到数据的局部结构，导致无法保持相邻点的数据相关性，从而影响聚类的效果。在非函数型数据聚类中，Deng 等(2011)[22]提出了基于图正则化的 NMF 模型(GNMF)，将拉普拉斯图引入到 NMF 框架中，它包含基本的流形结构，利用几何结构并考虑了局部不变性。然而基于图的 NMF 模型依赖于初始图，如果 KNN 构造的图过于简单，则不能很好地表示原始数据结构，为此，Huang 等(2017)[24]提出了局部自适应结构正则化非负矩阵分解算法(NMFAN)，为每个数据点分配自适应地最优邻居，以此构造最优数据图，更好的利用了数据的流形结构；刘威等(2023)[62]通过先验信息构造约束图矩阵，引入 PCP-SDP 方法，提出了一种基于约束图正则的块稀疏对称非负矩阵分解。

基于上述启发，为了处理函数型数据中所存在的这些问题，我们将 GNMF 与函数型聚类方法结合，并使用双随机矩阵从输入的低质量图中自适应地学习更好的相似图，提出了一种基于图正则化矩阵分解的函数型数据聚类框架。本章的后续内容组织如下：3.2 提出了所提出 BSMFFC 模型的框架；3.3 推导了一种有效的优化算法来求解所提出的目标函数；3.4 为实验，利用随机模拟数据 I，Growth 成长数据集和 TIMIT 语音数据集来验证所提出的 BSMFFC 算法的优越的聚类性能，并且对收敛速度和参数灵敏度都进行了分析；3.5 为实例应用，利用 2018 年

北京市二氧化氮(NO2)污染物日均浓度数据对监测站点进行聚类应用；3.6 为本章小结。

3.2 算法模型

假设有一个原始的数据矩阵$\boldsymbol{Y} \in \mathbb{R}^{m \times n}$，将图正则化项引入基于非负矩阵分解的函数型聚类算法中，则有：

$$\min_{\boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0} \left\| \boldsymbol{Y} - \boldsymbol{\Phi} \boldsymbol{U} \boldsymbol{V}^T \right\|_F^2 + \lambda \mathrm{tr}\left(\boldsymbol{V}^T \boldsymbol{L} \boldsymbol{V} \right) \tag{3.1}$$

为了构造高质量的数据图来提高聚类性能，我们通过 KNN 算法给定一个初始的低质量图$\boldsymbol{H}$，学习基于$\boldsymbol{H}$的高质量图$\boldsymbol{W}$。一般来说，$\boldsymbol{W}$应该具有以下性质。首先，$\boldsymbol{W}$描述了每两个数据样本成为邻居的概率。所以每一行$\boldsymbol{W}$的和是 1，即，$\boldsymbol{W}\mathbf{1} = \mathbf{1}$。那么$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W} = \boldsymbol{I} - \boldsymbol{W}$，其中，$\boldsymbol{I}$是单位矩阵。第二，$\boldsymbol{W}$代表每对点的相似度。所以它必须是非负对称的，也就是说，$\boldsymbol{W} = \boldsymbol{W}^T$，$\boldsymbol{W} \geq 0$。有了在相似矩阵$\boldsymbol{W}$上的上述约束(即，$\boldsymbol{W} \geq 0$，$\boldsymbol{W} = \boldsymbol{W}^T$，$\boldsymbol{W}\mathbf{1} = \mathbf{1}$)，又称为双随机矩阵[63]。此外，根据图论[64]，每个顶点不允许单独连接。所以我们加上图约束$\mathrm{diag}(\boldsymbol{W}) = 0$。最后，我们可以通过求解从初始输入图$\boldsymbol{H}$中学习一个双随机矩阵作为新图，则有：

$$\begin{aligned} &\min_{\boldsymbol{W}} \|\boldsymbol{W} - \boldsymbol{H}\|_F^2 \\ &\text{s.t.} \quad \boldsymbol{W} \geq 0,\ \boldsymbol{W} = \boldsymbol{W}^T,\ \boldsymbol{W}\mathbf{1} = \mathbf{1},\ \mathrm{diag}(\boldsymbol{W}) = \mathbf{0} \end{aligned} \tag{3.2}$$

将式(3.1)与式(3.2)结合，则提出了以下基于双随机图正则化矩阵分解的函数型聚类算法(BSMFFC)：

$$\begin{aligned} &\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} \frac{1}{2} \left\| \boldsymbol{Y} - \boldsymbol{\Phi} \boldsymbol{U} \boldsymbol{V}^T \right\|_F^2 + \lambda \mathrm{tr}\left(\boldsymbol{V}^T \boldsymbol{L} \boldsymbol{V} \right) + \alpha \|\boldsymbol{W} - \boldsymbol{H}\|_F^2 + \frac{\beta}{2} \|\boldsymbol{U}\|_F^2 \\ &\text{s.t.} \quad \boldsymbol{V} \geq 0,\ \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I},\ \boldsymbol{W} \geq 0,\ \boldsymbol{W} = \boldsymbol{W}^T,\ \boldsymbol{W}\mathbf{1} = \mathbf{1},\ \mathrm{diag}(\boldsymbol{W}) = \mathbf{0} \end{aligned} \tag{3.3}$$

其中，λ、α和β为正则化参数，λ平衡流形学习项，α控制图学习项的权重，β为调节参数。其中，正交约束$\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$是为了保证算法有唯一解，消除算法的不确定性，去掉对于$\boldsymbol{U}$的约束可以让算法适用于混合数据。

3.3 求解算法

3.3.1 优化求解

基于增广拉格朗日乘子法 (Augmented Lagrange Method，ALM)[65]，我们开发了一种新的迭代算法求解式(3.3)，便于同时更新学习双随机图矩阵$\boldsymbol{W}$和聚类指示矩阵$\boldsymbol{V}$。为了易于求解式(3.3)，引入辅助变量$\boldsymbol{Z}$和$\boldsymbol{E}$，则式(3.3)可等价写为如下问题：

$$\begin{aligned}&\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{Z},\boldsymbol{E},\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{E}\|_F^2+\lambda \mathrm{tr}\left(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{Z}\right)+\alpha\|\boldsymbol{W}-\boldsymbol{H}\|_F^2+\frac{\beta}{2}\|\boldsymbol{U}\|_F^2\\ &\text{s.t.}\ \ \boldsymbol{Z}\geq 0,\ \boldsymbol{Z}=\boldsymbol{V},\ \boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I},\ \boldsymbol{E}=\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T\\ &\qquad\boldsymbol{W}\geq 0,\ \boldsymbol{W}=\boldsymbol{W}^T,\ \boldsymbol{W}\boldsymbol{1}=\boldsymbol{1},\ \mathrm{diag}(\boldsymbol{W})=\boldsymbol{0}\end{aligned} \tag{3.4}$$

式(3.4)的增广拉格朗日函数为

$$\begin{aligned}&\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{Z},\boldsymbol{E},\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{E}\|_F^2+\lambda \mathrm{tr}\left(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{Z}\right)+\alpha\|\boldsymbol{W}-\boldsymbol{H}\|_F^2+\frac{\beta}{2}\|\boldsymbol{U}\|_F^2\\ &\qquad+\frac{\mu}{2}\left\|\boldsymbol{Z}-\boldsymbol{V}+\frac{\boldsymbol{\Lambda}}{\mu}\right\|_F^2+\frac{\mu}{2}\left\|\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T-\boldsymbol{E}+\frac{\boldsymbol{\Gamma}}{\mu}\right\|_F^2+f(\boldsymbol{\Lambda},\boldsymbol{\Gamma})\\ &\text{s.t.}\ \ \boldsymbol{Z}\geq 0,\ \boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I},\ \boldsymbol{W}\geq 0,\ \boldsymbol{W}=\boldsymbol{W}^T,\ \boldsymbol{W}\boldsymbol{1}=\boldsymbol{1},\ \mathrm{diag}(\boldsymbol{W})=0\end{aligned} \tag{3.5}$$

其中，μ是控制两个等式约束$\boldsymbol{Z}=\boldsymbol{V}$和$\boldsymbol{E}=\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T$的罚系数。$\boldsymbol{\Lambda}\in\mathbb{R}^{n\times k}$和$\boldsymbol{\Gamma}\in\mathbb{R}^{m\times n}$是拉格朗日乘数，且$f(\boldsymbol{\Lambda},\boldsymbol{\Gamma})=\frac{1}{2\mu}\left(\|\boldsymbol{\Lambda}\|_F^2+\|\boldsymbol{\Gamma}\|_F^2\right)$。

目标函数式(3.5)关于变量$\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{Z}$、$\boldsymbol{E}$和$\boldsymbol{W}$是非凸的，求解全局最优解是困难的。根据最优化理论[66]，可以优化仅含一个变量的目标函数，同时固定其他变量。因此，目标函数式(3.5)可分解成几个子问题进行简化考虑，采用交替迭代策略求解，直到收敛。

(1) 更新$\boldsymbol{E}$。固定其他变量，将式(3.5)简化为关于$\boldsymbol{E}$的优化子问题，对应的拉格朗日函数为

$$f(\boldsymbol{E})=\frac{1}{2}\|\boldsymbol{E}\|_F^2+\frac{\mu}{2}\left\|\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T-\boldsymbol{E}+\frac{\boldsymbol{\Gamma}}{\mu}\right\|_F^2 \tag{3.6}$$

式(3.6)关于$\boldsymbol{E}$求偏导，令$\frac{\partial f(\boldsymbol{E})}{\partial \boldsymbol{E}}=0$，有

$$\frac{\partial f(\boldsymbol{E})}{\partial \boldsymbol{E}}=(1+\mu)\boldsymbol{E}-\mu\left(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T+\frac{\boldsymbol{\Gamma}}{\mu}\right)=0$$

从而

$$\boldsymbol{E}=\frac{\mu}{1+\mu}\left(\boldsymbol{E}+\frac{\boldsymbol{\Gamma}}{\mu}\right) \tag{3.7}$$

(2) 更新$\boldsymbol{U}$。固定其他变量，关于的$\boldsymbol{U}$简化优化子问题为

$$\min_{\boldsymbol{U}}\frac{\mu}{2}\left\|\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{T}-\boldsymbol{E}+\frac{\boldsymbol{\Gamma}}{\mu}\right\|_{F}^{2}+\frac{\beta}{2}\|\boldsymbol{U}\|_{F}^{2} \tag{3.8}$$

式(3.8)的拉格朗日函数

$$f(\boldsymbol{U})=\frac{\mu}{2}\left\|\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{T}-\boldsymbol{E}+\frac{\boldsymbol{\Gamma}}{\mu}\right\|_{F}^{2}+\frac{\beta}{2}\|\boldsymbol{U}\|_{F}^{2}$$

关于$\boldsymbol{U}$求偏导，并令其为零，利用正交约束$\boldsymbol{V}^{T}\boldsymbol{V}=\boldsymbol{I}$得到

$$\boldsymbol{U}=\left(\boldsymbol{\Phi}^{T}\boldsymbol{\Phi}+\frac{\beta}{\mu}\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}^{T}\left(\boldsymbol{Y}-\boldsymbol{E}+\frac{\boldsymbol{\Gamma}}{\mu}\right)\boldsymbol{V} \tag{3.9}$$

需要说明的是，对$\boldsymbol{V}$施加正交约束$\boldsymbol{V}^{T}\boldsymbol{V}=\boldsymbol{I}$，除了保证解的唯一性之外，在求解$\boldsymbol{U}$时也避免了求大矩阵的逆，从而降低了计算成本。

(3) 更新$\boldsymbol{V}$。固定其他变量，将式(3.5)简化为关于$\boldsymbol{V}$的优化子问题

$$\min_{\boldsymbol{V}^{T}\boldsymbol{V}=\boldsymbol{I}}\lambda\operatorname{Tr}\left(\boldsymbol{V}^{T}\boldsymbol{L}\boldsymbol{Z}\right)+\frac{\mu}{2}\left\|\boldsymbol{Z}-\boldsymbol{V}+\frac{\boldsymbol{\Lambda}}{\mu}\right\|_{F}^{2}+\frac{\mu}{2}\left\|\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{T}-\boldsymbol{E}+\frac{\boldsymbol{\Gamma}}{\mu}\right\|_{F}^{2} \tag{3.10}$$

将目标函数式(3.10)展开，去掉与$\boldsymbol{V}$无关的项，得到

$$\begin{aligned}\min_{\boldsymbol{V}^{T}\boldsymbol{V}=\boldsymbol{I}}\frac{\mu}{2}\operatorname{tr}&\left(\boldsymbol{V}\boldsymbol{V}^{T}-2\boldsymbol{V}^{T}\boldsymbol{Z}-\frac{2}{\mu}\boldsymbol{V}^{T}\boldsymbol{\Lambda}+\frac{2\lambda}{\mu}\boldsymbol{V}^{T}\boldsymbol{L}\boldsymbol{Z}\right.\\&\left.+\left(\boldsymbol{V}^{T}\boldsymbol{E}^{T}\boldsymbol{\Phi}\boldsymbol{U}-\boldsymbol{V}^{T}\boldsymbol{X}^{T}\boldsymbol{\Phi}\boldsymbol{U}-\frac{2}{\mu}\boldsymbol{V}^{T}\boldsymbol{\Gamma}^{T}\boldsymbol{\Phi}\boldsymbol{U}\right)\right)\end{aligned} \tag{3.11}$$

可以将式(3.11)写为如下简便形式

$$\begin{aligned}&\min_{\boldsymbol{V}\boldsymbol{V}^{T}=\boldsymbol{I}}\|\boldsymbol{V}-\boldsymbol{P}\|_{F}^{2}\\&\Longleftrightarrow\min_{\boldsymbol{V}\boldsymbol{V}^{T}=\boldsymbol{I}}\operatorname{tr}\left(\boldsymbol{V}\boldsymbol{V}^{T}-2\boldsymbol{V}^{T}\boldsymbol{P}+\boldsymbol{P}\boldsymbol{P}^{T}\right)\\&\Longleftrightarrow\max_{\boldsymbol{V}\boldsymbol{V}^{T}=\boldsymbol{I}}\operatorname{tr}(\boldsymbol{V}^{T}\boldsymbol{P})\end{aligned} \tag{3.12}$$

其中，

$$\boldsymbol{P}=\left(\boldsymbol{Z}+\frac{\boldsymbol{\Lambda}}{\mu}\right)+\left(\boldsymbol{Y}^{T}-\boldsymbol{E}^{T}+\frac{\boldsymbol{\Gamma}^{T}}{\mu}\right)\boldsymbol{\Phi}\boldsymbol{U}-\frac{\lambda}{\mu}\boldsymbol{L}\boldsymbol{Z} \tag{3.13}$$

下面通过目标函数式(3.12)求解$\boldsymbol{V}$。参考 Yu 和 Shi(2003)[67]中的定理 1。根据奇异值分解，设式(3.13)中的矩阵$\boldsymbol{P}$可分解为

$$\boldsymbol{P}=\boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{F}^{T}$$

其中，$\boldsymbol{G}\in\mathbb{R}^{n\times n}$，$\boldsymbol{\Sigma}\in\mathbb{R}^{n\times k}$，$\boldsymbol{F}\in\mathbb{R}^{k\times k}$，$k<n$。从而有

$$\operatorname{tr}(\boldsymbol{V}^{T}\boldsymbol{P})=\operatorname{tr}(\boldsymbol{V}^{T}\boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{F}^{T})=\operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{F}^{T}\boldsymbol{V}^{T}\boldsymbol{G})=\operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{M})=\sum_{i}\sigma_{ii}m_{ii}$$

其中，$\boldsymbol{M}=\boldsymbol{F}^T\boldsymbol{V}^T\boldsymbol{G}\in\mathbb{R}^{k\times n}$，且易证明$\boldsymbol{M}$是正交的。因此，$m_{ij}\in[-1,1]$。注意到$\sigma_{ii}\geq 0$，从而

$$\mathrm{tr}(\boldsymbol{V}^T\boldsymbol{P})=\sum_i\sigma_{ii}m_{ii}\leq\sum_i\sigma_{ii} \tag{3.14}$$

当$\boldsymbol{M}=\boldsymbol{I}_{k\times n}$时，式(3.14)取“=”成立，即$\boldsymbol{F}^T\boldsymbol{V}^T\boldsymbol{G}=\boldsymbol{I}_{k\times n}$，这意味着

$$\boldsymbol{V}=\boldsymbol{G}\boldsymbol{F}^T \tag{3.15}$$

因此，式(3.12)的解$\boldsymbol{V}$可由式(3.15)求出，即$\boldsymbol{V}=\boldsymbol{G}\boldsymbol{F}^T$，这里$\boldsymbol{G}$和$\boldsymbol{F}$分别是式(3.13)中矩阵$\boldsymbol{P}$的 SVD 分解的左、右奇异矩阵。

(4) 更新$\boldsymbol{Z}$。固定其他变量，优化关于$\boldsymbol{Z}$的目标函数如下：

$$\min_{\boldsymbol{Z}\geq 0}\lambda\mathrm{tr}\left(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{Z}\right)+\frac{\mu}{2}\left\|\boldsymbol{Z}-\boldsymbol{V}+\frac{\boldsymbol{\Lambda}}{\mu}\right\|_F^2 \tag{3.16}$$

将式(3.16)展开，去掉与$\boldsymbol{Z}$无关的项，得到

$$\min_{\boldsymbol{Z}\geq 0}\mathrm{tr}\left(\boldsymbol{Z}^T\boldsymbol{Z}-2\boldsymbol{Z}^T\boldsymbol{V}+\frac{2}{\mu}\boldsymbol{Z}^T\boldsymbol{\Lambda}+\frac{2\lambda}{\mu}\boldsymbol{Z}^T\boldsymbol{L}\boldsymbol{V}\right)$$

式(3.16)等价于

$$\min_{\boldsymbol{Z}\geq 0}\left\|\boldsymbol{Z}-\left(\boldsymbol{V}-\frac{\boldsymbol{\Lambda}}{\mu}-\frac{\lambda}{\mu}\boldsymbol{L}\boldsymbol{V}\right)\right\|_F^2\Longleftrightarrow\min_{\boldsymbol{Z}\geq 0}\|\boldsymbol{Z}-\boldsymbol{Q}\|_F^2 \tag{3.17}$$

其中，$\boldsymbol{Q}=\boldsymbol{V}-\frac{\boldsymbol{\Lambda}}{\mu}-\frac{\lambda}{\mu}\boldsymbol{L}\boldsymbol{V}$，求解式(3.17)易得

$$\boldsymbol{Z}=\max\{\boldsymbol{Q},\,\boldsymbol{0}\} \tag{3.18}$$

(5) 更新$\boldsymbol{W}$。固定其他变量，考虑关于$\boldsymbol{W}$的子问题

$$\begin{aligned}&\min_{\boldsymbol{W}}\ \lambda\mathrm{tr}\left(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{Z}\right)+\alpha\|\boldsymbol{W}-\boldsymbol{H}\|_F^2\\&\text{s.t. }\boldsymbol{W}\geq 0,\ \boldsymbol{W}=\boldsymbol{W}^T,\ \boldsymbol{W}\boldsymbol{1}=\boldsymbol{1},\ \mathrm{diag}(\boldsymbol{W})=\boldsymbol{0}\end{aligned} \tag{3.19}$$

设$\boldsymbol{R}=\boldsymbol{H}+\frac{\lambda}{2\alpha}\boldsymbol{V}\boldsymbol{Z}^T$，式(3.19)可以改写为

$$\min_{\boldsymbol{W}}\|\boldsymbol{W}-\boldsymbol{R}\|_F^2,\quad\text{s.t. }\boldsymbol{W}\geq 0,\ \boldsymbol{W}=\boldsymbol{W}^T,\ \boldsymbol{W}\boldsymbol{1}=\boldsymbol{1},\ \mathrm{diag}(\boldsymbol{W})=\boldsymbol{0} \tag{3.20}$$

为了解式(3.20)，可以分成如下两个子问题

$$\min_{\boldsymbol{W}}\|\boldsymbol{W}-\boldsymbol{R}\|_F^2,\quad\text{s.t. }\boldsymbol{W}=\boldsymbol{W}^T,\ \boldsymbol{W}\boldsymbol{1}=\boldsymbol{1} \tag{3.21}$$

$$\min_{\boldsymbol{W}}\|\boldsymbol{W}-\boldsymbol{R}\|_F^2,\quad\text{s.t. }\boldsymbol{W}\geq 0,\ \mathrm{diag}(\boldsymbol{W})=\boldsymbol{0} \tag{3.22}$$

交替求解式(3.21)和式(3.22)，并相互投射彼此的解。首先获得子问题式(3.21)的最优解$\boldsymbol{W_1}$，并将$\boldsymbol{W_1}$视为子问题式(3.22)的$\boldsymbol{R}$。其次获得最优解子问题式(3.22)的$\boldsymbol{W_2}$，将$\boldsymbol{W_2}$看做子问题式(3.21)的$\boldsymbol{R}$。然后反复迭代执行以上两个过程直到收敛。由 Von Neumann 连续投影引理[68]可以保证上述求解策略的收敛性。该引理从理

论上证明了互投影策略的解最终收敛于原问题式(3.20)的全局最优解。

参考 Zass 和 Shashua(2007)[69]的文献，子问题式(3.21)的最优解是

$$\boldsymbol{W}_1 = \boldsymbol{S} + \frac{n + \mathbf{1}^T \boldsymbol{S} \mathbf{1}}{n^2} \boldsymbol{\mathcal{I}}^T - \frac{1}{n} \boldsymbol{S} \boldsymbol{\mathcal{I}}^T - \frac{1}{n} \boldsymbol{\mathcal{I}}^T \boldsymbol{S} \tag{3.23}$$

其中，$\boldsymbol{S} = \frac{\boldsymbol{R} + \boldsymbol{R}^T}{2}$，$\mathbf{1}$是所有元素都为 1 的向量，而$\boldsymbol{\mathcal{I}}$是所有元素都为 1 的方矩阵。同时，子问题式(3.22)很容易通过以下方法解决

$$\boldsymbol{W}_2 = \max\{\boldsymbol{R}, \mathbf{0}\}, \quad \operatorname{diag}(\boldsymbol{W}_2) = \mathbf{0} \tag{3.24}$$

