摘要

为研究我国中老年人群慢性病共病模式，探索慢性病间的关联子群结构，本文采用中国健康与养老追踪调查（China health and retirement longitudinal study, CHARLS）2013、2015、2018 年三年的数据，包括中国 25 个省份的 45 岁以上中老年人群患 10 种慢性病情况。然而，由于不同来源的数据集在生活方式、文化习俗等存在差异，多源数据异质性不可避免。若忽略数据之间的异质结构，直接合成所有数据集建模，会导致估计有偏、聚类不准确等问题。因此，如何有效识别多源数据的异质结构、整合同质数据，推断慢性病间的关联性是本文的主要研究目的。

为此，本文提出一种基于矩阵低秩分解的双重亚组（聚类）学习方法（**L**ow **R**ank assisted **D**ouble-task **S**ubgroup **A**nalysis, **LRDSA**）。首先，将不同省份以及不同慢性病下的参数排列为矩阵。进一步，采用低秩分解，提取出矩阵在省份维度以及慢性病维度的特征向量。最后，基于 Fused Lasso 融合函数，学习两个维度特征向量的聚类结构，实现多源数据的异同结构识别，整合同质数据，学习慢性病间的关联子群结构。

较传统的慢性病关联性学习，本文采用基于数据驱动的 Fused Lasso 融合聚类，不需提前给定聚类数目，在实际数据中具有更强的实用性。此外，本文数值模拟以及实际数据分析结果表明，提出的双重亚组分析方法具有更高的参数估计精度以及聚类准确度。特别地，针对 CHARLS 数据，本文发现，慢性病中心脏病与血脂异常属于一类，高血压与关节炎或风湿病属于一类，哮喘和肝脏疾病属于一类；中老年人患高血压和关节炎或风湿病这一类慢性病的风险显著高于其他类慢性病，患肝脏疾病、哮喘这一类慢性病的风险低于其他类慢性病。省份维度，广东、山西、福建聚类为一类，黑龙江、内蒙、甘肃聚类为一类；广东、山西、福建这一类省份患慢性病风险显著低于其他类省份，黑龙江、内蒙古、甘肃这一类省份患慢性病风险显著高于其

他类省份，这些高风险区域基本处于我国西北区域。

关键词：慢性病关联分析；多源数据；亚组分析

ABSTRACT

In order to study the comorbidity pattern of chronic diseases in Chinese middle-aged and elderly population and explore the structure of the associated subgroups among chronic diseases, this paper adopts the data of the China health and retirement longitudinal study (CHARLS) in 2013, 2015 and 2018, including 10 chronic diseases suffered by middle-aged and elderly people over 45 years old in 25 provinces of China. However, due to the differences in lifestyle, cultural customs and other aspects of data sets from different sources, multi-source data heterogeneity is inevitable. If the heterogeneous structure between data is ignored and all data sets are directly synthesized for modeling, problems such as biased estimation and inaccurate clustering will be caused. Therefore, how to effectively identify the heterogeneous structure of multi-source data and integrate homogeneous data to infer the correlation between chronic diseases is the main research purpose of this paper.

Therefore, this paper proposes a new method named Low Rank assisted Double-task Subgroup Analysis (LRDSA) based on matrix low-rank decomposition. First, the coefficients of different provinces and chronic diseases were arranged as a matrix. Furthermore, low-rank decomposition was used to extract the feature vectors of the matrix in the province dimension and the chronic disease dimension. Finally, on the basis of Fused Lasso function, we learn fused structure of two-dimensional feature vectors to recognize the similarities and differences of multi-source data, to realize the integration of multi-source data and the association analysis of chronic disease.

Compared with the traditional chronic disease association learning, the data-driven Fused Lasso method was used in this paper, which does not require the number of clusters to be specified in advance, so it is more useful in the actual data. In addition, numerical simulation and actual data analysis results show that the proposed dual subgroup analysis has higher parameter estimation accuracy and clustering accuracy. In particular, according to CHARLS data, this paper found that among c

hronic diseases, heart disease and dyslipidemia belong to the same category, hypert ension and arthritis or rheumatism belong to the same category, asthma and liver di sease belong to the same category. The risk of hypertension and arthritis or rheuma tism are significantly higher than other chronic diseases, and the risk of liver disea se and asthma is lower than other chronic diseases. In terms of provinces, Guangdo ng, Shanxi and Fujian are grouped into one category, while Heilongjiang, Inner M ongolia and Gansu are grouped into one category. The risk of chronic disease in G uangdong, Shanxi and Fujian is significantly lower than that in other provinces, w hile that in Heilongjiang, Inner Mongolia and Gansu provinces is significantly hig her than that in other provinces. These high-risk regions are basically in the northw est of China.