(6) 更新参数μ、$\boldsymbol{\Lambda}$和$\boldsymbol{\Gamma}$。在依次更新变量$\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{Z}$、$\boldsymbol{E}$和$\boldsymbol{W}$后，采用简单梯度上升更新 ALM 算法中拉格朗日乘子$\boldsymbol{\Lambda}$和$\boldsymbol{\Gamma}$，有

$$\begin{aligned} \boldsymbol{\Lambda} &= \boldsymbol{\Lambda} + \mu(\boldsymbol{Z} - \boldsymbol{V}) \\ \boldsymbol{\Gamma} &= \boldsymbol{\Gamma} + \mu\left(\boldsymbol{X} - \boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{E}\right) \\ \mu &= \rho\mu \end{aligned} \tag{3.25}$$

其中，$\rho > 1$是控制收敛速度的参数，ρ越大，收敛所需的迭代次数越少。

3.3.2 算法流程

综上分析，我们提出的基于双随机图正则化矩阵分解的函数型聚类算法(BSMFFC)的整体优化过程如算法 3.1 所示。

算法 3.1　基于双随机图正则化矩阵分解的函数型聚类算法(BSMFFC)

输入：数据矩阵$\boldsymbol{Y}$，聚类数r，基底矩阵Φ，初始图$\boldsymbol{H}$，参数λ，α和β。

过程：

1:初始化：$\varepsilon = 10^{-2}$，$t = 1$，$maxiter = 100$，用 K-means 初始化$\boldsymbol{V}_0$

2:for $t = 1, 2, \cdots, \text{maxiter}$

3: 固定其他变量，根据式(3.7)更新$\boldsymbol{E}$。

4: 固定其他变量，根据式(3.9)更新$\boldsymbol{U}$。

5: 固定其他变量，通过式(3.15)更新$\boldsymbol{V}$。

6: 固定其他变量，根据式(3.18)更新$\boldsymbol{Z}$。

7: 固定其他变量，根据式(3.23)和式(3.24)更新$\boldsymbol{W}$。

8: 根据式(3.25)更新 ALM 的参数μ，$\boldsymbol{\Lambda}$和$\boldsymbol{\Gamma}$。

9: if $t > maxiter$, $\|\boldsymbol{Z} - \boldsymbol{V}\|_\infty < \varepsilon, \left\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{E}\right\|_\infty < \varepsilon$

10: break

11: end if

12: end for

输出：$\boldsymbol{U}$，$\boldsymbol{V}$，$\boldsymbol{W}$，簇划分$\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$

算法 3.1 的收敛性依赖于 ALM 框架的收敛性，关于 ALM 框架的收敛性讨论和证明过程可参见 Lin 等(2007)[65]的文献。因此，目标函数式(3.3)将单调减少

到一个平稳点，保证了 BSMFFC 至少能找到一个局部最优解。

3.3.3 计算复杂度分析

本节进一步研究 BSMFFC 算法的计算时间复杂度问题。其中，n代表样本总量，m代表特征数量，r代表类别数，p代表 B-样条基底数量。按照算法 3.1 的流程，参数$\boldsymbol{E}$、$\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{Z}$和$\boldsymbol{W}$的的更新操作是决定算法计算复杂度的关键。如果迭代次数为t，则参数$\boldsymbol{E}$、$\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{Z}$和$\boldsymbol{W}$的时间复杂度分别为$O(mnr)$、$O(npr)$、$O(n^2r)$、$O(n^2r)$和$O(n^2r)$。通常情况下，迭代达到一定次数之后算法会停止。因此，算法 3.1 进行t次迭代，其更新的总时间复杂度为$O(n^2rt+nprt)$。

3.4 模拟实验

为了评估我们所提 BSMFFC 算法在聚类任务中的有效性，在模拟数据集 I、Growth 成长数据集和 TIMIT 语音数据集上与 6 种先进的聚类算法进行了比较，包括经典的聚类算法(K-means)，鲁棒图正则化 NMF 算法(RMNMF)[23]；鲁棒双随机矩阵图正则化 NMF 算法(RBSMF)[25]，基于 NMF 的函数型数据聚类算法(FNMF)[14]；函数型聚类两步法(TA)[70]；函数型聚类一步法(FCOF)[13]。

本实验中的所有代码均在 MATLAB R2018a 软件实现，实验的计算机环境为：12th Gen Intel(R) Core(TM) i5-12500H 2.50 GHz，内存 16GB，Windows11 64 位操作系统。

3.4.1 实验数据

（1）随机模拟数据 I

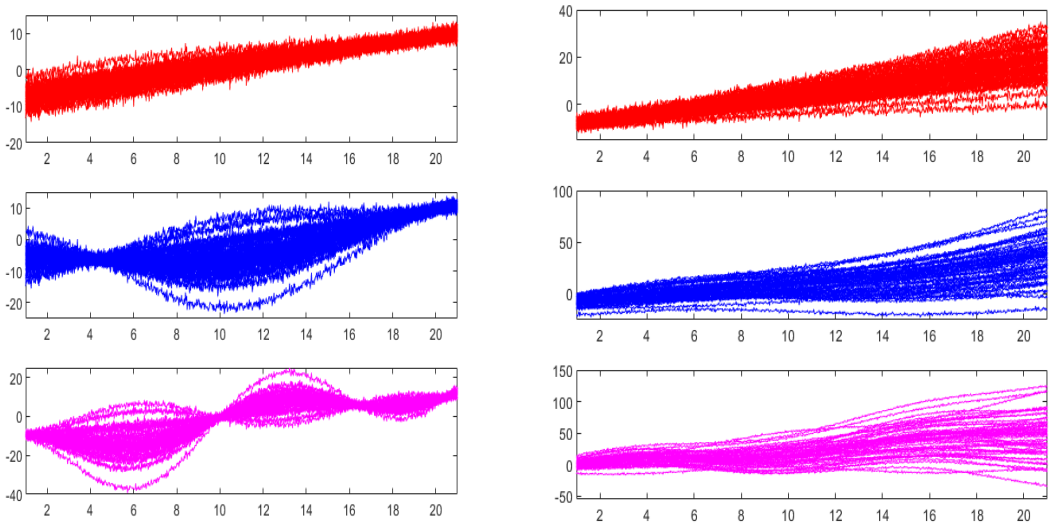
根据 Jacques 和 Preda(2014)[2]的研究方法，模拟生成了两组 3 类的数据集$X_1(t)$和$X_2(t)$。函数型数据通过将三角函数和多项式函数进行线性组合的方法所生成，公式如下：

$$X_1(t)=-\frac{21}{2}+t+kU_1cos\left(k\frac{t}{10}\right)+kU_1sin\left(k+\frac{t}{10}\right)+\epsilon(t)$$

$$X_2(t)=-\frac{21}{2}+t+kU_1sin\left(k\frac{t}{10}\right)+kU_2cos\left(k+\frac{t}{10}\right)+kU_3\left(\left(\frac{t}{10}\right)^2+\frac{t}{10}+1\right)+\epsilon(t)$$

其中，U_i是随机变量，$U_i\sim N(1,1)$，而$\epsilon(t)\sim N(0,1)$是高斯白噪声。自变量$t\in[1,21]$，

一共取 1001 个离散时间点，k分别取 1、3、5，则每个变量可以生成三类数据。每类随机生成 50 条曲线，共计 150 条曲线。



(a) $X_1(t)$ (b) $X_2(t)$

图 3.1 随机模拟数据 I

（2）Growth 成长数据

R 扩展包 fda 提供的 Growth 成长数据集，包括 39 名男孩和 54 名女孩 31 个阶段从 1 岁到 18 岁的身高数据。在图 3.2 中，样本和类别标签通过横坐标表示年龄，通过纵坐标表示身高，以不同颜色或线型的方式区分不同性别。

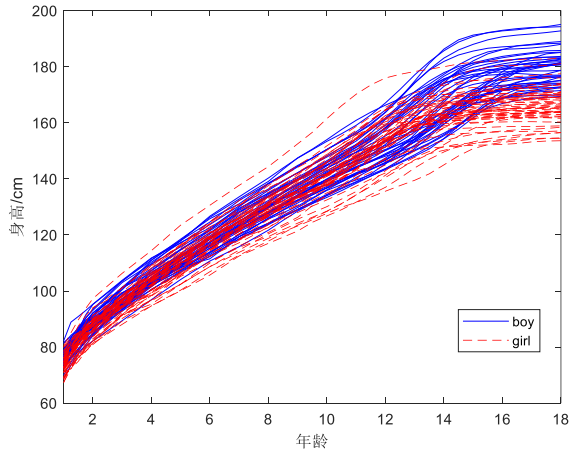


图 3.2 Growth 成长数据

（3）TIMIT 语音数据

TIMIT 语音数据库是由来自美国 8 个不同的地区共计 630 位说话者朗读 10 个特定句子所组成的，该数据库已经成为语音识别领域的标准，广泛应用于语音识别研究方面。本文采用了 TIMIT 语音数据库中名为 SA1 的语音数据作为分析对象。该数据集包含了数字信号处理后的音素数据以及相应的音素类别标签。图 3.3 所示的样本示例呈现了原始数据和类别标签，不同的线型代表不同的音素类

别。

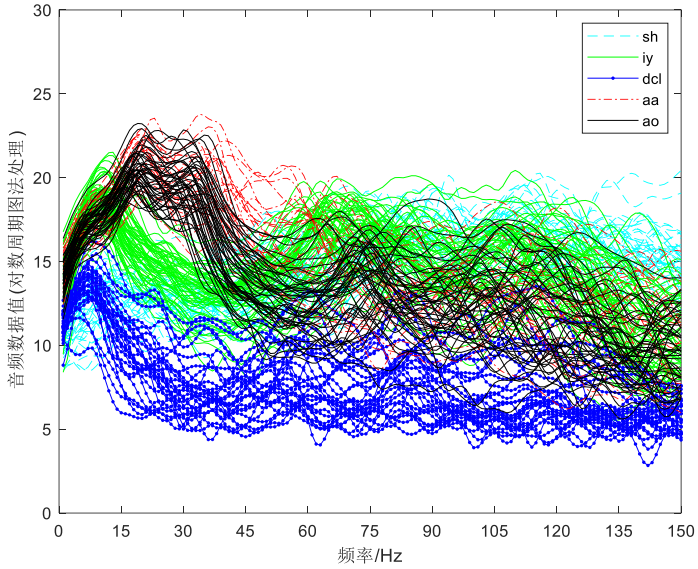


图 3.3 TIMIT 语音数据及类别标签的一个样本示例

3.4.2 参数设置

在 BSMFFC 方法中，选定的参数分别为：(1) 针对生成的随机模拟数据 I，包括$X_1(t), X_2(t)$，每个变量分别包含 3 类数据，类别数$r=3$；Growth 成长数据中性别为男孩和女孩，则取类别数$r=2$；在进行 TIMIT 数据集聚类时，该数据集包含 5 种不同的音素，因此类别数应被设定为 5；(2) 基于 3 次 B-样条基底的设定，变量$X_1(t), X_2(t)$的基底个数分别为 30；Growth 数据集和 TIMIT 数据集的基底数量则分别被设置为 20 和 24。基于以上的参数设定，将模拟数据集$X_1(t)$、$X_2(t)$随机生成 100 次；Growth 数据集对其进行 100 次相互独立的重复试验，TIMIT 语音数据集每次抽取了 300 个样本并进行单一样本独立处理，重复实验 100 次。

为了衡量聚类效果，我们将采用 4 个广泛使用的评价指标：ACC、NMI、PUR 和 RI，这些指标值均介于 0 到 1 之间，数值越大表明聚类性能越好。

3.4.3 实验结果

对 BSMFFC 算法和多种聚类算法进行了对比分析，其中包括非函数型聚类算法 K-means、RMNMF 和 RBSMF，以及函数型聚类算法 TA、FCOF 和 FNMF。分别在随机模拟数据 I、Growth 成长数据和 TIMIT 语音数据集上实验，并记录

了四项聚类评价指标的结果平均值和标准差，详见表 3.1。

表 3.1 7 种不同算法在 4 个数据集上聚类结果比较 (100 次抽样结果均值±标准差)

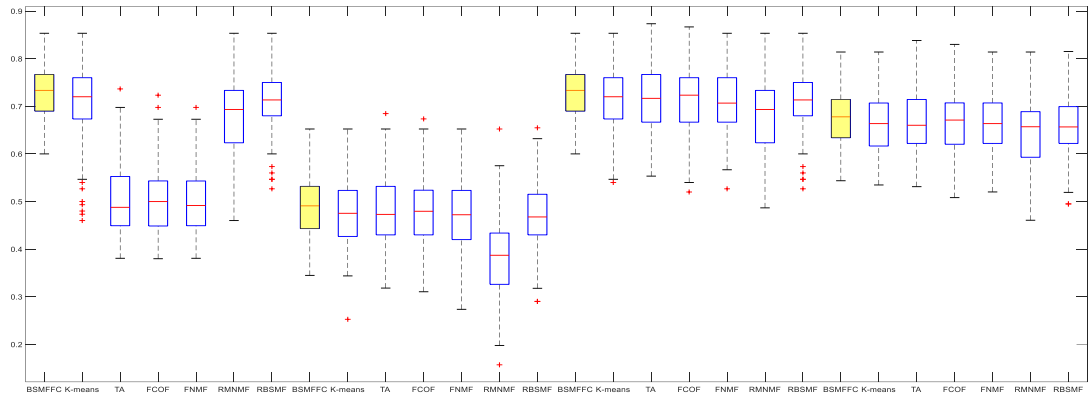
数据集 \ 评价指标 / 方法		ACC(%)	NMI(%)	PUR(%)	RI(%)
$X_1(t)$	K-means	69.75±7.48	43.77±6.53	70.21±2.41	64.55±5.62
	TA	50.48±7.51	48.02±7.47	71.31±7.27	66.71±6.63
	FCOF	50.32±7.26	47.90±7.36	70.30±7.26	66.56±6.64
	FNMF	50.12±6.89	47.22±7.42	70.87±6.98	66.51±6.20
	RMNMF	67.79±8.38	40.96±8.69	67.85±8.23	64.74±5.09
	RBSMF	68.09±8.77	43.84±5.79	68.08±6.42	64.22±5.15
	BSMFFC	**73.03±5.58**	**49.01±6.70**	**73.03±5.58**	**67.65±6.00**
$X_2(t)$	K-means	56.65±4.72	24.01±5.68	57.67±4.05	60.53±2.78
	TA	42.94±3.04	24.42±5.86	58.07±4.16	60.91±2.80
	FCOF	42.81±2.84	23.70±5.59	57.95±3.71	60.82±2.68
	FNMF	44.58±4.03	24.96±6.39	59.27±4.87	62.34±3.65
	RMNMF	56.24±6.94	27.54±7.15	59.09±5.85	63.02±4.78
	RBSMF	58.04±5.09	35.47±6.41	61.45±4.79	62.33±4.05
	BSMFFC	**58.85±5.07**	**38.39±4.45**	**62.04±2.90**	**64.22±3.13**
Growth	K-means	66.14±1.20	7.07±1.14	66.14±1.21	54.74±0.81
	TA	54.68±0.68	6.65±1.07	65.66±1.12	54.42±0.72
	FCOF	57.04±2.40	11.59±4.77	69.14±3.46	57.09±2.81
	FNMF	56.68±2.32	10.06±4.20	68.61±3.34	56.67±2.71
	RMNMF	88.51±2.27	49.77±7.39	88.51±2.27	79.54±3.61
	RBSMF	61.48±4.08	4.40±5.02	61.78±3.51	52.44±2.12
	BSMFFC	**92.37±1.20**	**64.75±2.14**	**92.37±1.20**	**85.77±2.03**
TIMIT	K-means	77.36±7.61	72.56±7.21	79.59±5.91	87.81±3.83
	TA	71.94±6.51	65.18±2.13	80.96±4.38	88.68±2.50
	FCOF	74.72±9.10	70.66±3.01	82.94±4.59	89.96±4.31
	FNMF	74.88±8.12	72.01±2.25	**83.24±2.03**	89.85±3.08
	RMNMF	72.31±4.60	61.76±5.25	74.22±4.32	84.85±2.06
	RBSMF	75.53±6.93	67.47±4.31	77.52±5.14	86.12±3.44
	BSMFFC	**82.33±4.30**	**75.55±3.96**	82.97±3.31	**90.04±2.13**

注：粗体表示比较结果最优

通过表 1 可以发现，在随机模拟数据$X_1(t)$、$X_2(t)$、Growth 成长数据和 TIMIT 语音数据的实验中，BSMFFC 算法在 ACC、NMI、PUR 以及 RI 上，相对于其他几种聚类方法而言聚类性能更为卓越。

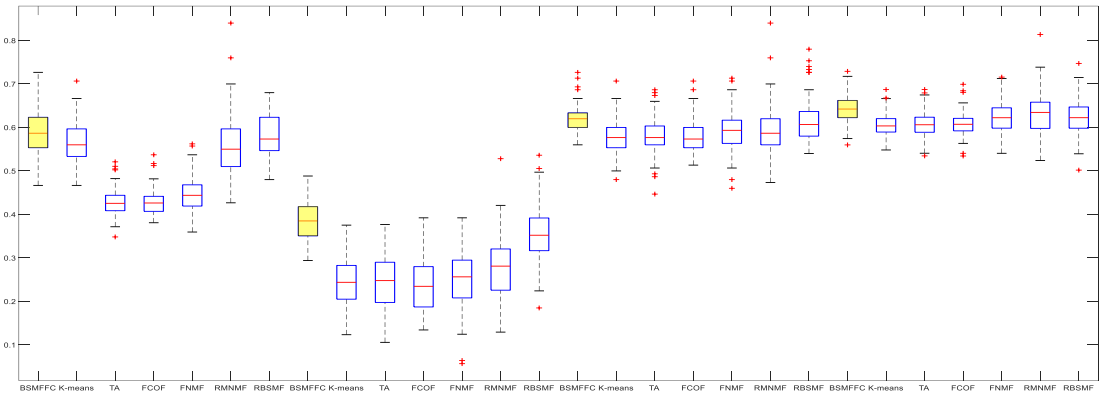
为了评价 BSMFFC 算法的稳定性，在进行了 100 次聚类的基础上，绘制了 4 组数据对应的评价指标箱线图，具体如图 3.4-图 3.7 所示。研究结果表明，BSMFFC 算法的聚类效果在平均水平上明显优于 RMNMF、TA、FCOF 和 FNMF 算法，略微优于 K-means 和 RBSMF 算法，这也从另一方面印证了表 3.1 所得出的结论。同时，BSMFFC 算法在稳定性方面表现出的优异性也值得一提，相比 K-means、TA、FCOF 和 FNMF 算法，其算法稳定性更为出色，在与 RMNMF

和 RBSMF 算法相比时差别不大。因此，在综合表现方面，BSMFFC 算法显然更为突出。



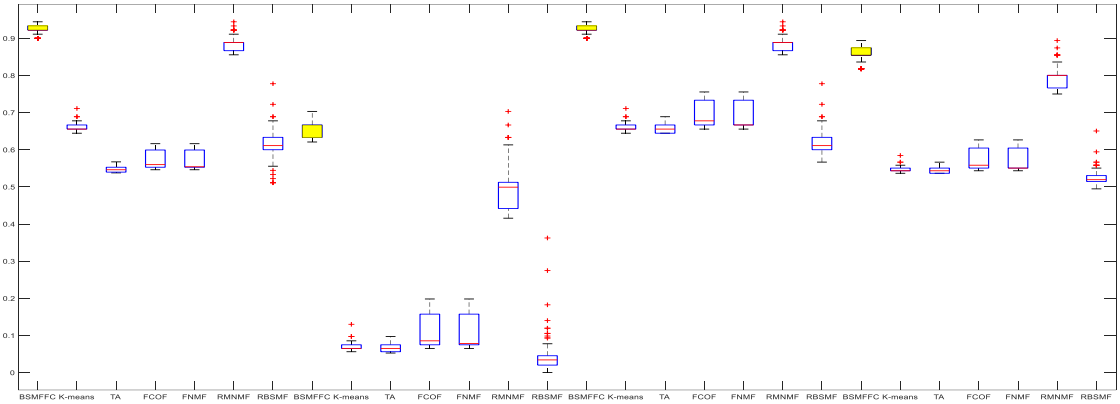
ACC NMI PUR RI

图 3.4 模拟数据X_1的 100 次聚类结果箱线图



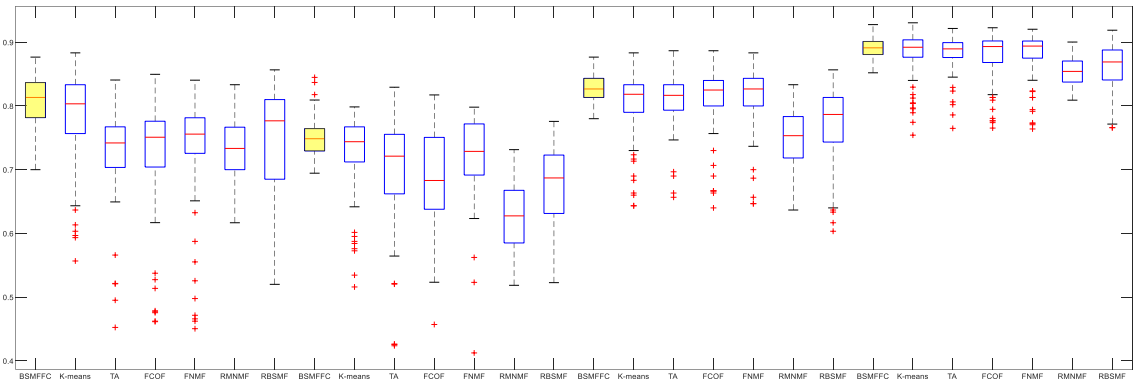
ACC NMI PUR RI

图 3.5 模拟数据X_2的 100 次聚类结果箱线图



ACC NMI PUR RI

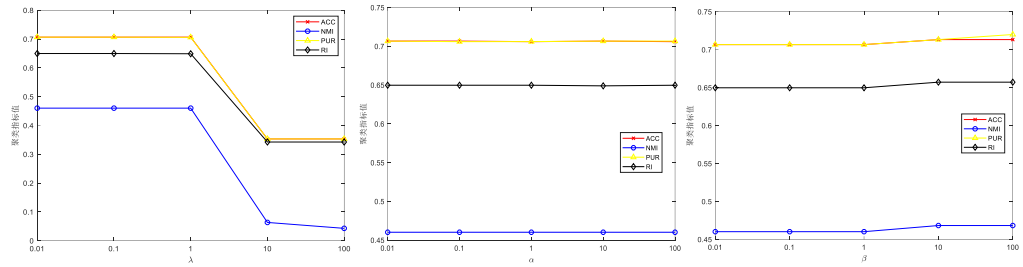
图 3.6 Growth 成长数据的 100 次聚类结果箱线图



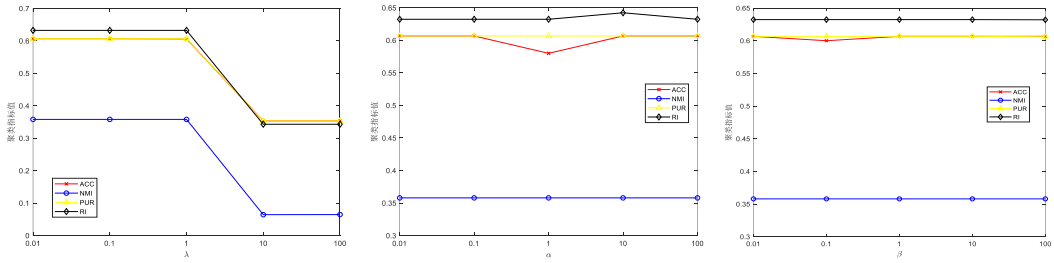
ACC NMI PUR RI

图 3.7 TIMIT 语音数据 100 次聚类结果箱线图

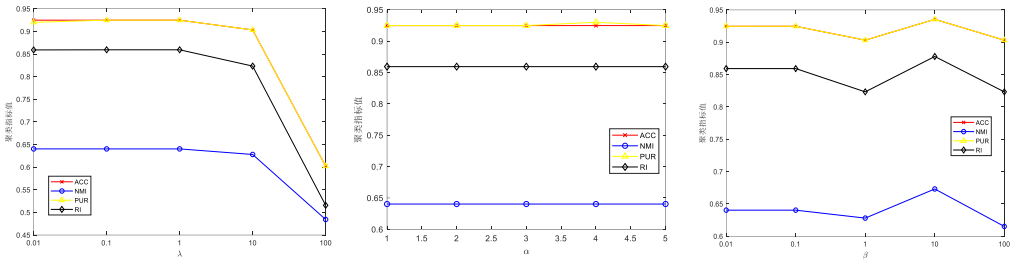
接着，我们通过大量的实验研究所提 BSMFFC 算法对超参数α, λ和β的敏感性。在考虑某一参数时，其余两个参数设定为最优值。例如，当分析α时，λ和β取最优值。图 3.8 展示了参数α, λ和β在{0.01，0.1，1，10，100}范围内变化时 4 个数据集的 ACC、NMI、PUR 和 RI 曲线。可以看出，超参数的取值对聚类效果的影响是有差异的，但 BSMFFC 算法对参数α, λ和β总体上是相对稳健的。特别地，当参数α, λ和β在{0.01，0.1，1}范围内时 BSMFFC 算法在 4 个数据集上的聚类结果均较好。



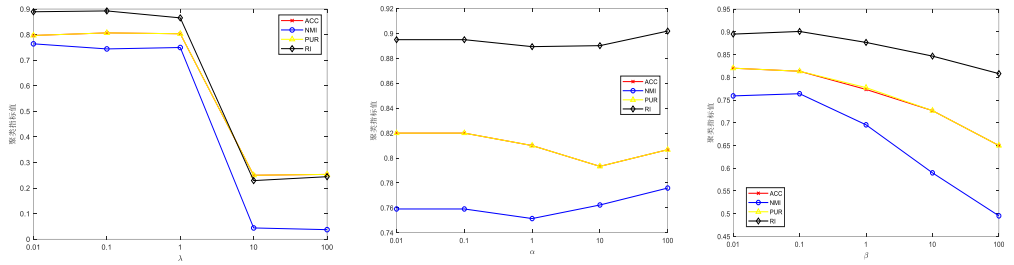
(a) $X_1(t)$



(b) $X_2(t)$



(c) Growth 成长数据



(d) TIMIT 数据

图 3.8　BSMFFC 算法在 4 个数据集上参数敏感性分析

同时，式(3.5)中正则化参数λ平衡流形学习项，KNN 算法中K的取值决定相似图的构造，α控制图学习项的权重，而聚类性能在很大程度上取决于数据的图矩阵。因此，我们在实验中充分考虑了超参数λ、α和K值对聚类性能的影响，分别选取λ和α在$\{0.01, 0.1, 1, 10, 100\}$范围内，$K = \{3, 5, 7, 9, 11\}$，依次在 4 个数据集上进行聚类，结果如图 3.9-图 3.12 所示。

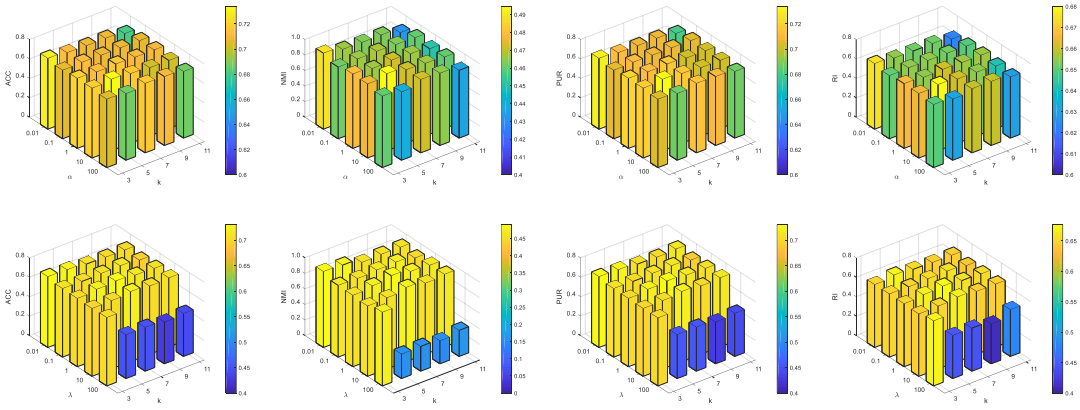
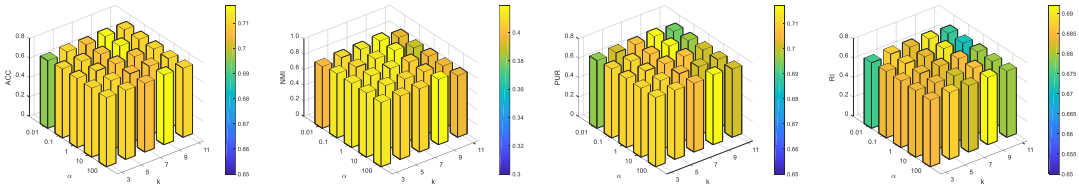


图 3.9　　模拟数据$X_1(t)$的聚类效果



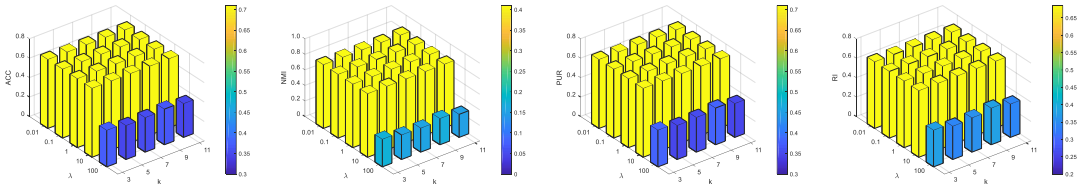


图 3.10 模拟数据$X_2(t)$的聚类效果

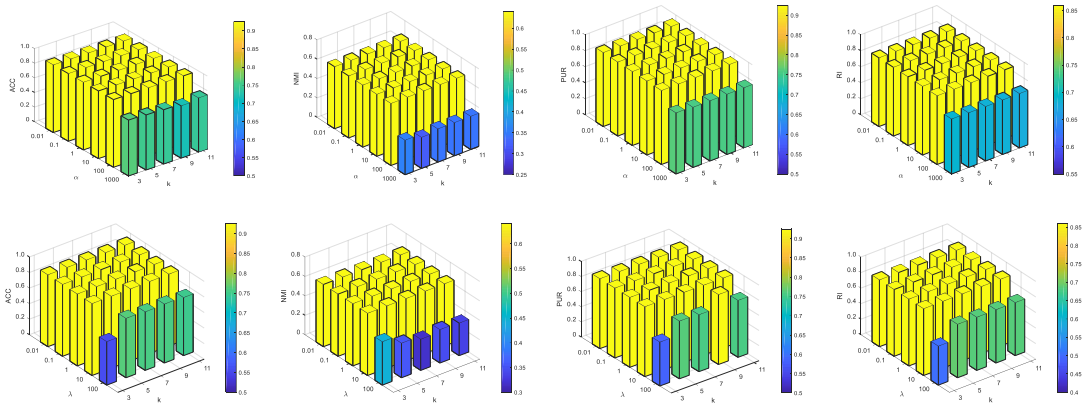


图 3.11 Growth 成长数据的聚类效果

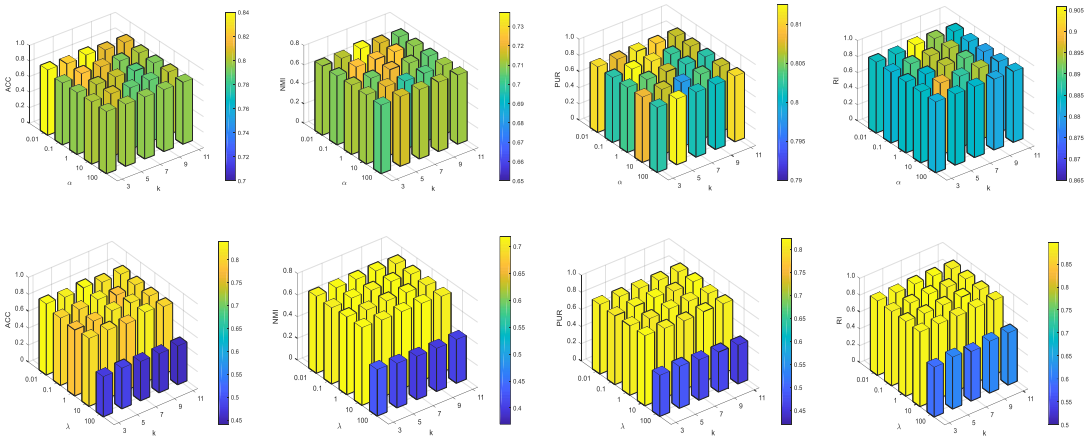


图 3.12 TIMIT 语音数据的聚类效果

图 3.9-图 3.12 中，α与k、λ与K对聚类性能的影响分别呈上下两层依次显示。可以看出，聚类性能与超参数λ、α和K的取值是密切相关的，不同的取值组合下，聚类的结果也不尽相同。当λ固定时，K值变大，聚类指标值相应减小。此外当K固定时，可以直观看出，λ和α值取 100 时，聚类性能下降明显。因此，结合图 3.8 可得知，当λ和α取$\{0.01, 0.1, 1, 10\}$时，聚类性能较好，较为稳健。

最后，进一步探究了该算法的收敛速率。根据图 3.13 所示，在 4 个给定数据集中，目标函数值随着迭代次数的增加呈现出明显的单调递减趋势。而且在 4 个数据集上，目标函数值大约在 20 次迭代左右就可达到收敛状态。总体而言，我们的实验结果表明 BSMFFC 优化算法的收敛速率非常快，具有较高的求解效率。

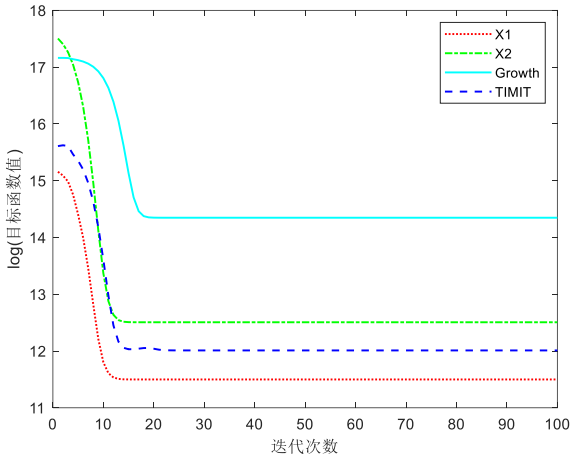


图 3.13 BSMFFC 算法的收敛性

3.5 实例应用—以北京市污染物小时浓度监测站点聚类为例

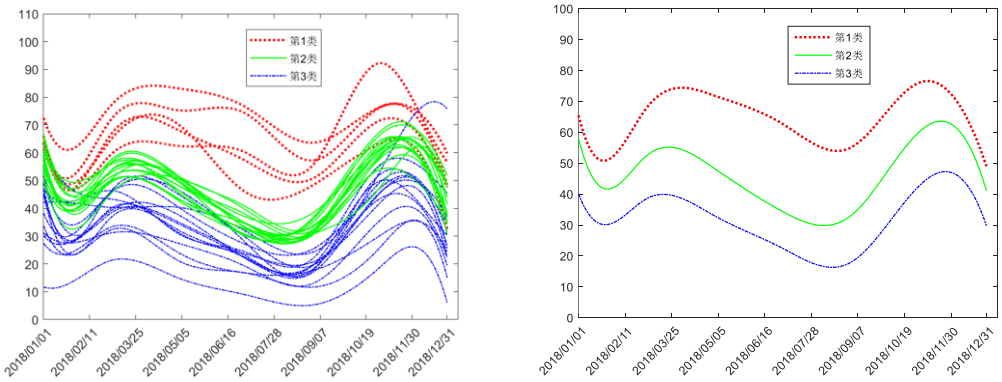
为了进一步验证 BSMFFC 算法的实际应用效果，本文以 2018 年 1 月 1 日至 2018 年 12 月 31 日北京市监测站点所记录 NO2 污染物小时浓度数据为例，采用所提算法对其进行聚类。北京市空气质量监测站点记录的监测站点相关基本信息包括站名、类型、经纬度坐标等参数，详见表 3.2。

表 3.2 北京市空气质量监测站点基本信息

站点类别	编号	站点名称(经纬度坐标)	编号	站点名称(经纬度坐标)	编号	站点名称(经纬度坐标)
城市环境评价站点	1	东四(116.42, 39.93)	9	北部新区(116.17, 40.09)	17	云岗(116.15, 39.82)
	2	良乡(116.14, 39.74)	10	丰台花园(116.28, 39.86)	18	昌平镇(116.23, 40.22)
	3	西城官园(116.34, 39.93)	11	顺义新城(116.66, 40.13)	19	双峪(116.11, 39.94)
	4	万寿西宫(116.35, 39.88)	12	古城(116.18, 39.91)	20	海淀万柳(116.29, 39.99)
	5	夏都(115.97, 40.45)	13	天坛(116.41, 39.89)	21	怀柔镇(116.63, 40.33)
	6	农展馆(116.46, 39.94)	14	黄村(116.40, 39.72)	22	密云镇(116.83, 40.37)
	7	平谷镇(117.10, 40.14)	15	亦庄(116.51, 39.80)	23	奥体中心(116.40, 39.98)
	8	香山(116.21, 40.00)	16	通州北苑(116.660, 39.89)		
区域背景传输站点	24	京西北(115.990, 40.370)	26	京东(117.120, 40.100)	28	京南(116.300, 39.520)
	25	京东北(116.910, 40.500)	27	京东南(116.780, 39.710)	29	京西南(116.000, 39.580)
交通污染控制站点	30	前门(116.400, 39.900)	32	西直门(116.350, 39.950)	34	东四环(116.480, 39.940)
	31	永定门(116.390, 39.880)	33	南三环(116.370, 39.860)		
城市清洁对照站点	35	定陵(116.220, 40.29)				

注：站点信息为 2016 年北京市空气质量监测点分类调整后的命名方式。

对 34 个监测站点的数据进行分析，排除了“城市清洁对照点”。在对数据进行了缺失插补和异常处理(黄恒君，2014)[7]的基础上，采用 BSMFFC 算法对 NO_2 污染物小时浓度数据进行聚类分析，相关的参数设定如下：(1) 采用等距节点 3 次 B-样条基底拟合曲线，并取基底个数为 9；(2) 由于空气质量监测站点分为 3 类，因此聚类数设置为$r=3$。BSMFFC 算法的聚类结果展示在图 3.14 中。



(a)曲线聚类结果　　(b)类均值曲线

图 3.14　2018 年北京市 NO_2 浓度聚类结果

图 3.14 中，采用不同的线型和颜色展示了聚类信息，其中红色点线表示第 1 类，绿色实线表示第 2 类，蓝色虚线表示第 3 类。其中，图 14(a)展示了北京市 34 个空气检测站点的 NO2 污染物小时浓度水平的曲线聚类结果，图 14(b)则呈现了 NO2 污染物小时浓度的类均值曲线。总体上看，分类具有明显的类别特征，NO2 污染物浓度从高到低依次呈现出趋势。第 2 类和第 3 类则呈现出明显的年周期波动特征，在 3 月末和 11 月末 NO2 污染物浓度达到峰值，第 1 类中站点的 NO2 污染物浓度基本一直高位运行。

为了进一步说明聚类结果的空间布局，我们将图 3.14 中北京市 34 个空气质量监测站点的聚类结果可视化呈现在北京市地图上，如图 3.15 所示：

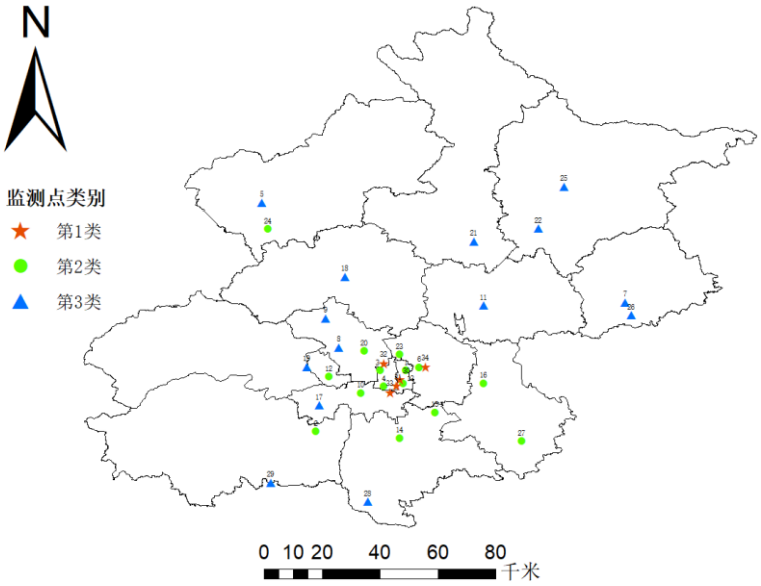


图 3.15 北京市空气质量监测站点的聚类结果空间分布图

通过与表 2 中北京市现有监测站点类型的对比，可以发现，图 3.15 中的聚类结果显示出第 1 类涵盖了所有交通污染控制站点，且主要分布于北京市中心城区；第 2 类涵盖了大部分城市环境评价站点；第 3 类则涵盖除 24 号京西北站点和 27 号京东南站点之外的绝大部分区域背景传输站点。此外，根据图 3.15 中呈现的聚类空间分布图所显示的结果，在空气质量监测站点的空间布局识别方面，所提出的 BSMFFC 算法显示出了较高的实用性和适用性。值得注意的是，该图所展现的聚类特征呈现出一定程度的环状模式，其中最外环、中环和内环分别对应第 3 类、第 2 类和第 1 类。而基于具体的实例应用结果来看，所提供的数据支持可以有效唯空气质量监测点的设定和布局调整提供帮助。综上所述，BSMFFC 算法的实用性和适用性十分广泛，具有一定的应用前景。

3.6 本章小结

本章提出了一种全新的函数型聚类算法—基于双随机图正则化矩阵分解（BSMFFC）的方法。相较于传统算法，BSMFFC算法增加了自适应图学习策略，通过动态学习相似图且利用双随机矩阵，弥补了学习图的不足之处，提高了聚类的适用性。此外，BSMFFC算法也实现了并行计算，结合了图的更新和矩阵分解过程。此外，本文还采用了增广拉格朗日乘子法（ALM）来推导更新规则，分析了其收敛性、计算复杂性和参数敏感性等重要问题。实验研究显示，BSMFFC算法对随机模拟数据I、Growth数据和TIMIT语音数据的处理显示出非凡的有效

性与显著优势。此外，在北京城市空气质量监测站点数据的聚类分析中，也充分证实了该算法在真实场景应用的可行性与合理性，展现出巨大的应用价值和发展潜能。

4 基于鲁棒图正则化矩阵分解的函数型聚类算法

本章提出了用于函数型数据聚类的鲁棒流形非负矩阵分解算法（Functional Robust Manifold Nonnegative Matrix Factorization，FRMNMF）。具体地说，采用基于$l_{2,1}$范数的损失函数来抑制噪声或异常值的影响，利用流形的局部不变性质，充分保持了相邻数据的局部几何结构。本研究采用交替迭代法对求解过程进行迭代更新，并从收敛性与计算复杂度两个维度对所提方法进行了深入剖析。通过实证分析，无论是在合成数据集还是真实数据集上，该算法均展现出了优异的性能表现，与几种先进的方法相比，FRMNMF 在函数型聚类任务中具有可行性和鲁棒性。