Keywords: Chronic Disease Association Analysis; Multi-source Data; Subgro up Analysis

目录

1.绪论

1.1 研究背景

慢性病即慢性非传染性疾病，并不指某一种特定的疾病，是对一类患病周期长、治疗难、易反复的疾病的概括性称呼。主要的慢性病包括心脑血管疾病、糖尿病、癌症、慢性呼吸系统疾病等。2022 年，世界卫生组织发布报告显示，心脏病、癌症和糖尿病等非传染性疾病已经代替了传染病，成为“全球首位杀手”，约 75%的死因是非传染性疾病。中老年人群体更是慢性病患病的高风险人群，有学者[1]研究发现同时患多种慢性病的现象在中老年人群体中十分普遍，60 岁以上的中老年人中超过一半的人患有两种及以上的慢性病，并且伴随着年龄的增长，发病率会越来越高，各种慢性病显然已成为中老年人健康的一大威胁。

某些慢性病之间往往是共存的，患有一种慢性疾病可能会增加其他慢性疾病的风险，控制一种慢性疾病的恶化，也会减少其他慢性疾病的患病风险。这些慢性病互相是对方的风险因子，产生着多米诺骨牌效应。已有研究表明，肥胖会增加患 2 型糖尿病或者高血压的风险[2]，2 型糖尿病或者高血压会破坏血管壁，导致动脉粥样硬化[3]，进而诱发冠心病或中风[4]等心脑血管疾病。因此发现慢性病之间的关联性是缓解和解决慢性非传染性疾病发病的有效手段，同时也成为实现健康老龄化的必经之路。

国内外关于中老年人慢性病关联性的相关研究已有很多，很多学者运用统计学、机器学习等方法建立模型，探索慢性病的关联结构，得到多条有用的关联规则。但是很少有研究是基于多源数据环境进行分析，多源数据由于收集数据的来源不同，往往呈现异构性。数据的异构性是指数据的内在结构或者外在表现形式不同。整合不同来源数据集带来的好处是毋庸置疑的，越多的数据会提供越多的信息，不同结构的数据之间可以互相借用信息，会大幅提高参数估计的统计性质。但是对这样的数据集建模，也面临很多挑战。

如果简单粗暴的把所有样本合并，无视不同来源数据之间的异构性，会得到有偏、不科学的结论；反之如果认为不同来源的数据完全异构，对每个来源的数据单独建模，会使得模型待估参数增多，模型过于冗杂，同时也忽视不同来源数据之间的关联性。如何准确识别多源异构数据的关联性和差异性，是多源数据建模面临的重要问题。在研究慢性病之间关联性时，也会面临多源数据，例如来自全国各个省份的中国健康与养老追踪调查（China Health and Retirement Longitudinal Study），即 CHARLS。CHARLS 是由北京大学主持的，针对全国 45 岁及以上的中老年人，覆盖全国 28 个省的高质量微观数据。我国幅员辽阔，地区间自然条件迥异，不同地区从经济、社会以及城镇化水平等方面的差异不容忽视。另外，不同地区自然资源、气候条件、文化底蕴、宗教信仰等因素形成了各具风格的饮食习惯，构成了丰富多彩、各具特色的中国饮食文化。虽然交通工具的发展使得不同地区的物资可以共享、人口能够广泛流动，各个地区的饮食口味互相包容，差异不断缩小[5]。但是在省份这样大面积的空间范围上，各个省份仍然有独特的饮食口味。因此，在现代社会医学模式下，不同地区中老年人面临的各种可能风险因子有很大差异，进而导致慢性病患病率也会不同。如果忽略地区的异质性，只考虑慢性病，会得出不全面、不科学的结论。已有研究在做中老年人慢性病分析时，大多只是将地区作为解释变量引入模型，很少有将来自不同地区数据看作多个数据集来进行整合分析，同时考虑不同数据集的同质、异质结构，因此如何发现不同来源数据集中老年人慢性病关联子群结构是一个重要的问题。