4.1 问题概述

Kong 等(2011)[21]基于非负矩阵分解，利用$l_{2,1}$范数定义损失函数，提出鲁棒非负矩阵分解，即 RNMF。鉴于 RNMF 方法存在的优势，本章将 RNMF 算法引入函数型数据聚类分析中，并与图正则化项相结合，讨论鲁棒的图正则化函数型数据聚类方法，旨在提高算法的可用范围，提高算法鲁棒性和聚类性能。

后续的相关内容安排如下：4.2 构建所提出 FRMNMF 模型的框架；4.3 推导给出有效的优化迭代算法求解所提出的目标函数；4.4 为实验，利用随机模拟数据，Growth 成长数据集、CanadianWeather 数据集和 Fatspectrum 数据集来验证所提出的 FRMNMF 算法的优越的聚类性能，并且对收敛速度和参数灵敏度都进行了分析；3.5 为实例应用，利用 2000-2022 年的城镇居民人均可支配收入数据进行聚类应用；3.6 为本章小结。

4.2 目标函数

考虑使用$l_{2,1}$范数来增强 FNMF 聚类的鲁棒性，并引入局部流形正则化，提出了将流形学习和 RNMF 相结合的基于鲁棒非负矩阵分解的函数型聚类算法(FRMNMF)。具体地说，为深入探究高维空间数据$\boldsymbol{Y}$的几何特征，本章融入了 RNMF 方法中的关键式—拉普拉斯算子，此举是为了捕捉数据流形的固有局部性结构，并保持该性质在降维过程中的连续性。据此基础，FRMNMF 模型的目

标函数得以构建：

$$\min_{\boldsymbol{U}\geq 0,\,\boldsymbol{V}\geq 0}\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T\|_{2,1}+\frac{\lambda}{2}\mathrm{tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V}) \tag{4.1}$$

其中，$\boldsymbol{Y}\in\mathbb{R}^{m\times n}$，$\boldsymbol{\Phi}\in\mathbb{R}^{m\times p},\boldsymbol{U}\in\mathbb{R}^{p\times r}$，$\boldsymbol{V}^T=(\boldsymbol{v_1},\boldsymbol{v_2},\ldots,\boldsymbol{v_n})\in\mathbb{R}^{r\times n}$，$\boldsymbol{v_j}\in\mathbb{R}^r$。

正则化项$\mathcal{R}=\frac{\lambda}{2}\mathrm{tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V})$可以在低维表示$\boldsymbol{v}_i$和$\boldsymbol{v}_j$中保留$\boldsymbol{y}_i$和$\boldsymbol{y}_j$具有相似特征的局部信息，从而大大提高其表示能力和聚类性能。

4.3 求解算法

4.3.1 优化求解

对于目标函数式(4.1)的优化问题，根据 Lee 和 Seung[71]推导迭代乘法算法的公式。令目标函数式(4.1)拉格朗日函数为

$$\begin{aligned}\boldsymbol{J}&=\frac{1}{2}\mathrm{tr}(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{G}(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)^T+\frac{\lambda}{2}\mathrm{tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V})\\&\quad-\mathrm{tr}(\boldsymbol{\Lambda_1}\boldsymbol{U}^T)-\mathrm{tr}(\boldsymbol{\Lambda_2}\boldsymbol{V}^T)\\&:=f(\boldsymbol{U},\boldsymbol{V})+g(\boldsymbol{V})-\mathrm{tr}(\boldsymbol{\Lambda_1}\boldsymbol{U}^T)-\mathrm{tr}(\boldsymbol{\Lambda_2}\boldsymbol{V}^T)\end{aligned} \tag{4.2}$$

其中，$\boldsymbol{G}=\mathrm{diag}(G_{11},G_{22},\ldots,G_{nn})$，且

$$G_{jj}=1/\left(\sum_{i=1}^{m}(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}})_{ij}^2\right)^{\frac{1}{2}}=1/\|\boldsymbol{y_j}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{v_j}\|_2 \tag{4.3}$$

这里，为了避免对角矩阵$\boldsymbol{G}$的对角线元素出现分母为0的情形，我们通常做如下处理:

$$G_{jj}=1/\left[\left(\sum_{i=1}^{m}(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}})_{ij}^2\right)^{\frac{1}{2}}+\epsilon\right]=1/\left[\|\boldsymbol{y_j}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{v_j}\|_2+\epsilon\right],\quad\epsilon>0 \tag{4.4}$$

同时

$$\begin{aligned}f(\boldsymbol{U},\boldsymbol{V})&=\frac{1}{2}\mathrm{tr}(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}})\boldsymbol{G}(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}})^T\\&=\frac{1}{2}\mathrm{tr}(\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}^{\boldsymbol{T}}-\boldsymbol{2}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{G}\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U}+\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{G}\boldsymbol{V}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{\Phi}^{\boldsymbol{T}})\end{aligned} \tag{4.5}$$

$$g(\boldsymbol{V})=\frac{\lambda}{2}\mathrm{tr}(\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{L}\boldsymbol{V}) \tag{4.6}$$

函数$f(\boldsymbol{U},\boldsymbol{V})$分别关于$\boldsymbol{U}$、$\boldsymbol{V}$求偏导数，$g(\boldsymbol{V})$关于$\boldsymbol{V}$求偏导数得

$$\frac{\partial f}{\partial\boldsymbol{U}}=-\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{V}+\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{G}\boldsymbol{V}$$

$$\frac{\partial f}{\partial \boldsymbol{V}} = -\boldsymbol{G}\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U} + \boldsymbol{G}\boldsymbol{V}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U}$$

$$\frac{\partial g}{\partial \boldsymbol{V}} = \boldsymbol{\lambda}\boldsymbol{L}\boldsymbol{V}$$

式(4.2)关于$\boldsymbol{U}$，$\boldsymbol{V}$求偏导，令$\frac{\partial J}{\partial \boldsymbol{U}} = 0$，$\frac{\partial J}{\partial \boldsymbol{V}} = 0$，有

$$\frac{\partial J}{\partial \boldsymbol{U}} = -\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{V} + \boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{G}\boldsymbol{V} - \boldsymbol{\Lambda_1} = 0$$

$$\frac{\partial J}{\partial \boldsymbol{V}} = -\boldsymbol{G}\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U} + \boldsymbol{G}\boldsymbol{V}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U} + \boldsymbol{\lambda}\boldsymbol{L}\boldsymbol{V} - \boldsymbol{\Lambda_2} = 0$$

从而

$$\boldsymbol{\Lambda_1} = \boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{G}\boldsymbol{V} - \boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{V}$$

$$\boldsymbol{\Lambda_2} = \boldsymbol{G}\boldsymbol{V}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U} - \boldsymbol{G}\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U} + \boldsymbol{\lambda}\boldsymbol{L}\boldsymbol{V}$$

结合 KKT 条件$\boldsymbol{\Lambda_1} \odot \boldsymbol{U} = 0$，$\boldsymbol{\Lambda_2} \odot \boldsymbol{V} = 0$，得$\boldsymbol{U}$和$\boldsymbol{V}$的更新规则为

$$\boldsymbol{U}_{ij} \leftarrow \boldsymbol{U}_{ij}\frac{(\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{V}))_{ij}}{(\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{G}\boldsymbol{V})_{ij}} \tag{4.7}$$

$$\boldsymbol{V}_{ij} \leftarrow \boldsymbol{V}_{ij}\frac{(\boldsymbol{G}\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U} + \boldsymbol{\lambda}\boldsymbol{W}\boldsymbol{V})_{ij}}{(\boldsymbol{G}\boldsymbol{V}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{U} + \boldsymbol{\lambda}\boldsymbol{D}\boldsymbol{V})_{ij}} \tag{4.8}$$

综合分析得出，式(4.4)、式(4.7)与式(4.8)依次使用，对$\boldsymbol{G}$、$\boldsymbol{U}$和$\boldsymbol{V}$进行连续交替迭代更新，从而可对优化模型式(4.1)进行求解。

4.3.2 算法流程

综合迭代求解更新公式，我们提出的基于鲁棒图正则化矩阵分解的函数型聚类算法(FRMNMF)的相关具体执行步骤详见算法 4.1。

算法 4.1 基于鲁棒图正则化矩阵分解的函数型聚类算法(FRMNMF)

输入：数据矩阵$\boldsymbol{Y}$、基底矩阵$\boldsymbol{\Phi}$、惩罚参数λ和类别数r
过程：
1：构造拉普拉斯矩阵$\boldsymbol{L}$
2：初始化$\boldsymbol{U}^0$、$\boldsymbol{V}^0$
3：for $\mathrm{t} = 1, 2, \cdots, \mathrm{maxiter}$
4：　　固定$\boldsymbol{U}^{t-1}$、$\boldsymbol{V}^{t-1}$，根据式(4.4)更新$\boldsymbol{G}$
5：　　固定$\boldsymbol{V}^{t-1}$，根据式(4.7)更新$\boldsymbol{U}^{t-1}$
6：　　固定$\boldsymbol{U}^{t}$，根据式(4.8)更新$\boldsymbol{V}^{t-1}$
7：　　if 式(4.1)收敛
8：　　　　break
9：　　end if
10：end for
输出：$\boldsymbol{U}$、$\boldsymbol{V}$和$\boldsymbol{G}$，簇划分$\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$

4.3.3 收敛性证明

下面我们采用标准辅助函数方法[71]，证明更新规则式(4.7)和式(4.8)的收敛性。

定理 1 固定$\boldsymbol{V}$，依更新规则式(4.7)更新$\boldsymbol{U}$，目标函数式(4.1)是单调递减的；固定$\boldsymbol{U}$，依更新规则式(4.8)更新$\boldsymbol{V}$，目标函数式(4.1)是单调递减的。

命题 1 矩阵不等式（Ding *et.al*[72]）

$$\operatorname{tr}(\boldsymbol{S^T A S B}) \leq \sum_{i=1}^{n}\sum_{l=1}^{r}(\boldsymbol{AS'B})_{il}\frac{\boldsymbol{S}_{il}^2}{\boldsymbol{S}'_{il}}$$

其中$\boldsymbol{A}\in\mathbb{R}_+^{n\times n}$，$\boldsymbol{B}\in\mathbb{R}_+^{r\times r}$为对称阵，$\boldsymbol{S}\in\mathbb{R}_+^{n\times r}$，$\boldsymbol{S'}\in\mathbb{R}_+^{n\times r}$。当且仅当$\boldsymbol{S}=\boldsymbol{S'}$时取等号。

定理 1 证明 不妨记

$$\mathcal{F}(\boldsymbol{V}) := \left\|\boldsymbol{Y}-\boldsymbol{\Phi U V^T}\right\|_{2,1} + \frac{\lambda}{2}\operatorname{tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V}) \tag{4.9}$$

$$F(\boldsymbol{V}) := \frac{1}{2}\operatorname{tr}(\boldsymbol{Y}-\boldsymbol{\Phi U V^T})\boldsymbol{G}(\boldsymbol{Y}-\boldsymbol{\Phi U V^T})^T + \frac{\lambda}{2}\operatorname{tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V}) \tag{4.10}$$

利用命题 1，$F(\boldsymbol{V})$可展为

$$\begin{aligned} F(\boldsymbol{V}) &= \frac{1}{2}\operatorname{tr}\left(\boldsymbol{YGY^T}-\boldsymbol{2V^TGY^T\Phi U}\right)+\frac{\lambda}{2}\operatorname{tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V}) \\ &\quad+\frac{1}{2}\operatorname{tr}\left(\boldsymbol{GV}(\boldsymbol{\Phi U})^{\boldsymbol{T}}(\boldsymbol{\Phi U})\boldsymbol{V^T}\right) \\ &\leq \frac{1}{2}\operatorname{tr}\left(\boldsymbol{YGY^T}-\boldsymbol{2V^TGY^T\Phi U}\right)+\frac{\lambda}{2}\operatorname{tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V}) \\ &\quad+\frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{r}\frac{\left(\boldsymbol{GV'}(\boldsymbol{\Phi U})^{\boldsymbol{T}}(\boldsymbol{\Phi U})\right)_{il}\boldsymbol{V}_{il}^2}{\boldsymbol{V}'_{il}} \\ &:= H(\boldsymbol{V},\boldsymbol{V'}) \end{aligned} \tag{4.11}$$

并且$\boldsymbol{F(V)}=H(\boldsymbol{V},\boldsymbol{V})$，利用辅助函数方法，则$H(\boldsymbol{V},\boldsymbol{V'})$是$F(\boldsymbol{V})$的辅助函数。

如果$\boldsymbol{V}^{t+1}$，使得

$$\boldsymbol{V}^{t+1} = \arg\min_{\boldsymbol{V}} H(\boldsymbol{V},\boldsymbol{V^t}) \tag{4.12}$$

有

$$F(\boldsymbol{V}^{t+1}) = H(\boldsymbol{V}^{t+1},\boldsymbol{V}^{t+1}) \leq H(\boldsymbol{V}^{t+1},\boldsymbol{V}^t) \leq H(\boldsymbol{V}^t,\boldsymbol{V}^t) = F(\boldsymbol{V}^t)$$

则

$$F(\boldsymbol{V}^{t+1}) \leq F(\boldsymbol{V}^t) \leq \cdots \leq F(\boldsymbol{V}^0)$$

即$F(\boldsymbol{V})$关于$\boldsymbol{V}$单调递减。

事实上，$H(\boldsymbol{V},\boldsymbol{V'})$关于$\boldsymbol{V}$的梯度为

$$\frac{\partial H(\boldsymbol{V},\boldsymbol{V'})}{\partial \boldsymbol{V}_{il}} = -(\boldsymbol{GY^T\Phi U})_{il} + \frac{\left(\boldsymbol{GV'(\Phi U)^T(\Phi U)}\right)_{il}\boldsymbol{V}_{il}}{\boldsymbol{V'_{il}}} + (\boldsymbol{LV})_{il} \tag{4.13}$$

包含二阶导数

$$\frac{\partial^2 H(\boldsymbol{V},\boldsymbol{V'})}{\partial \boldsymbol{V}_{il}\boldsymbol{V}_{js}} = \left(\frac{\left(\boldsymbol{GV'(\Phi U)^T(\Phi U)}\right)_{il}}{\boldsymbol{V'_{il}}} + \boldsymbol{L}_{il}\right)\delta_{ij}\delta_{ls} \tag{4.14}$$

的 Hessian 矩阵是半正定的。因此，$H(\boldsymbol{V},\boldsymbol{V'})$是关于$\boldsymbol{V}$的凸函数。令

$$\frac{\partial H(\boldsymbol{V},\boldsymbol{V'})}{\partial \boldsymbol{V}_{il}} = 0$$

注意到，$\boldsymbol{V}\to\boldsymbol{V}^{t+1}$，$\boldsymbol{V'}\to\boldsymbol{V}^{t}$，可得$H(\boldsymbol{V},\boldsymbol{V'})$的全局最小值

$$\boldsymbol{V}^*_{ij} = \boldsymbol{V}'_{ij}\frac{(\boldsymbol{GY^T\Phi U})_{ij}}{(\boldsymbol{GV'(\Phi U)^T(\Phi U)+LV})_{ij}} \tag{4.15}$$

注意到

$$\begin{aligned} F(\boldsymbol{V}^t) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{Y}-\boldsymbol{\Phi U}(\boldsymbol{V^t})^{\boldsymbol{T}})\boldsymbol{G}(\boldsymbol{Y}-\boldsymbol{\Phi U}(\boldsymbol{V^t})^{\boldsymbol{T}})^T + \frac{\lambda}{2}\mathrm{tr}((\boldsymbol{V^t})^T\boldsymbol{LV^t}) \\ &= \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{m}(\boldsymbol{Y}-\boldsymbol{\Phi U}(\boldsymbol{V^t})^{\boldsymbol{T}})_{ij}^2 G_{jj} + \frac{\lambda}{2}\mathrm{tr}((\boldsymbol{V^t})^T\boldsymbol{LV^t}) \\ &= \frac{1}{2}\sum_{j=1}^{n}\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^t})\right\|_2^2 G_{jj} + \frac{\lambda}{2}\mathrm{tr}((\boldsymbol{V^t})^T\boldsymbol{LV^t}) \\ &= \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{1}{G_{jj}} + \frac{\lambda}{2}\mathrm{tr}((\boldsymbol{V^t})^T\boldsymbol{LV^t}) \end{aligned} \tag{4.16}$$

其中$G_{jj}=1/\|\boldsymbol{y_j}-\boldsymbol{\Phi U v_j^t}\|_2$。

类似地

$$\begin{aligned} F(\boldsymbol{V}^{t+1}) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{Y}-\boldsymbol{\Phi U}(\boldsymbol{V^{t+1}})^T)\boldsymbol{G}(\boldsymbol{Y}-\boldsymbol{\Phi U}(\boldsymbol{V^{t+1}})^T)^T + \frac{\lambda}{2}\mathrm{tr}((\boldsymbol{V^{t+1}})^T\boldsymbol{LV^{t+1}}) \\ &= \frac{1}{2}\sum_{j=1}^{n}\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2^2 G_{jj} + \frac{\lambda}{2}\mathrm{tr}((\boldsymbol{V^{t+1}})^T\boldsymbol{LV^{t+1}}) \end{aligned} \tag{4.17}$$

则

$$F(\boldsymbol{V}^{t+1}) - F(\boldsymbol{V}^t) = \frac{1}{2}\sum_{j=1}^{n}\left(\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2^2 G_{jj} - \frac{1}{G_{jj}}\right) \tag{4.18}$$

又

$$\begin{aligned} \mathcal{F}(\boldsymbol{V}^{t+1}) - \mathcal{F}(\boldsymbol{V}^t) :=& \left\|\boldsymbol{Y}-\boldsymbol{\Phi U}(\boldsymbol{V^{t+1}})^{\boldsymbol{T}}\right\|_{2,1} - \left\|\boldsymbol{Y}-\boldsymbol{\Phi U}(\boldsymbol{V^t})^{\boldsymbol{T}}\right\|_{2,1} \\ =& \sum_{j=1}^{n}\left(\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2 - \left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^t})\right\|_2\right) \\ =& \sum_{j=1}^{n}\left(\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2 - \frac{1}{G_{jj}}\right) \end{aligned} \tag{4.19}$$

式(4.19)减去式(4.18)得

$$\begin{aligned}
&\left(\mathcal{F}(\boldsymbol{V}^{t+1})-\mathcal{F}(\boldsymbol{V}^{t})\right)-\left(F(\boldsymbol{V}^{t+1})-F(\boldsymbol{V}^{t})\right)\\
&=\sum_{j=1}^{n}\left(\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2-\frac{1}{2}\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2^2-\frac{1}{G_{jj}}\right)\\
&=\sum_{j=1}^{n}\frac{G_{jj}}{2}\left(\frac{2\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2}{G_{jj}}-\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2^2-\frac{1}{G_{jj}^2}\right)\\
&=\sum_{j=1}^{n}-\frac{G_{jj}}{2}\left(\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2^2-2\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2\frac{1}{G_{jj}}+\frac{1}{G_{jj}^2}\right)\\
&=\sum_{j=1}^{n}-\frac{G_{jj}}{2}\left(\left\|\boldsymbol{y_j}-\boldsymbol{\Phi U}(\boldsymbol{v_j^{t+1}})\right\|_2-\frac{1}{G_{jj}}\right)^2\\
&\leq 0
\end{aligned}\tag{4.20}$$

从而

$$\mathcal{F}(\boldsymbol{V}^{t+1})-\mathcal{F}(\boldsymbol{V}^{t})\leq F(\boldsymbol{V}^{t+1})-F(\boldsymbol{V}^{t})$$

又$F(\boldsymbol{V}^{t+1})-F(\boldsymbol{V}^{t})\to 0$，所以

$$\left\|\boldsymbol{Y}-\boldsymbol{\Phi U}(\boldsymbol{V^{t+1}})^{\boldsymbol{T}}\right\|_{2,1}-\left\|\boldsymbol{Y}-\boldsymbol{\Phi U}(\boldsymbol{V^{t}})^{\boldsymbol{T}}\right\|_{2,1}\leq 0$$

同理可证

$$\left\|\boldsymbol{Y}-\boldsymbol{\Phi U^{t+1}V^{T}}\right\|_{2,1}-\left\|\boldsymbol{Y}-\boldsymbol{\Phi U^{t}V^{T}}\right\|_{2,1}\leq 0$$

因此，$\mathcal{F}(\boldsymbol{V}^{t+1})-\mathcal{F}(\boldsymbol{V}^{t})\leq 0$，即$\mathcal{F}(\boldsymbol{V})$关于$\boldsymbol{V}$单调递减。当更新次数$t$不断增大时，$\mathcal{F}(\boldsymbol{V})$至少会收敛于一个局部极小值。

4.3.4 计算复杂度分析

进一步探讨 FRMNMF 算法的计算复杂度问题。样本量为n，特征数量为m，基底的个数为p，聚类的个数为r。由此可知，在算法 4.1 中，主要的计算复杂度在进行$\boldsymbol{G}$、$\boldsymbol{U}$和$\boldsymbol{V}$的更新迭代过程中，具体如下：

(1) 根据式(4.4)，更新$\boldsymbol{G}$的时间复杂度为$O(mpr+mnr+mn+m+n)$。

(2) 在$\boldsymbol{U}$的更新过程中，时间复杂度为$O(mnp+pr+p^2r+p^2m+pn^2+npr)$。

(3) 在$\boldsymbol{V}$的更新过程中，时间复杂度为$O(mnp+n^2r+npr+mn^2+nr)$。

每次迭代的计算复杂度为$O(mnr)$，其中迭代次数记作t。因此，当进行t次迭代更新后，该算法的计算复杂度为$O(tmnr)$。

4.4 模拟实验

为了验证本文所提出的 FRMNMF 算法的有效性和优越性，我们进行了一系列模拟实验。在保持参数设置基本一致的情况下，我们将本文算法与其他六种聚类算法进行了比较：Lee 和 Seung(2001)的 NMF 算法[15]、Kong 等(2011)的 RNMF 算法[21]、黄恒君等(2019)的 FCOF 算法[13]、高海燕等(2020)的 FNMF 算法[14]、Jacques 和 Preda(2013)的 Funclust 算法[5]和 Schmutz 等(2020)的 FunHDDC 方法[57]进行比较。在合成数据(随机模拟数据)和真实数据(Growth 数据、CanadianWeather 数据和 FatSpectrum 数据)上分别进行实验，评估了所提出的 FRMNMF 算法的聚类性能。

本实验中的所有代码均在 R4.3.1 软件中实现，实验的计算机环境为：12th Gen Intel(R) Core(TM) i5-12500H 2.50 GHz，内存 16GB，Windows11 64 位操作系统。

4.4.1 实验数据

(1) 随机模拟数据 II

数据按照模型

$$\boldsymbol{Y}=\boldsymbol{\Phi}(\boldsymbol{MU}+\boldsymbol{E}_1)+\boldsymbol{E}_2+a\boldsymbol{1}$$

进行模拟。其中，曲线离散采样点$n=500$、样本曲线数量$N=100$。

具体而言，数据生成和分析步骤如下：(1) 生成类中心矩阵$\boldsymbol{M}$和类标签矩阵$\boldsymbol{U}$。其中，$\boldsymbol{\mu}_1$和$\boldsymbol{\mu}_2$的前 10 个元素分别按照$N(0.5,1)$和$N(-0.5,1)$两种不同方式生成，以凸显类别间的差异；$\boldsymbol{\mu}_1$和$\boldsymbol{\mu}_2$的后 40 个元素则通过相同的方式生成，即$N(0,0.01)$，不含任何类别信息，作为随机误差项。类标签矩阵$\boldsymbol{U}$的元素按照等概率方式生成。(2) 生成独立分布的随机误差项向量$\boldsymbol{E}_1$和$\boldsymbol{E}_2$，其中向量按照相应的分布形成，分别为$N(0,1)$和$N(0,0.1)$。(3) 生成全 1 矩阵$\boldsymbol{1}$，常数a取 5，确保生成的随机模拟数据为正数矩阵。(4) 根据如下公式生成数据矩阵$\boldsymbol{Y}$。数据生成试验重复 1000 次，其中一次数据生成的示例如图 1 所示：

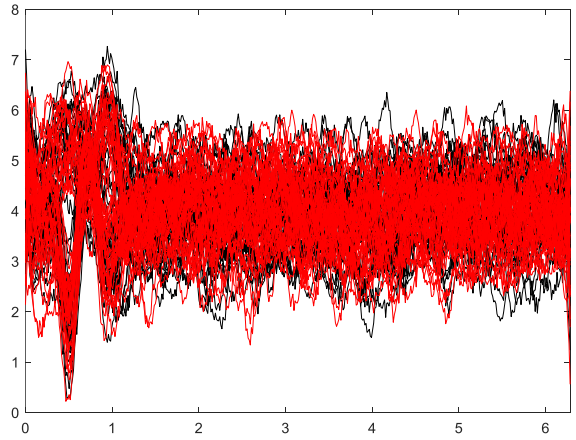


图 4.1 随机模拟数据 II

(2) Growth 成长数据

Growth 数据集源自伯克利生长研究，主要对儿童及青少年身高增长轨迹进行聚类分析，用以检验聚类效果是否能够显现性别间的差异性。该数据集包含了 1 至 18 岁之间，共 31 个时间节点的 54 位女性和 39 位男性儿童的身高数据，并对样本进行性别类别的标注。参考图 4.2 呈现的数据可知，X 轴表示年龄，Y 轴表示身高；此外，不同色彩和线条类型则代表着不同性别的分类信息。

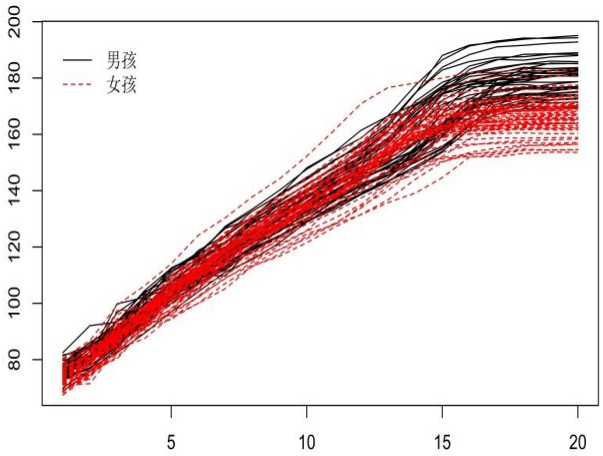


图 4.2 Growth 数据

(3) CanadianWeather 数据集

CanadianWeather 数据集来自 R 包 fda，此数据集记录了 1960 年至 1994 年期间，加拿大 35 个不同地点的日平均气温和降水量。本实验只采取了一年中每天的平均日降雨量，四舍五入至 0.1 毫米。标签为 4 个气候区，主要为：大西洋，太平洋，大陆和北极。图中，横轴表示日期信息，纵轴为降雨量的变化情况。同时，通过不同的颜色以及线型的区分，直观地刻画了各自所代表的气候区域划分。

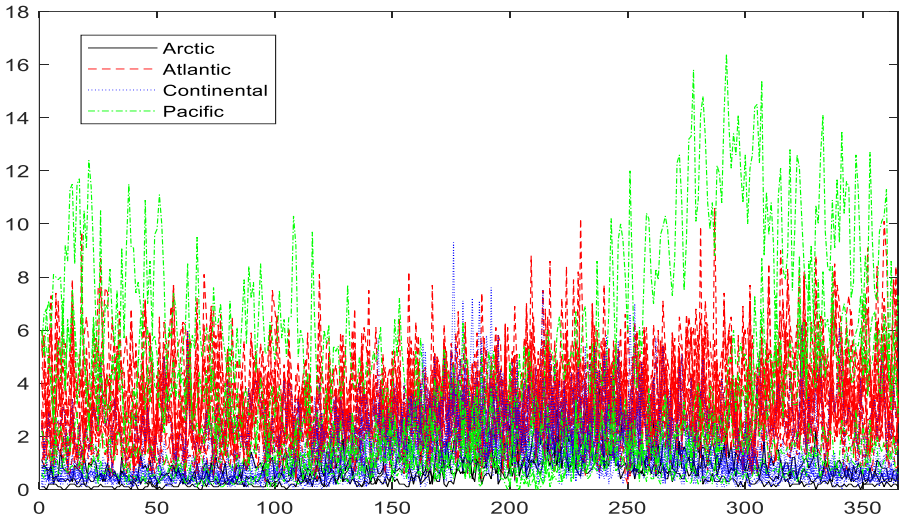


图 4.3 CanadianWeather 数据

(4) Fatspectrum 数据集

Fatspectrum 数据集又称为脂肪光谱数据集，该数据首次在 Kalivas(1997)[73] 中进行了分析，并可在 R 包 fds 中获得。此数据集包括了 100 个小麦样品的近红外反射光谱，这些数据通过近红外传输(NIT)原理记录在波长范围 850 - 1050nm 的 Tecator 红外食品和饲料分析仪上，每隔 2nm 记录一次，还包含了相关的响应变量脂肪含量。由于这些样品没有自然分组，我们将其人工分为两组，类似于 Ferraty 和 Vieu(8.4.2 节)[74]：我们将含脂肪量小于 20 的 n1 = 138 条曲线分配给第 1 组，其余 n2 = 77 条曲线分配给第 2 组。原始数据及类别标签的样本示例如图 4 所示。

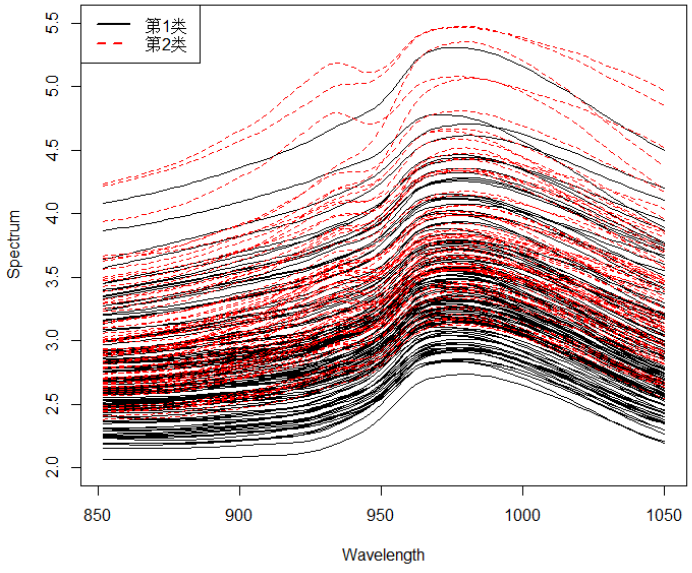


图 4.4 Fatspectrum 数据

4.4.2 参数设置

在进行聚类研究的过程中，对参数进行了如下配置：(1) 以三次等间距节点B-样条基函数为工具进行曲线逼近，通过对基函数个数的精准操控来达到对曲线光滑度的调节，具体在 4 组数据中分别设定基函数的个数为 8、20、65、48；(2) 在随机模拟数据 II 中包含的两种类型数据，$r=2$；在 Growth 数据集中，根据性别维度进行分别处理，该性别维度涵盖男性与女性两个分类，则$r=2$；CanadianWeather 数据集中包含了大西洋、太平洋、大陆和北极四个气候区，故取聚类数$r=4$；Fatspectrum 数据集人工手动分为 2 类，故聚类数$r=2$。

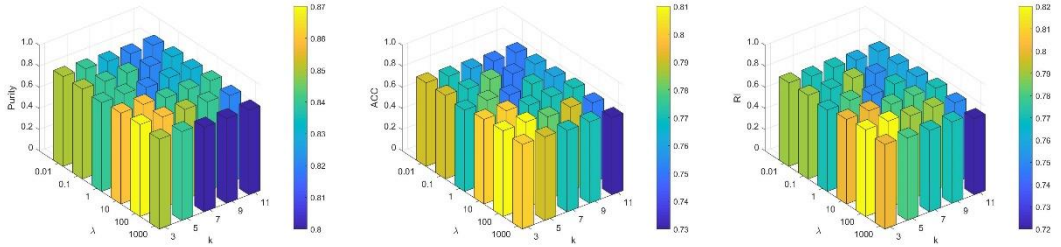
在参数设定基本一致的基础上，利用随机模拟数据 II、Growth 数据、CanadianWeather 数据和 Fatspectrum 数据，将上述 9 种方法进行比较。

4.4.3 实验结果

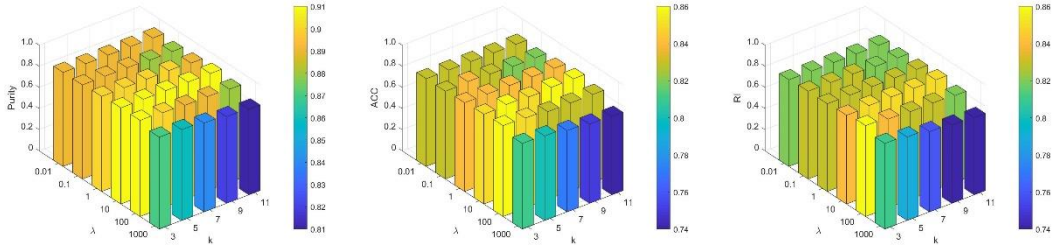
将本章提出的 FRMNMF 算法在随机模拟数据 II、Growth 数据、CanadianWeather 数据和 Fatspectrum 数据上进行实验，本算法采用 PUR、ACC 以及 RI 三种评估指标对聚类效果进行定量分析。为了克服算法初始化设置对聚类成果的潜在影响，本实验每种情景下重复实施 50 次，并选取评价指标的平均值来全面评估聚类算法的性能。

式(4.1)中正则化参数λ平衡流形学习项，KNN 算法中K的取值决定相似图的构造，而聚类性能很大程度上取决于数据相似度矩阵。因此，我们在实验中充分考虑了超参数λ和K值对聚类性能的影响，分别选取$\lambda=\{0.01,0.1,1,10,100,1000\}$，$K=\{3,5,7,9,11\}$。在研究$\lambda$时，$K$的参数被设定为最优值。当研究$K$时，同样的设置也适用。

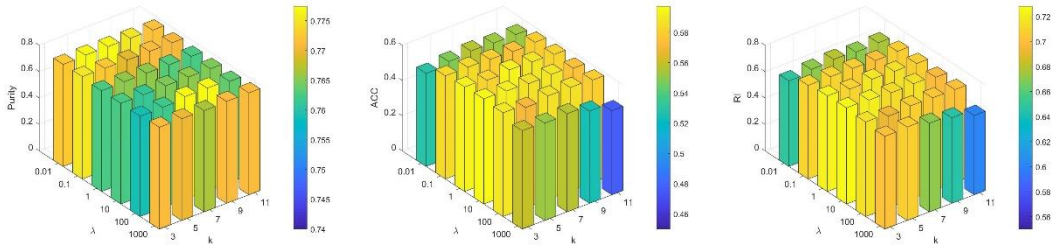
依次在随机模拟数据 II、Growth 数据、CanadianWeather 数据和 Fatspectrum 数据上进行模拟实验，聚类效果随参数λ和K的变化趋势如图 4.5 所示：



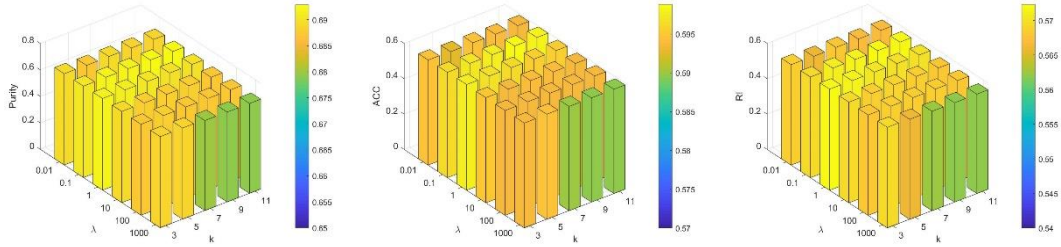
(a) 随机模拟数据 II



(b) Growth 数据



(c) CanadianWeather 数据



(d) Fatspectrum 数据

图 4.5 聚类效果随参数的变化趋势

在图 4.5 中，可以直观看到，FRMNMF 算法的聚类性能与超参数λ和K的取值密切相关，不同组合下，聚类效果各不相同。大多数情况下，当λ固定时，聚类效果随着K值的增大而递减。事实上，由于我们采用 KNN 定义相似矩阵来刻画数据的内在局部几何结构，其前提假设为邻居样本点具有相同的类标签。显然，该假设成立的概率会随着K的增大而减小。此外，注意到当K固定、$\lambda = 1000$时，聚类效果明显降低。但从整体来看，当$\lambda = \{0.01, 0.1, 1, 10, 100\}$时，对 FRFNMF 聚类结果的影响并不是很明显，这意味着该算法具有较好的鲁棒性。

其次，将 FRMNMF 算法与现有 6 种聚类算法（NMF、RNMF、FCOF、FNMF、Funclust、FunHDDC）的聚类评价指标进行比较，结果如表 4.1 所示：

表 4.1 FRMNMF 在随机模拟数据 II 的聚类指标（均值%±标准差%）

聚类指标 / 聚类方法	PUR	ACC	RI
NMF	76.97±12.93	67.23±13.47	67.54±13.61
RNMF	76.95±12.50	67.71±14.39	67.89±14.30
FCOF	70.16±14.03	61.45±13.41	61.54±13.30
FNMF	74.26±14.11	64.73±14.54	65.24±14.49
Funclust	57.46±7.05	51.42±7.15	51.32±7.12
FunHDDC	75.88±14.37	66.98±15.02	66.97±15.12
FRMNMF	**86.43±11.50**	**80.14±16.71**	**79.75±17.04**

表 4.2 FRMNMF 在 Growth 数据的聚类指标（均值%±标准差%）

聚类指标 / 聚类方法	PUR	ACC	RI
NMF	76.02±5.29	67.99±5.69	67.24±5.64
RNMF	58.42±0.37	50.83±0.27	50.08±0.28
FCOF	69.60±3.69	57.76±2.57	57.50±3.05
FNMF	70.28±3.79	58.23±2.64	58.06±3.13
Funclust	58.06±0.00	50.44±0.29	50.01±0.65
FunHDDC	58.06±0.00	51.02±0.26	50.29±0.22
FRMNMF	90.42±3.07	85.32±3.87	84.56±3.83

表 4.3 FRMNMF 在 CanadianWeather 数据的聚类指标（均值%±标准差%）

聚类指标 / 聚类方法	PUR	ACC	RI
NMF	74.99±1.65	42.32±2.51	54.15±3.04
RNMF	76.60±0.90	42.78±3.04	55.79±4.03
FCOF	80.06±0.40	38.37±4.13	49.71±4.52
FNMF	**80.17±0.69**	39.53±7.01	50.99±7.66
Funclust	56.29±1.32	36.95±2.53	56.91±4.51
FunHDDC	53.54±6.90	35.25±5.26	58.49±4.66
FRMNMF	77.67±1.01	**54.07±4.17**	**66.25±4.26**

表 4.4 FRMNMF 在 FatSpectrum 数据的聚类指标（均值%±标准差%）

聚类指标 / 聚类方法	PUR	ACC	RI
NMF	66.99±0.80	57.50±0.87	54.78±0.99
RNMF	64.47±0.21	54.27±0.29	47.95±0.13
FCOF	68.78±0.46	59.32±0.24	56.82±0.35
FNMF	68.60±0.40	59.22±0.21	56.72±0.30
Funclust	64.28±0.29	55.07±0.55	52.82±0.73
FunHDDC	68.63±0.42	59.24±0.22	56.74±0.32
FRMNMF	68.92±0.47	59.52±0.47	56.85±0.52

注：粗体表示比较结果为优。FunHDDC 算法包括 6 个子模型：$a_{kj}b_kQ_kd_k$、$a_{kj}bQ_kd_k$、$a_kb_kQ_kd_k$、$ab_kQ_kd_k$、$a_kbQ_kd_k$和abQ_kd_k。FunHDDC 的结果为 6 个子模型中的最优结果，

自表 4.1 至表 4.4 的实验结果可以得知，本文提出的 FRMNMF 聚类算法在四个不同数据集上均表现出较为优越的聚类性能，这一表现不仅远超相比较的聚类方法，而且在 Growth 数据集的测试结果中表现尤其优异。

最后，对 FRMNMF 算法的收敛性问题进行深入探讨。如 4.3.2 节所示，经过严密的理论推导，本研究已对 FRMNMF 算法的收敛性特性做出了系统性证实。为了更直观地证明该算法的收敛性，图 4.6 为在多个数据集上应用 FRMNMF 算法的收敛曲线，涉及的数据集包括模拟数据集 II、Growth 数据集、CanadianWeather 数据集以及 FatSpectrum 数据集。

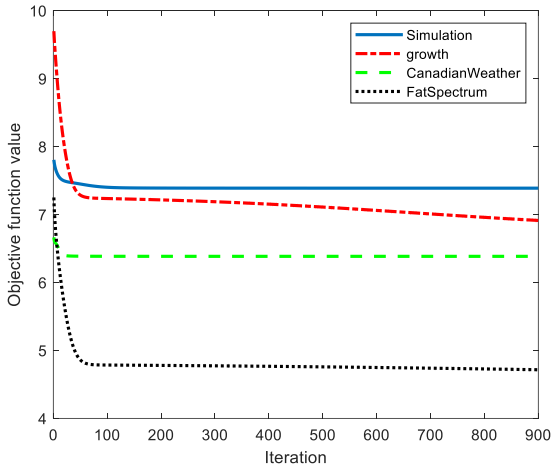


图 4.6 FRMNMF 算法的收敛曲线

图中 x 轴均代表算法的迭代次数，y 轴代表目标函数值取对数。由上图可以看出，所提出的方法在 4 个数据集上可快速收敛，迭代次数在 100 多次便可以达

到收敛状态。

4.5 实例应用—以城镇居民人均可支配收入数据为例

本小节在函数型数据视角下，对我国 31 个省市 2000-2022 年的城镇居民人均可支配收入进行聚类分析，所采用的数据样本来源于《中国统计年鉴》，数据集完整，避免了缺失值处理。然而，鉴于所得数据量值较为庞大，且呈现显著的变异性，故本文采取了对数转换与数据标准化等预处理技术以优化后续分析的效度和稳健性，即：

$$\tilde{y}_{ij} = \log_{10} y_{ij}$$

$$\overline{y}_{ij} = \frac{\tilde{y}_{ij} - \min\tilde{y}_{ij}}{\max\tilde{y}_{ij} - \min\tilde{y}_{ij}}$$

为了更加形象地呈现经过加工处理后的城市居民人均可支配收入数据的发展趋势，我们以年份时间为变量，绘制各省数据折线图，如图 4.7 所示。同时采取箱线图来刻画 2000-2022 年 31 个省份城镇居民可支配收入的总体趋势，如图 4.8 所示。

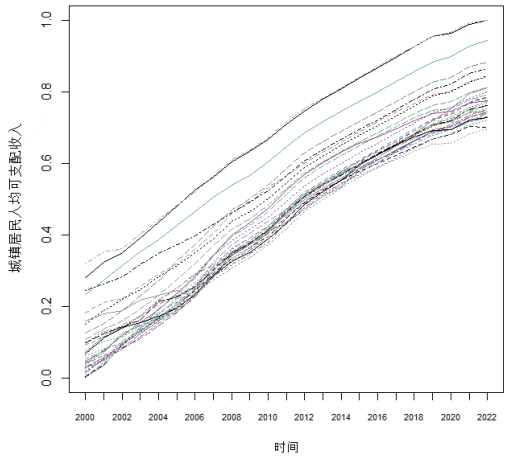


图 4.7 2000-2022 年数据折线图

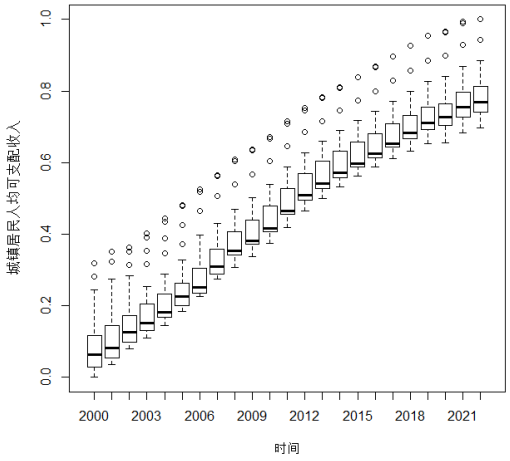


图 4.8 2000-2022 年 31 省市数据箱线图

为了保证拟合效果的光滑性，采用等距节点 3 次 B-样条基底来拟合曲线，拟合曲线如图 4.9 所示。

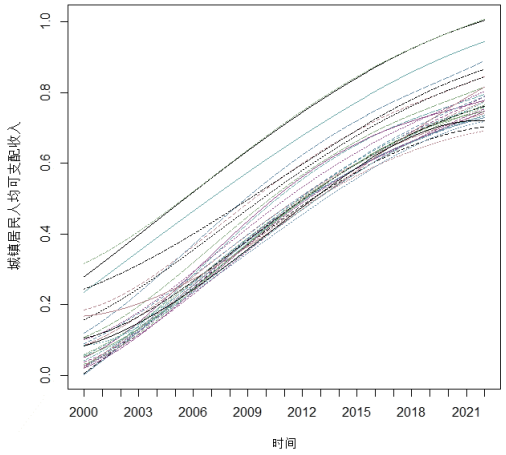


图 4.9 31 个省城镇居民人均可支配收入拟合曲线

图 4. 10 为部分省市的城镇居民人均可支配收入的拟合曲线。

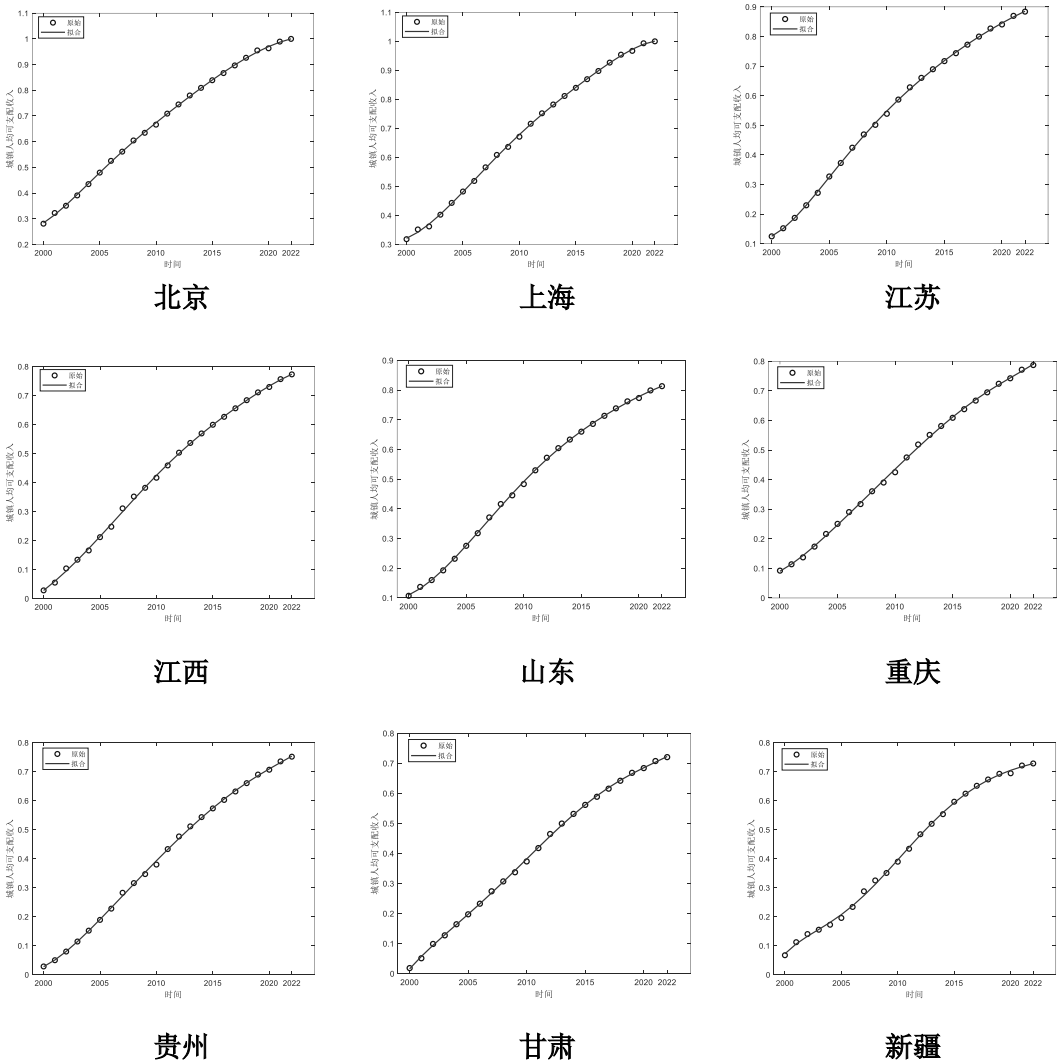


图 4.10 部分省市城镇居民人均可支配收入拟合曲线

利用本章所提出的算法，采用 FRMNMF 算法对数据矩阵进行函数型聚类分析，选取类别数$r=5$进行聚类，得到聚类结果如表 4.5 所示。

表 4.5 城镇居民人均可支配收入聚类结果分类表

类别	省级行政区
第 1 类	上海、浙江、北京
第 2 类	辽宁、湖南、山东、四川、内蒙古
第 3 类	天津、福建、重庆、江苏、广东
第 4 类	广西、湖北、河北、山西、吉林、江西、河南、海南、云南、安徽
第 5 类	黑龙江、西藏、陕西、甘肃、青海、新疆、宁夏、贵州

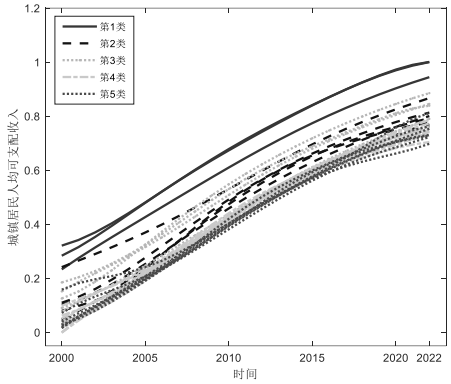


图 4.11 城镇居民人均可支配收入聚类结果图

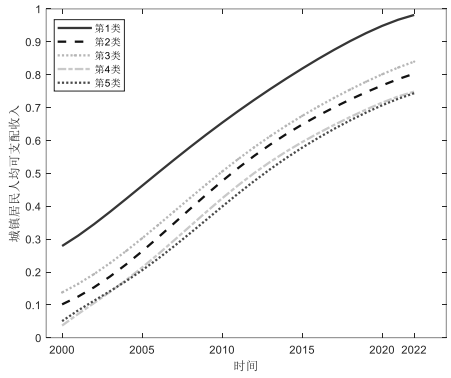


图 4.12 城镇居民人均可支配收入类中心图

图 4.11 和图 4.12 分别为各省市城镇居民人均可支配收入的函数型聚类结果曲线图和类中心曲线图。图中的聚类信息以颜色和线型的不同进行区分。

从图 4.11 和 4.12 可以看出，近 20 年来，我国 31 个省市的城镇居民收入一直处于增长状态，且第 1 类的上海市、浙江省、北京市的城镇居民可支配收入水平明显超越其他级别城市。相较之下，第 5 类的黑龙江省、西藏自治区、陕西省、甘肃省、青海省、新疆维吾尔自治区等地区的城镇人均可支配收入不仅处于低位，其增长速度亦相对缓慢。针对这一现状，相关区域需针对本地实际情况，定制合理的收入提升策略，同时应加大与周边收入水平较高地区在交通、商贸等方面的互联互通，以便加速这些省级行政单元城镇居民收入的增长，进而促进区域经济的均衡发展。

4.6 本章小结

本章提出了一种基于鲁棒流形非负矩阵分解的函数型聚类算法，该算法采用$l_{2,1}$范数来衡量矩阵分解的重构误差，对噪声和异常值具有较小的敏感性。考虑到数据的流形结构以及其局部不变性，我们通过引入拉普拉斯矩阵来构建低维空

间$\boldsymbol{V}$，以便能够有效地表达高维空间$\boldsymbol{Y}$中的数据几何特征。我们为函数型数据聚类构建了一个鲁棒的基于图的 NMF 模型。给出$\boldsymbol{U}$、$\boldsymbol{V}$和$\boldsymbol{G}$的更新公式，证明了该算法的收敛性，同时对其时间复杂度进行了定量分析。分别在随机模拟数据集 II、Growth 数据集、CanadianWeather 数据集和 Fatspectrum 数据集上的实验表明，FRMNMF 算法的性能优于其他聚类算法，并且能够快速收敛。

5 自加权不完整多视角函数型聚类算法

对于存在数据缺失的多视角函数型数据，本章提出了一种适用于缺失多视角函数型数据的聚类算法。依托于非负矩阵分解理论框架，本算法融合了函数型聚类、矩阵填充及多视角聚类方法，创新性地构建了一种集插补与聚类于一体的一步式模型。给出模型交替迭代更新公式，并对其收敛性进行了严谨的证明。同时，深入分析了模型求解过程中的计算复杂性。通过模拟实验对该模型的可行性进行了检验，并且针对北京市空气污染物小时浓度数据的实证分析结果显示，该研究提出的 AIMFC 算法不仅可以妥善处理含随机缺失数据的情形，而且具有较为出色的聚类性能。

5.1 问题概述

在现实应用中，函数型数据通常也会以多元形式出现，而多视角函数型聚类方法面临着问题，现实世界中观测到的多视角函数型数据总是存在缺失数据的情况，缺少数据可能会使传统的方法难以直接使用，这就导致了在函数型数据视角下的不完全多视图聚类问题。

对于多视角缺失数据的处理方法目前的研究主要有两类：一是视角中数据不完整，对于某些视角中部分数据点的信息缺失，一般利用其他视角的完整信息对缺失视角进行插补，如武森等(2012)[75] 计算所有对象的总体差异，以数据聚类的结果进行缺失数据填补；Tao 和 Hou 等(2019)[76] 提出了用于不完备多视角图学习的联合嵌入学习和低秩近似框架。二是视角缺失，在某些视角中数据点的信息完全缺失，但是在其他视角中没有数据点缺失，可以充分利用视角内相似性和视角间互补性进行数据插补，如 Liu 和 Li(2021)[77]用学习的共识聚类矩阵来插补由不完整视图生成的每个不完整基础矩阵；Wen 和 Zhang(2019)[78] 利用视图之间的互补信息、各视图的局部信息来学习样本的潜在表示，可以处理不完整和完整的多视图聚类。

本文主要针对于第一种情况展开研究，为了解决这个问题，提出一种适用于缺失函数型数据集的自加权多视图聚类软方法(AIMFC)，能够同时完成数据插补和多视角聚类。

后续的相关内容安排如下：5.2 构建所提出 AIMFC 模型的框架；5.3 给出迭代优化算法求解所提出的目标函数；5.4 为实验部分，利用随机模拟数据 III，将数据处理为缺失数据，验证所提出的 AIMFC 算法的聚类性能，并且对收敛速度和时间复杂度都进行了分析；5.5 为实例应用，利用 2016 年北京市空气污染物小时浓度数据进行数据插补和聚类应用；5.6 为本章小结。

5.2 算法模型

对于数据矩阵$\boldsymbol{Y}$，给定不完备指示矩阵$\boldsymbol{O}$:

$$\boldsymbol{O}_{ij}=\begin{cases}0, & \boldsymbol{Y}_{ij}\text{值缺失;}\\ 1, & \boldsymbol{Y}_{ij}\text{值可观测}\end{cases} \tag{5.1}$$

在函数型数据分析视角下，结合非负矩阵分解，则有：

$$\begin{gathered}\min_{\boldsymbol{U}_v,\boldsymbol{V}_v}\|\boldsymbol{O}\odot(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{U}\boldsymbol{V}^T)\|_F^2\\ \text{s.t. }\boldsymbol{U}\geq 0,\boldsymbol{V}\geq 0\end{gathered} \tag{5.2}$$

上式是关于单视角函数型缺失数据聚类问题的描述，因此，将此方法延伸至多视角领域则有：

$$\begin{gathered}\min_{\boldsymbol{U}_v,\boldsymbol{V}_v}\sum_{v=1}^{n_v}\|\boldsymbol{O}_v\odot(\boldsymbol{Y}_v-\boldsymbol{\Phi}_v\boldsymbol{U}_v\boldsymbol{V}_v^T)\|_F^2\\ \text{s.t. }\boldsymbol{U}_v\geq 0,\boldsymbol{V}_v\geq 0\end{gathered} \tag{5.3}$$

其中，$\boldsymbol{Y}_v$表示第v个视角带有缺失值的数据矩阵，$\boldsymbol{O}_v$表示第v个视角的不完备指示矩阵，$\boldsymbol{\Phi}_v$表示第v个视角的基底矩阵，$\boldsymbol{V}_v$反映了第v个视角下函数型数据的聚类特性。为了充分考虑到数据视角间信息的互补性和一致性，构造损失函数：

$$\mathcal{N}(\boldsymbol{V}_v,\boldsymbol{V}^*)=\|\boldsymbol{V}_v-\boldsymbol{V}^*\|_F^2 \tag{5.4}$$

其中，共识度矩阵$\boldsymbol{V}^*$用于整合不同视角的潜在聚类特征；另一方面，$\mathcal{N}(\boldsymbol{V}_v,\boldsymbol{V}^*)$则用于衡量$\boldsymbol{V}_v$和$\boldsymbol{V}^*$之间的差异。基于此，合并各个视角的聚类特征后进行具体分析。则有：

$$\begin{gathered}\min_{\boldsymbol{U}_v\boldsymbol{V}_v,\boldsymbol{V}^*}\sum_{v=1}^{n_v}\{\|\boldsymbol{O}_v\odot(\boldsymbol{Y}_v-\boldsymbol{\Phi}_v\boldsymbol{U}_v\boldsymbol{V}_v^T)\|_F^2+\alpha_v\|\boldsymbol{V}_v-\boldsymbol{V}^*\|_F^2\}\\ \text{s.t. }\boldsymbol{U}_v\geq 0,\boldsymbol{V}_v\geq 0,\boldsymbol{V}^*\geq 0\end{gathered} \tag{5.5}$$

其中，α_v为调节参数。为了求解更稳定，保证结果的光滑性，将正则化项加入到

模型中，即对$\boldsymbol{V}$添加$l_{2,1}$范数惩罚，取第v个视角的惩罚项为$\|\boldsymbol{V}_v\|_{2,1}$，则对所有的视角的惩罚为

$$\sum_{v=1}^{n_v} \beta_v \|\boldsymbol{V}_v\|_{2,1} \tag{5.6}$$

其中，$\beta_v(v=1,2,\cdots,n_v)$表示第$v$个视角的惩罚参数。

不同的视角具有不同的可用性，但是共同的潜在特征矩阵$\boldsymbol{V}^*$并不直接包含关于各个视角的可用性的信息。每个视角的数据可用性信息存在差异，这些信息可以区分不同的观点，也会影响聚类插补效果。因此只利用矩阵$\boldsymbol{V}^*$很难区分不同视角的可用性，因此，我们为每个视角定义分配权重因子w_v，且自适应地为每个视角分配适当的权重，在迭代时可以自动更新权重值以获得更好的聚类效果。

综合上述，可以得到自加权不完整多视角函数型聚类算法(AIMFC)：

$$\begin{aligned} &\min_{\boldsymbol{U}_v\boldsymbol{V}_v,\boldsymbol{V}^*} \sum_{v=1}^{n_v} \{w_v \|\boldsymbol{O}_v \odot (\boldsymbol{Y}_v - \boldsymbol{\Phi}_v \boldsymbol{U}_v \boldsymbol{V}_v^T)\|_F^2 + \alpha_v \|\boldsymbol{V}_v - \boldsymbol{V}^*\|_F^2 + \beta_v \|\boldsymbol{V}_v\|_{2,1}\} \\ &\text{s.t. } \boldsymbol{U}_v \geq 0, \boldsymbol{V}_v \geq 0, \boldsymbol{V}^* \geq 0 \end{aligned} \tag{5.7}$$

其中：

$$w_v = \frac{1}{2\|\boldsymbol{O}_v \odot (\tilde{\boldsymbol{Y}}_v - \boldsymbol{\Phi}_v \boldsymbol{U}_v \boldsymbol{V}_v^T)\|_F} \tag{5.8}$$

5.3 求解算法

5.3.1 优化求解

对于目标函数式(5.7)采用迭代更新算法求解，具体过程如下：

（1）固定$\boldsymbol{V}^*$，最小化式求解$\boldsymbol{U}_v$和$\boldsymbol{V}_v$

对于给定的视角v，目标函数式可简化为：

$$\begin{aligned} &\min_{\boldsymbol{U}_v,\boldsymbol{V}_v} w_v \|\boldsymbol{O}_v \odot (\boldsymbol{Y}_v - \boldsymbol{\Phi}_v \boldsymbol{U}_v \boldsymbol{V}_v^T)\|_F^2 + \alpha_v \|\boldsymbol{V}_v - \boldsymbol{V}^*\|_F^2 + \beta_v \|\boldsymbol{V}\|_{2,1} \\ &\text{s.t. } \boldsymbol{U}_v \geq 0, \boldsymbol{V}_v \geq 0 \end{aligned} \tag{5.9}$$

①固定$\boldsymbol{V}^*$和$\boldsymbol{V}_v$：

拉格朗日函数为：

$$\mathcal{L}_1 = w_v \|\boldsymbol{O}_v \odot (\boldsymbol{Y}_v - \boldsymbol{\Phi}_v \boldsymbol{U}_v \boldsymbol{V}_v^T)\|_F^2 - \mathrm{tr}(\boldsymbol{\Lambda}_v \boldsymbol{U}_v^T)$$

其中，$\boldsymbol{\Lambda}_v$为$\boldsymbol{U}_v$的拉格朗日乘子矩阵。对于$\mathcal{L}_1$关于$\boldsymbol{U}_v$求偏导，并令$\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{U}_v} = 0$，得到：

$$\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{U}_v} = -2w_v\boldsymbol{\Phi}_v^T(\boldsymbol{O}_v \odot \boldsymbol{Y}_v)\boldsymbol{V}_v + 2w_v\boldsymbol{\Phi}_v^T(\boldsymbol{O}_v \odot \boldsymbol{\Phi}_v\boldsymbol{U}_v\boldsymbol{V}_v^T)\boldsymbol{V}_v - \boldsymbol{\Lambda}_v = 0$$

从而有：

$$\boldsymbol{\Lambda}_v = -2\boldsymbol{\Phi}_v^T(\boldsymbol{O}_v \odot \boldsymbol{Y}_v)\boldsymbol{V}_v + 2\boldsymbol{\Phi}_v^T(\boldsymbol{O}_v \odot \boldsymbol{\Phi}_v\boldsymbol{U}_v\boldsymbol{V}_v^T)\boldsymbol{V}_v$$

利用 KKT 条件$\boldsymbol{\Lambda}_v \odot \boldsymbol{U}_v = 0$，可得矩阵$\boldsymbol{U}_v$的更新公式：

$$\boldsymbol{U}_{vij} \leftarrow \boldsymbol{U}_{vij}\frac{(w_v\boldsymbol{\Phi}_v^T(\boldsymbol{O}_v \odot \boldsymbol{Y})_v\boldsymbol{V}_v)_{ij}}{(w_v\boldsymbol{\Phi}_v^T(\boldsymbol{O}_v \odot \boldsymbol{\Phi}_v\boldsymbol{U}_v\boldsymbol{V}_v^T)\boldsymbol{V}_v)_{ij}} \tag{5.10}$$

②固定$\boldsymbol{V}^*$和$\boldsymbol{U}_v$：

相应的拉格朗日函数为：

$$\mathcal{L}_2 = w_v\|\boldsymbol{O}_v \odot (\boldsymbol{Y}_v - \boldsymbol{\Phi}_v\boldsymbol{U}_v\boldsymbol{V}_v^T)\|_F^2 + \alpha_v\|\boldsymbol{V}_v - \boldsymbol{V}^*\|_F^2 + \beta_v\|\boldsymbol{V}_v\|_{2,1} - \mathrm{tr}(\boldsymbol{\Gamma}_v\boldsymbol{V}_v^T)$$

其中，$\boldsymbol{\Gamma}_v$为$\boldsymbol{V}_v$的拉格朗日乘子矩阵。对于$\mathcal{L}_2$关于$\boldsymbol{V}_v$求偏导，并令$\frac{\partial \mathcal{L}_2}{\partial \boldsymbol{V}_v} = 0$，得到：

$$\begin{aligned}\frac{\partial \mathcal{L}_2}{\partial \boldsymbol{V}_v} =& -2w_v(\boldsymbol{O}_v \odot \boldsymbol{Y}_v^T)\boldsymbol{\Phi}_v\boldsymbol{U}_v + 2w_v(\boldsymbol{O}_v \odot \boldsymbol{V}_v\boldsymbol{U}_v^T\boldsymbol{\Phi}_v^T)\boldsymbol{\Phi}_v\boldsymbol{U}_v \\ &+ 2\alpha_v(\boldsymbol{V}_v - \boldsymbol{V}^*) + \frac{\beta_v\boldsymbol{V}_v}{\sqrt{(\sum_k(V_{vik})^2)}} - \boldsymbol{\Gamma}_v \\ &= 0\end{aligned}$$

从而有：

$$\begin{aligned}\boldsymbol{\Gamma}_v =& -2w_v(\boldsymbol{O}_v^T \odot \boldsymbol{Y}_v^T)\boldsymbol{\Phi}_v\boldsymbol{U}_v + 2w_v(\boldsymbol{O}_v^T \odot \boldsymbol{V}_v\boldsymbol{U}_v^T\boldsymbol{\Phi}_v^T)\boldsymbol{\Phi}_v\boldsymbol{U}_v \\ &+ 2\alpha_v(\boldsymbol{V}_v - \boldsymbol{V}^*) + \frac{\beta_v\boldsymbol{V}_v}{\sqrt{\sum_k(V_{vik})^2}}\end{aligned}$$

利用 KKT 条件$\boldsymbol{\Gamma}_v \odot \boldsymbol{V}_v = 0$则矩阵$\boldsymbol{V}_v$的更新公式为：

$$\boldsymbol{V}_{vij} \leftarrow \boldsymbol{V}_{vij}\frac{(w_v(\boldsymbol{O}_v^T \odot \boldsymbol{Y}_v^T\boldsymbol{\Phi}_v\boldsymbol{U}_v)_{ij} + \alpha_v\boldsymbol{V}_{ij}^*}{(w_v(\boldsymbol{O}_v^T \odot \boldsymbol{V}_v\boldsymbol{U}_v^T\boldsymbol{\Phi}_v^T)\boldsymbol{\Phi}_v\boldsymbol{U}_v)_{ij} + \alpha_v\boldsymbol{V}_{vij} + \frac{\beta_v\boldsymbol{V}_{vij}}{2\sqrt{\sum_k(\boldsymbol{V}_{vik})^2}}} \tag{5.11}$$

（2）固定$\boldsymbol{U}_v$和$\boldsymbol{V}_v$，求解矩阵$\boldsymbol{V}^*$

令$\boldsymbol{\Psi}$为$\boldsymbol{V}^*$的拉格朗日乘子矩阵，则相应的拉格朗日函数为：

$$\mathcal{L}_3 = \sum_{v=1}^{n_v}\alpha_v\|\boldsymbol{V}_v - \boldsymbol{V}^*\|_F^2 - \mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{V}^{*T})$$

对于$\mathcal{L}_3$关于$\boldsymbol{V}^*$求偏导，并令$\frac{\partial \mathcal{L}_3}{\partial \boldsymbol{V}^*} = 0$，则有：

$$\begin{aligned}\frac{\partial \mathcal{L}_3}{\partial \boldsymbol{V}^*} &= \frac{\partial \mathrm{tr}\left[\alpha_v(\boldsymbol{V}_v^T\boldsymbol{V}_v - 2\boldsymbol{V}^{*T}\boldsymbol{V}_v + \boldsymbol{V}^{*T}\boldsymbol{V}^*) - \boldsymbol{\Psi}\boldsymbol{V}^{*T}\right]}{\partial \boldsymbol{V}^*} \\ &= \sum_{v=1}^{n_v} -2\alpha_v\boldsymbol{V}_v + \sum_{v=1}^{n_v} 2\alpha_v\boldsymbol{V}^* - \boldsymbol{\Psi} \\ &= 0\end{aligned}$$

从而可以得到：

$$\boldsymbol{\Psi} = 2\sum_{v=1}^{n_v} \alpha_v (\boldsymbol{V}^* - \boldsymbol{V}_v)$$

利用 KKT 条件$\boldsymbol{\Psi} \odot \boldsymbol{V}^* = 0$，可得矩阵$\boldsymbol{V}^*$的更新规则为：

$$\boldsymbol{V}^* = \frac{\sum_{v=1}^{n_v} \alpha_v \boldsymbol{V}_v}{\sum_{v=1}^{n_v} \alpha_v} \tag{5.12}$$

根据式(5.10)、式(5.11)、式(5.12)分别迭代更新$\boldsymbol{V}_v$、$\boldsymbol{U}_v$和$\boldsymbol{V}^*$，即可实现目标函数式(5.7)的求解算法。

5.3.2 算法流程

本算法所提出的自加权不完整多视角函数型聚类算法(AIMFC)的一般求解框架具体阐述如下：

算法 5.1 自加权不完整多视角函数型聚类算法(AIMFC)

输入：数据矩阵$\boldsymbol{Y}_v$、聚类数r、基底矩阵$\boldsymbol{\Phi}_v$、参数α_v和β_v，最大迭代次数maxiter
过程：
1：初始化$\boldsymbol{U}_v^{(0)}$、$\boldsymbol{V}_v^{(0)}$和$\boldsymbol{V}^{*(0)}$
2：构造不完备指示矩阵$\boldsymbol{O}_v$
3：for $i = 1, 2, \cdots, \text{maxiter}$
4：for $v = 1, 2, \cdots, n_v$
5：固定$\boldsymbol{V}^{*(t-1)}$和$\boldsymbol{V}_v^{(t-1)}$，根据式(5.16)更新$\boldsymbol{U}_v^{(t-1)}$
6：固定$\boldsymbol{V}^{*(t-1)}$和$\boldsymbol{U}_v^{(t)}$，根据式(5.17)更新$\boldsymbol{V}_v^{(t-1)}$
7：end for
8：固定$\boldsymbol{V}_v^{(t)}$和$\boldsymbol{U}_v^{(t)}$，根据式(5.18)更新$\boldsymbol{V}^{*(t-1)}$
9：if 式收敛
10：break
11：end if
12：end for
输出：$\boldsymbol{V}_v^{(t)}$、$\boldsymbol{U}_v^{(t)}$、$\boldsymbol{V}^{*(t)}$、簇划分$c = \{c_1, c_2, \cdots, c_k\}$

5.3.3 时间复杂度分析

AIMFC 算法的时间复杂度主要包括以下四部分：

① 对于$\boldsymbol{V}_v$的更新(式(5.10))，计算的时间复杂度为：$O(npr + npm + nr)$，则更新n_v个视角的变量$\boldsymbol{V}_v$总时间复杂度为：$O(n_v(npr + npm + nr))$。

② 根据式(5.11)，计算$\boldsymbol{U}_v$的时间复杂度为：$O(npr + npm + p^2m + p^2r + pr)$，

则更新n_v个视角的变量$\boldsymbol{U}_v$总时间复杂度为：$O(n_v(npr+npm+p^2m+p^2r+pr))$。

③ 对于$\boldsymbol{V}^*$的更新(式(5.12))，计算的时间复杂度为$O(n_v nr)$。

④ 根据式(5.8)，计算w_v的时间复杂度为$O(pmr+mnr)$

则总时间复杂度为$O(In_v(npr+(np+pr+nr+n+p^2)m+r(p^2+p+n)))$。

5.4 模拟实验

为进一步验证 AIMFC 算法的性能优越性，本研究在参数设定一致的前提下，将该算法与当前处理含缺失数据单视角聚类情景中的 WNMF[79]、特征连接方法 ConcatKmeans，以及用于多视角函数型缺失数据聚类的 MFWNMF 方法展开对比。以 PUR、ACC 和 RI 为评价指标，对不同算法在聚类任务中的表现进行综合评估。

本实验中的所有代码在 MATLAB R2018a 及 R3.4.1 软件中实现，实验的计算机环境为：12th Gen Intel(R) Core(TM) i5-12500H 2.50 GHz，内存 16GB，Windows11 64 位操作系统。

5.4.1 实验数据

本章的模拟实验选取随机模拟数据 I (3.4.1(1))进行，将数据处理为非负，得到随机模拟数据 III，如图 5.1 所示。

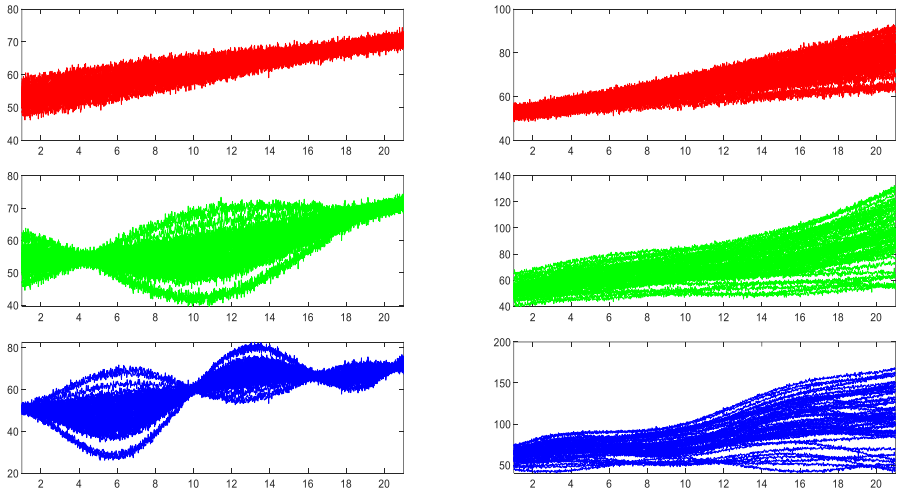


图 5.1 随机模拟数据 III

5.4.2 参数设置

在 AIMFC 算法的应用过程中，需对其参数进行设置，其中包括以下两个：(1) 随机模拟数据 III 所涉及的视角数为 2，每个视角分别包含 3 类数据，因此，视角数$n_v = 2$，$r = 3$；(2) 采用 3 次 B-样条基底进行曲线的拟合处理，各个视角的基底数量为30。WNMF、ConcatKmeans 和 MFWNMF 的相关参数设置与 AIMFC 算法保持一致，WNMF 方法在两个视角上分别进行实验，选取最优的结果；ConcatKmeans 方法先用均值来填充缺失值，接着将两个视角连接，直接在连接数据上进行 Kmeans 算法。

5.4.3 聚类结果

因为所提出的算法针对于不完备数据，因此需要对于随机模拟数据 III 做随机缺失处理。为探究所提出的 AIMFC 算法在不同数据缺失比例条件下的聚类性能，因此设计了一系列实验。在本实验中，将 AIMFC 算法分别应用到数据缺失率为 10%、20%、30%、40%及 50%的数据集上，并与 ConcatKmeans、WNMF 和 MFWNMF 算法做了横向比较。实验结果经详细分析后，如表 5.1 至表 5.3 所示，表明了不同缺失比例下 AIMFC 算法的聚类结果。

表 5.1 AIMFC 在随机模拟数据 III 的 ACC（均值±标准差）

方法 / 缺失率	ConcatKmeans	WNMF	MFWNMF	AIMFC
10%	0.48±0.0420	0.50±0.0384	0.51±0.0611	**0.51±0.0342**
20%	0.50±0.0792	0.49±0.0491	0.51±0.0366	**0.51±0.0386**
30%	0.48±0.0650	0.49±0.0456	0.49±0.0432	**0.51±0.0538**
40%	0.49±0.0689	0.48±0.0673	0.49±0.0527	**0.49±0.0416**
50%	0.46±0.0506	0.48±0.0408	0.49±0.0495	**0.49±0.0452**

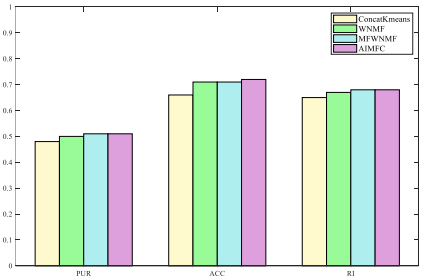
表 5.2 AIMFC 在随机模拟数据 III 的 PUR（均值±标准差）

方法 / 缺失率	ConcatKmeans	WNMF	MFWNMF	AIMFC
10%	0.66±0.0584	0.71±0.0374	0.71±0.0483	**0.72±0.0248**
20%	0.67±0.0924	0.70±0.0466	0.71±0.0297	**0.72±0.0367**
30%	0.65±0.0772	**0.70±0.0384**	0.70±0.0411	0.70±0.0458
40%	0.68±0.0855	0.68±0.0582	0.69±0.0437	**0.69±0.0419**
50%	0.64±0.0736	0.68±0.0378	**0.69±0.0476**	0.68±0.0469

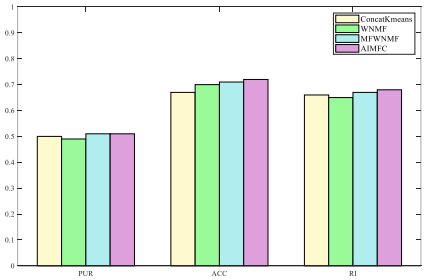
表 5.3 AIMFC 在随机模拟数据 III 的 RI（均值±标准差）

方法 / 缺失率	ConcatKmeans	WNMF	MFWNMF	AIMFC
10%	0.65±0.0373	0.67±0.0334	0.68±0.0512	**0.68±0.0277**
20%	0.66±0.0702	0.65±0.0456	0.67±0.0305	**0.68±0.0311**
30%	0.65±0.0538	0.66±0.0398	0.66±0.0380	**0.68±0.0417**
40%	0.66±0.0635	0.65±0.0580	0.66±0.0315	**0.66±0.0305**
50%	0.63±0.0483	0.65±0.0361	0.65±0.0440	**0.66±0.0390**

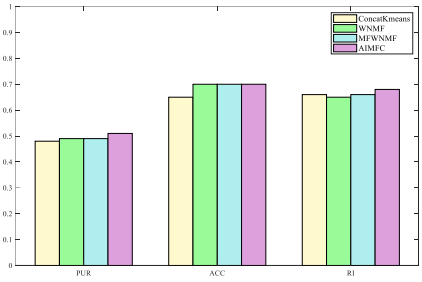
从表 5.1、表 5.2 和表 5.3 可以看出，对于视角中数据不完备的情况，我们提出的 AIMFC 算法聚类效果较好，且对于缺失率较大的数据上，聚类值相对稳定；经实验证明，ConcatKmeans 方法在处理多视角数据方面的聚类效果不尽如人意，这说明仅仅通过简单拼接多个视角的数据特征并不能有效地挖掘出各个视角中互补的特征。其次，AIMFC 算法优于单视角聚类算法 WNMF，说明单视角学习不能用于处理多视角数据，多视角学习更有利于充分整合来自每个视图的信息，可以更加有效的将数据划分；最后，对于 MFWNMF 算法来说，我们提出的 AIMFC 算法标准差更小，因此聚类结果更加稳定。



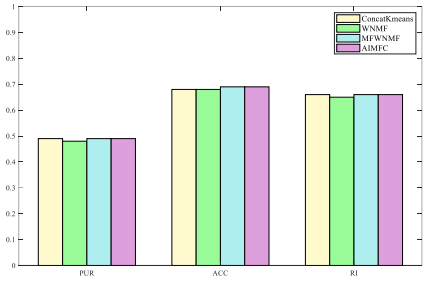
(a) 10%



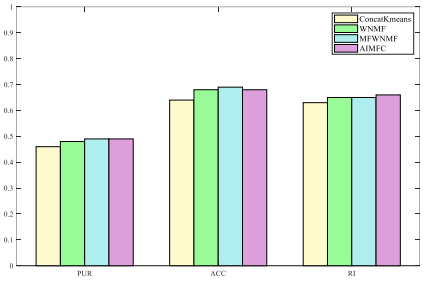
(b)20%



(c)30%



(d)40%



(e)50%

图 5.2 不同聚类方法在不同缺失率的聚类结果比较

把表 5.1-5.3 根据缺失率绘制成柱状图如图 5.2 所示，从图 5.2 可以更加直观的看出，除了 MFWNMF 算法与 AIMFC 算法的聚类指标值大多数相同，我们所提出的算法 AIMFC 相较于 ConcatKmeans、WNMF 算法来说，在缺失率为 10%、20%、30%、40%、50%时不论是 PUR、ACC 以及 RI 的值都是最优的。

我们利用一种简单的方法来获得两个参数的最优组合进行实验。对模型中的参数α_v和β_v进行参数实验，研究参数对于聚类结果的影响。其中，因为随机模拟数据 III 有两个视角，所以我们取$\alpha_v = \frac{1}{2}$，分别取$\beta_v = \{0.01, 0.1, 1, 10, 100\}$，在缺失率分别为 10%、20%、30%、40%和 50%的随机模拟数据上进行聚类。为了消除初始化对于聚类结果的影响，重复进行 30 次实验，采用 PUR、ACC 和 RI 为聚类评价指标，图 5.3 为聚类效果随β_v和缺失率 Incomplete rate 变化的柱状图。

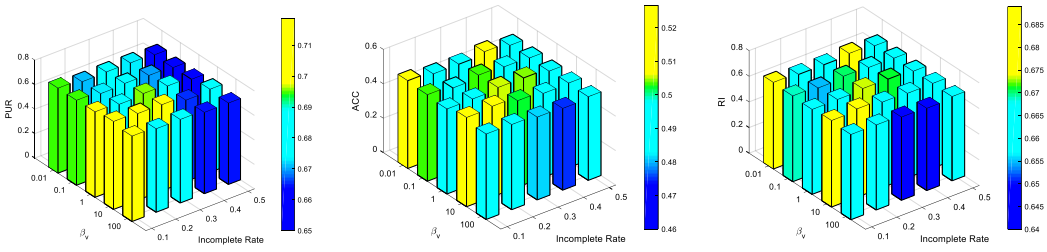


图 5.3 聚类效果随缺失率和β_v的变化趋势

如图 5.3 所示，Incomplete Rate 度量了数据集的缺失程度。通过分析可见，

不同惩罚系数值以及数据缺失率对聚类效果产生了显著影响，数据完整性较高时，所获得的聚类性能相对较佳。具体来说，在缺失率分别为 10%、20%、30%、40%以及 50%的随机模拟数据中，本研究提出的 AIMFC 算法展现出了较为一致的聚类结果，这在一定程度上证明了 AIMFC 算法在面对模型参数时展现出了良好的鲁棒性与稳健性。

我们研究 AIMFC 算法在随机模拟数据上的收敛性，图 5.4 为缺失率分别为 0.1、0.3、0.5 时的收敛曲线图，体现了目标函数值与迭代次数的关系。

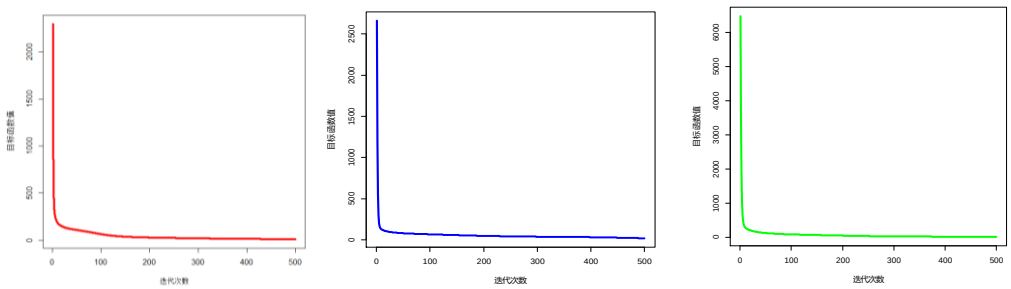


图 5.4 AIMFC 算法的收敛曲线

图 5.4 所示结果表明，AIMFC 算法具备较快的收敛特性，其在迭代不足 200 次的情况下即可实现收敛，且随着数据缺失率上升，能更迅速地使目标函数值达到稳态。

5.5 实例应用—以北京市空气质量监测站点数据为例

为了深入探究和实证 AIMFC 算法的实际应用效能，鉴于空气污染数据的函数特征，我们以 2016 年 1 月 1 日至 12 月 31 日北京市空气污染物小时浓度数据为研究对象，开展相应的应用研究。北京市一共有 35 个空气质量监测站点，检测的主要空气污染物主要包括 CO、NO_2、O_3、$PM_{2.5}$、PM_{10} 和 SO_2，所用污染物浓度数据来自中国环境监测总站(http://www.cnemc.cn/)。

由于数据采集仪器技术等不可抗因素的影响，实际获得的数据存在大量缺失，其中 6 个污染物小时浓度数据缺失率如表 5.4 所示：

表 5.4 北京市主要污染物小时浓度数据缺失率

污染物	CO	NO_2	O_3	$PM_{2.5}$	PM_{10}	SO_2
缺失率	5.84%	5.24%	6.11%	6.13%	35.31%	4.77%

根据表 5.5 所列出的信息可以将北京市的空气质量监测站点总体分为四个类

型，包括站点名称、站点类型、还有经纬度等基础信息。显然从表 5.5 中可以明显发现，“城市清洁对照站点”类仅有一个，因此，我们将剩下的 34 个监测站点所测得的 6 种污染物浓度数据进行聚类分析，并以监测站点类型作为标签，用以验证所得聚类结果的准确性。

表 5.5 北京市空气质量监测站点基本信息

站点类别	编号	站点名称(经纬度坐标)	编号	站点名称(经纬度坐标)	编号	站点名称(经纬度坐标)
城市环境评价站点	1	东四(116.42, 39.93)	9	北部新区(116.17, 40.09)	17	云岗(116.15, 39.82)
	2	良乡(116.14, 39.74)	10	丰台花园(116.28, 39.86)	18	昌平镇(116.23, 40.22)
	3	西城官园(116.34, 39.93)	11	顺义新城(116.66, 40.13)	19	双峪(116.11, 39.94)
	4	万寿西宫(116.35, 39.88)	12	古城(116.18, 39.91)	20	海淀万柳(116.29, 39.99)
	5	夏都(115.97, 40.45)	13	天坛(116.41, 39.89)	21	怀柔镇(116.63, 40.33)
	6	农展馆(116.46, 39.94)	14	黄村(116.40, 39.72)	22	密云镇(116.83, 40.37)
	7	平谷镇(117.10, 40.14)	15	亦庄(116.51, 39.80)	23	奥体中心(116.40, 39.98)
	8	香山(116.21, 40.00)	16	通州北苑(116.660, 39.89)		
区域背景传输站点	24	京西北(115.990, 40.370)	26	京东(117.120, 40.100)	28	京南(116.300, 39.520)
	25	京东北(116.910, 40.500)	27	京东南(116.780, 39.710)	29	京西南(116.000, 39.580)
交通污染控制站点	30	前门(116.400, 39.900)	32	西直门(116.350, 39.950)	34	东四环(116.480, 39.940)
	31	永定门(116.390, 39.880)	33	南三环(116.370, 39.860)		
城市清洁对照站点	35	定陵(116.220, 40.29)				

为了让实证应用能够顺利进行，实验中的参数设定为：(1) 涉及的监测站点共有 3 类，因此取聚类数为$k=3$；(2) 采用等距节点 3 次 B-样条基底来拟合曲线，基底个数均取为 30；(3) 视角数n_v随着聚类实验所需要的污染物类别(依据缺失率)进行取值，分别取$n_v=2,3,4,5,6$。

为了减少初始值对于聚类结果的影响，实验重复进行 30 次，采用 ACC、PUR、RI 作为聚类评价指标，如表 5.6 所示。

表 5.6 北京市污染物小时浓度数据聚类效果(均值±标准差)

评价指标 / 变量个数	ACC	PUR	RI
2	0.5007±0.0118	0.6786±0.0034	0.5056±0.0089
3	0.4974±0.0115	0.6792±0.0033	0.5022±0.0089
4	0.5040±0.0132	0.6796±0.0034	0.5078±0.0103
5	0.5021±0.0097	0.6782±0.0024	0.5064±0.0074
6	**0.5083±0.0083**	**0.6806±0.0043**	**0.5105±0.0061**

注：粗体表示较优的聚类结果。

依据缺失率的大小选取变量，则有表 5.6 中，2 变量表示 $PM_{2.5}$ 和 PM_{10}，3 变量代表 $PM_{2.5}$、PM_{10} 和 O_3，4 变量代表 $PM_{2.5}$、PM_{10}、O_3 和 CO，5 变量代表 $PM_{2.5}$、PM_{10}、O_3、CO 和 NO_2，6 变量代表 $PM_{2.5}$、PM_{10}、SO_2、CO、NO_2 和 O3 六种污染物。

通过观察表 5.6，可以得知针对所有 6 种污染物小时浓度数据所得到的聚类结果最佳。这说明 AIMFC 算法能够充分提取不同视角间的互补信息，并从而为数据插补提供更加有效的信息，进而提升聚类效果。与此同时，AIMFC 算法中的自加权项不仅可以用于处理缺失数据，而且还具备处理大规模缺失数据的能力，并且所得到的算法结果在稳定性方面表现出色。

5.6 本章小结

本章提出了一种可以用于不完备函数型数据的聚类方法 AIMFC。AIMFC 算法在非负矩阵分解的框架下，设定不完备指示矩阵O_v，将缺失数据插补、多视角学习和聚类方法相结合，为了更好地利用视角间的信息，引入自加权项自动更新每个视角的权重因子。给出了迭代更新算法并计算了时间复杂性。首先在随机模拟数据 III 上进行模拟实验，其次在北京市污染物小时浓度数据上进行实证应用，证明 AIMFC 算法的聚类性能较好，并且在不同缺失率上结果比较稳定。

6 结论与展望

本章对论文的主要研究结论进行总结，并对后续研究工作进行展望。

6.1 主要结论

本文在函数型数据分析的视角下，以非负矩阵分解的框架对一元、多元函数型数据展开聚类方法研究，针对实际应用中数据存在噪声、缺失值导致不完备等问题，本文主要提出了三项工作：

1. 在现有函数型聚类算法的基础上，提出基于双随机图正则化矩阵分解的函数型聚类算法(BSMFFC)，结合了图拉普拉斯和双随机矩阵，通过双随机矩阵动态更新图学习，充分利用数据的局部流形结构信息，构建的模型算法更适用于混合型函数型数据，尊重数据的空间几何机构。此外，基于增广拉格朗日乘子法，给出更新公式，并研究算法的时间复杂度。在随机模拟数据 I、Growth 数据、TIMIT 数据上的模拟实验结果表明：无论是 ACC、NMI、PUR 还是 RI，BSMFFC 算法的聚类表现都优于其他算法。最后，对北京市 NO2 小时浓度进行聚类得到的结果也有一定的实际应用参考价值。

2. 构建基于$l_{2,1}$范数的鲁棒函数型聚类算法(FRMNMF)，为了规避利用 Frobenius 范数定义损失函数所带来的影响，尝试用$l_{2,1}$来定义损失函数，衡量矩阵分解质量，同时结合流形学习，构建更具有鲁棒性的函数型聚类一步算法，以此提高聚类算法的性能。此外，通过使用交替迭代的优化算法，推导目标函数更新公式，并且对算法的计算时间复杂度进行了深入剖析。随后，在随机模拟数据 II、Growth 数据、CanadianWeather 数据、FatSpetrum 数据上的模拟实验可以表明，FRMNMF 算法的聚类效果和超参数和构造初始图中 KNN 算法中K的取值有关。且从聚类评价结果来看，不论是 PUR、ACC、RI，所提出的 FRMNMF 算法是优于其他算法的，说明利用$l_{2,1}$范数构造损失函数可以提高聚类性能。最后，在城镇居民人均可支配收入数据上开展聚类分析，以此验证算法的可行性、合理性及实际应用价值。

3. 构建用于不完整多视角函数型数据的聚类算法(AIMFC)，基于非负矩阵分解的研究框架，提出了一种将函数型矩阵填充和多视角聚类相结合的方法，以构

建一种同时具有数据插补和聚类并行功能的聚类一步法，同时通过定义每个视角的权重因子，自适应地为每个视角分配适当的权重，充分利用每一视角的数据信息；其次，引入乘法更新规则，给出迭代优化公式，分析该算法的时间复杂度。通过在随机模拟数据 III 上构造缺失值，并利用所提出的算法进行聚类分析，结果表明，AIMFC 算法能够有效处理带有缺失的数据，并且在惩罚参数的稳健性上具有一定优势。实际应用中，AIMFC 算法在处理北京市大气污染物小时浓度数据的聚类结果中便可发现，其具备对全变量均存在缺失的情形进行处理的能力，同时也可处理存在大量数据缺失的情形。

6.2 展望

尽管本文中所提出的三种算法在聚类任务中已经取得了较好的效果，但是仍然存在改进的空间：

首先，对于一元数据而言，在实际应用中，部分数据会存在少量的标签数据，因此，忽略标签信息导致聚类效果有待进一步提升，大量无标签的数据和带有标签的少量数据一起学习的机器学习方法被称为半监督学习，在函数型聚类方法中进一步引入半监督学习，有利于提高算法的聚类效果；其次，大多数多视角聚类方法仍然面临两个主要问题：一是数据存在缺失值，二是数据中包含噪声。因此，如何同时处理这两个问题均存在的多视角函数型数据聚类是接下来进一步要研究的重点内容。

在未来的工作中，我们可以尝试其他图学习方法，做一些方法改进。大多数方法只考虑一阶关系来构造相似矩阵，不能充分挖掘邻域信息，我们可考虑特征之间的高阶关系，利用一阶和二阶关系来获得最优邻域几何结构。此外，尝试将单视角函数型聚类算法推广到解决多视角聚类问题，同时可以把一些先验信息和标签数据整合到算法中，设计一种新的半监督函数型聚类算法，以上问题我们将在未来继续研究。

参考文献

[1] Bouveyron C, Brunet-Saumard C. Model-Based Clustering of High-Dimensional Data: A review[J]. Computational Statistics and Data Analysis. 2014, 71(1): 52-78.

[2] Jacques J, Preda C. Model-based clustering for multivariate functional data[J]. Computational Statistics & Data Analysis. 2014, 71(3): 92-106.

[3] Chiou J, Li P. Functional Clustering and Identifying Substructures of Longitudinal Data[J]. Journal of the Royal Statistical Society：Series B (Statistical Methodology). 2007, 69(4): 679-699.

[4] 王德青，朱建平，刘晓葳，等. 函数型数据聚类分析研究综述与展望[J]. 数理统计与管理. 2018, 37(01): 51-63.

[5] Jacques J, Preda C. Funclust: A curves clustering method using functional random variables density approximation[J]. Neurocomputing. 2013, 112(10): 164-171.

[6] Abraham C, Cornillon P A, Matzner Løber E, et al. Unsupervised Curve Clustering using B-Splines[J]. Scandinavian Journal of Statistics. 2003, 30(3): 581-595.

[7] 黄恒君. 基于B-样条基底展开的曲线聚类方法[J]. 统计与信息论坛. 2013, 28(09): 3-8.

[8] 许腾腾，王瑞，黄恒君. 一种加入类间因素的曲线聚类算法[J]. 智能系统学报. 2019, 14(02): 362-368.

[9] Yamamoto M, Hwang H. Dimension-Reduced Clustering of Functional Data via Subspace Separation[J]. Journal of Classification. 2017, 34(2): 294-326.

[10] Yamamoto M, Terada Y. Functional Factorial K-means Analysis[J]. Computational Statistics & Data Analysis. 2014, 79(4): 133-148.

[11] Coffey N, Hinde J, Holian E. Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data[J]. Computational Statistics & Data Analysis. 2014, 71(3): 14-29.

[12] Giacofci M, Lambert-Lacroix S, Marot G, et al. Wavelet-Based Clustering for Mixed-Effects Functional Models in High Dimension[J]. Biometrics. 2013, 69(1): 31-40.

[13] 黄恒君，高海燕，张梦瑶. 函数型聚类分析:基于距离的一步法框架[J]. 数理统计与管理. 2019, 38(06): 986-995.

[14] 高海燕，黄恒君，王宇辰. 基于非负矩阵分解的函数型聚类算法[J]. 统计研究. 2020, 37(08): 91-103.

[15] Lee D, Seung H. Learning the Parts of Objects by Non-Negative Matrix Factorization[J]. Nature. 1999, 6755(401): 788-791.

[16] Ding C, He X, Simon H D, et al. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering[C]. United States: 2005.

[17] Li S Z, Xin W H, Hong J Z, et al. Learning spatially localized, parts-based representation[C]. IEEE: 2001.

[18] Xiong Z, Zang Y, Jiang X, et al. Document Clustering with an Augmented Nonnegative Matrix Factorization Model[C]. Springer, 2014.

[19] Tao D, Tao D, Liu W, et al. Large Sparse Cone Non-negative Matrix Factorization for Image Annotation[J]. ACM Transactions on Intelligent Systems and Technology. 2017, 8(3): 1-21.

[20] Ding C, Li T, Jordan M I. Convex and Semi-Nonnegative Matrix Factorizations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010, 32(1): 45-55.

[21] Kong D, Ding C, Huang H. Robust nonnegative matrix factorization using L21-norm[C]. Glasgow, Scotland, UK: Association for Computing Machinery, 2011.

[22] Cai D, He X, Han J, et al. Graph Regularized Nonnegative Matrix Factorization for Data Representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011, 33(8): 1548-1560.

[23] Huang J, Nie F, Huang H, et al. Robust Manifold Nonnegative Matrix Factorization[J]. ACM Transactions on Knowledge Discovery from Data. 2014, 8(3): 1-21.

[24] Huang S, Xu Z, Wang F. Nonnegative matrix factorization with adaptive neighbors[C]. 2017.

[25] Wang Q, He X, Jiang X, et al. Robust Bi-Stochastic Graph Regularized Matrix Factorization for Data Clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022, 44(1): 390-403.

[26] 董文婷，尹学松，余节约，等. 鲁棒结构正则化非负矩阵分解[J]. 计算机应用研究. 2023, 40(03): 794-799.

[27] Haifeng L, Zhaohui W, Xuelong L, et al. Constrained Nonnegative Matrix Factorization for Image Representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012, 34(7): 1299-1311.

[28] Yi Y, Lai S, Wang W, et al. SDNMF: Semisupervised discriminative nonnegative matrix factorization for feature learning[J]. International Journal of Intelligent Systems. 2022, 37: 11547-11581.

[29] 陈君航，杨祖元，刘名扬，等. 基于正交约束的广义可分离非负矩阵分解算法[J]. 计算机工程. 2023, 49(08): 46-53.

[30] 侯兴荣，彭冲. 基于局部相似性学习的鲁棒非负矩阵分解[J]. 数据采集与处理. 2023, 38(05): 1125-1141.

[31] 高海燕，刘万金，黄恒君. 鲁棒自适应对称非负矩阵分解聚类算法[J]. 计算机应用研究. 2023, 40(04): 1024-1029.

[32] Bickel S, Scheffer T. Multi-view clustering[C]. 2004.

[33] Kumar A, Iii H D. A Co-training Approach for Multi-view Spectral Clustering[C]. In Proceedings of the International Conference on Machine Learning, 2011.

[34] Ye Y, Liu X, Yin J, et al. Co-regularized kernel k-means for multi-view clustering[Z]. 20161583-1588.

[35] Nie F, Shi S, Li X. Auto-weighted Multi-view Co-clustering via Fast Matrix Factorization[J]. Pattern Recognition. 2020, 102: 107207.

[36] Gönen M, Margolin A. Localized Data Fusion for Kernel k-Means Clustering with Application to Cancer Biology[Z]. 2014: 2, 1305-1313.

[37] Zhu X, Liu X, Li M, et al. Localized incomplete multiple kernel k-means[C].

Stockholm, Sweden: 2018.

[38] Wang S, Liu X, Zhu E, et al. Multi-view Clustering via Late Fusion Alignment Maximization[C]. 2019.

[39] Wang Y, Zhang W, Wu L, et al. Iterative Views Agreement: An Iterative Low-Rank based Structured Optimization Method to Multi-View Spectral Clustering[C]. 2016.

[40] Saha M. A Graph Based Approach to Multiview Clustering[C]. 2013.

[41] Nie F, Li J, Li X. Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semi-supervised classification[C]. 2016.

[42] Cai Y, Jiao Y, Zhuge W, et al. Partial multi-view spectral clustering[J]. Neurocomputing. 2018, 311: 316-324.

[43] 夏冬雪，杨燕，王浩，等. 基于邻域多核学习的后融合多视图聚类算法[J]. 计算机研究与发展. 2020, 57(08): 1627-1638.

[44] Chaudhuri K, Kakade S M, Livescu K, et al. Multi-view clustering via canonical correlation analysis[C]. 2009.

[45] Guo Y. Convex subspace representation learning from multi-view data[C]. 2013.

[46] Fan Y, He R, Hu B. Global and local consistent multi-view subspace clustering[C]. 2015.

[47] Wang D, Yin Q, He R, et al. Multi-view Clustering via Structured Low-rank Representation[C]. 2015.

[48] 吴峰，刘改，刘诗仪. 基于图信息的自监督多视角子空间聚类[J]. 计算机系统应用. 2022, 31(05): 377-381.

[49] Liu J, Wang C, Gao J, et al. Multi-View Clustering via Joint Nonnegative Matrix Factorization[Z]. 2013252-260.

[50] Zhang X, Zhao L, Zong L, et al. Multi-view Clustering via Multi-manifold Regularized Nonnegative Matrix Factorization[C]. 2014.

[51] 刘正，张国印，陈志远. 基于特征加权和非负矩阵分解的多视角聚类算法[J]. 电子学报. 2016, 44(03): 535-540.

[52] 宗林林，张宪超，赵乾利，等. 一种多流形正则化的多视图非负矩阵分解算

法[J]. 南京大学学报(自然科学). 2017, 53(03): 557-568.

[53] Huang S, Ren Y, Xu Z. Robust multi-view data clustering with multi-view capped-norm K-means[J]. Neurocomputing. 2018, 311: 197-208.

[54] 连佳琪，王毅刚，储志伟，等. 结构正则化多视图非负矩阵分解[J]. 计算机应用研究. 2022, 39(10): 3033-3038.

[55] 林虹燕，杜元花，周楠，等. 基于多视角自适应图正则的非负矩阵分解聚类[J]. 成都信息工程大学学报. 2023, 38(05): 526-534.

[56] Li C, Che H, Leung M, et al. Robust multi-view non-negative matrix factorization with adaptive graph and diversity constraints[J]. Information Sciences. 2023, 634: 587-607.

[57] Schmutz A, Jacques J, Bouveyron C, et al. Clustering multivariate functional data in group-specific functional subspaces[J]. Computational Statistics. 2020, 35(3): 1101-1131.

[58] Ieva F, Pigoli D, Paganoni A, et al. Multivariate Functional Clustering for the Morphological Analysis of ECG curves[J]. Journal of the Royal Statistical Society. 2013, 62(3): 401-418.

[59] 姚晓红，高海燕，吕家奇，等. 一种基于多视角学习的多元函数型聚类方法[J]. 数理统计与管理. 2022, 41(04): 689-702.

[60] Jia Y, Liu H, Hou J, et al. Semisupervised Adaptive Symmetric Non-Negative Matrix Factorization[J]. IEEE Transactions on Cybernetics. 2021, 51(5): 2550-2562.

[61] Sheng Y, Wang M, Wu T, et al. Adaptive local learning regularized nonnegative matrix factorization for data clustering[J]. Applied Intelligence. 2019, 49(6): 2151-2168.

[62] 刘威，邓秀勤，刘冬冬，等. 基于约束图正则的块稀疏对称非负矩阵分解[J]. 计算机科学. 2023, 50(07): 89-97.

[63] Wang F, Li P, König A C, et al. Improving clustering by learning a bi-stochastic data similarity matrix[J]. Knowledge and Information Systems. 2012, 32(2): 351-382.

[64] von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing.

2007, 17(4): 395-416.

[65] Lin Z, Chen M, Ma Y. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices[J]. Mathematical Programming. 2010, 9.

[66] Boyd S P, Vandenberghe L. Convex Optimization[M]. 2004: 1859.

[67] Yu, Shi. Multiclass spectral clustering[Z]. IEEE Computer Society: 2003313-319.

[68] Antonevich A. Functional Operators[M]. Linear Functional Equations. Operator Approach, Antonevich A, Basel:Birkhäuser Basel, 1996, 17-48.

[69] Schölkopf B, Platt J, Hofmann T. Doubly Stochastic Normalization for Spectral Clustering[M]. MIT Press, 2007, 1569-1576.

[70] Abraham C, Cirad-For Et U. Unsupervised Curve Clustering using B-Splines[J]. Scandinavian Journal of Statistics. 2001, 30(3): 581-595.

[71] Lee D, Seung H. Algorithms for Non-negative Matrix Factorization[J]. Adv. Neural Inform. Process. Syst. 2001, 13.

[72] Ding C H Q, Li T, Jordan M I. Convex and Semi-Nonnegative Matrix Factorizations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010, 32(1): 45-55.

[73] Kalivas J H. Two data sets of near infrared spectra[J]. Chemometrics and Intelligent Laboratory Systems. 1997, 37(2): 255-259.

[74] Ferraty F, Vieu P. Nonparametric Functional Data Analysis[M]. Springer New York, 2006.

[75] 武森，冯小东，单志广. 基于不完备数据聚类的缺失数据填补方法[J]. 计算机学报. 2012, 35(08): 1726-1738.

[76] Tao H, Hou C, Yi D, et al. Joint Embedding Learning and Low-Rank Approximation: A Framework for Incomplete Multiview Learning[J]. IEEE Transactions on Cybernetics. 2021, 51(3): 1690-1703.

[77] Liu X, Li M, Tang C, et al. Efficient and Effective Regularized Incomplete Multi-View Clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021, 43(8): 2634-2646.

[78] Wen J, Zhang Z, Xu Y, et al. Incomplete Multi-view Clustering via Graph Regularized Matrix Factorization[C]. Cham: Springer International Publishing. 2019：593-608, 2019.

[79] Kim Y, Choi S. Weighted nonnegative matrix factorization[C]. 2009.

攻读硕士学位期间承担的科研任务及主要成果

发表的论文：

刘万金,赵芳芳.自监督对称非负矩阵在GDP聚类分析中的应用[J].甘肃科技纵横,2022,51(07):69-73.

参与科研项目：

甘肃教育科技创新项目：大数据背景下基于非负矩阵分解的多视角聚类方法及应用研究(2020A-059)，2020.6—2022.9，已结项。

竞赛获奖：

“函数型数据视角下我国 GDP 的聚类分析及预测研究”荣获第五届全国应用统计专业学位研究生案例大赛全国三等奖，2022 年 8 月。

“空气污染物防治区域划分研究——以重庆市 $PM_{2.5}$ 为例”荣获 2022 年（第八届）全国大学生统计建模大赛省级三等奖，2022 年 8 月。

致 谢

写到这里，我的三年研究生生活就要进入尾声了。时间真的过得很快，翻开三年记忆的长卷，有胜利和成功的微笑，也有失败和不甘的泪水。让我从一个幼稚、懦弱的孩子逐渐蜕变成一个敢于承担的大人。这三年的回忆将永远封存在我的脑海里，即使以后翻开也热泪盈眶。

玉壶存冰心，朱笔写师魂。在此，首先要特别地感激我的导师高海燕老师，在论文的撰写过程中，从开题报告、中期检查直到论文完稿都悉心指导。感谢老师对我的包容与教导，感谢她对我的信任和用心，让我能圆满地给这三年回忆画上句号。

其次，非常感谢我的父母，我的家人。尽管他们只是普通的农民，却总是竭尽全力给我最好的。感谢他们在我选择读研的无条件支持，如果没有他们的付出，我也无法得到这些收获。感谢他们的信任，让我去改变既定的命运，终究，我只是站在家人肩膀上，看到了他们从未见过的世界。

感谢自己，感谢自己的坚持与努力。感谢为了实现梦想不断奋斗的自己，感谢相信读书是脱离贫穷出路的自己，感谢疯狂为考研而努力的 20 岁的自己。虽然在黑夜里独自前行，但一定会拥有星辰大海，希望未来的自己，继续保持对生活的热爱。还有那些未曾谋面却帮助过我的人，感谢他们成为我黑暗疲惫生活里的星光，对我而言他们是特殊的存在，让我变得善良、温柔而强大。

感谢指导过我的老师们，感谢我的室友，愿此去有繁花似锦，理想不伟大，只愿年过半百，归来仍是少年。