1.2 研究目的及意义

国内关于中老年人慢性病的研究起步较晚，研究主题和研究的内容相对而言还是比较局限。人口老龄化并不是单单指年龄变老这一过程，它还包含着人口与社会的不断互动。影响中老年人慢性病因素是复杂的、多方面的，除了基因、生物学特征外，经济发展、社会保障、饮食习惯、居住环境等都会影响着中老年人的生活习惯和方式，从而改变其患慢性病的风险大小。并且我国地大物博，不同省份之间不论是饮食习惯还是经济发展水平都差异较大，因此有必要对其进行较小颗粒度的空间分析。如果对所有的省份一起分

析，忽略彼此的异质性，会得出不科学、一刀切的结论；如果人为的将省份依照地理位置分成几个区域，可能会得出不准确的结论；如果把所有省份都分割开来，认为所有省份都是有差异，可能会损失样本，忽略省份之间的相似性。此外不同慢性病之间往往也不是完全独立的，是存在关联的，如果不细分具体的慢性病，只是分析是否患慢性病，会无视慢性病之间的异质性，得出不科学甚至是错误的结论；如果分开探讨所有的慢性病，又会忽略慢性病之间的相似性，不利于对慢性病进行综合的预防和治疗。

因此，在我国特有的社会经济文化背景下，充分理解中老年人健康涵义，借鉴吸收国际国内相关研究方法和成果，着眼于 10 种不同慢性病、25 个不同省份之间的相似性与差异性，发现以省份为单位的中老年人患慢性病的聚集性以及不平衡性，以及不同慢性病之间的聚类结构。深入刻画中老年人患慢性病的时空分布特征以及慢性病之间的关联性，为慢性病早期的发现、预防以及治疗提供借鉴，为疾病的区域化防控能够提供支撑，进而促进健康老龄化目标的实现。

本文的研究意义将从应用和方法两个方面来说明：

（1）应用方面

第一，进行较小颗粒度的空间分析，发现不同空间单位中老年人患慢性病的情况，为疾病的区域化防控能够提供指导，协调配置基本公共设施和服务，缩小区域健康差异。

第二，细分 10 种不同的慢性病，全面、综合、系统的分析中老年人慢性病情况。发现不同慢性病的关联子群结构，为慢性病早期的发现、预防以及治疗提供借鉴，有助于慢性病的综合防治。

第三，考虑慢性病维度和省份维度的双重聚类，区别于以往单一视角的研究，一定程度上是对慢性病研究的补充。

（2）方法方面

本文提出一种基于矩阵低秩分解的双重亚组（聚类）学习方法 **LRDSA**。首先，将不同省份以及不同慢性病下的参数排列为二维矩阵。进一步，采用低秩分解，提取出矩阵在省份维度以及慢性病维度的特征向量。最后，结合 Fused Lasso 融合函数以及 ADMM 算法，学习两个维度特征向量的聚类结构，实现多源数据的整合以及慢性病关联性分析。

与异质和同质模型相比，本文提出的双重亚组分析具有更高的参数估计精度以及聚类准确度。此外，本文采用基于数据驱动的 Fused Lasso 融合聚类，不需要提前给定聚类数目，在实际数据中具有更强的实用性，丰富了多源数据异同结构识别的研究方法。

1.3 研究内容及章节安排

本文重点介绍了 Lasso 模型、Fused Lasso 模型以及 ADMM 算法的相关理论，基于多源数据异质性不可避免的背景，提出一种基于矩阵低秩分解的双重亚组（聚类）学习方法 **LRDSA**。并通过数值模拟验证此方法的有效性，将其与同质、异质模型进行对比，结果表明此方法具有更高的参数估计稳定性和聚类精度。在我国已处于并将长期处于深度老龄化以及中国各区域生活方式、文化习俗、饮食习惯存在不可忽视的差异的现实背景下，将此方法应用于中老年人慢性病关联性研究。根据实证结果，得出研究结论以及提出相应的政策建议，寻求中老年人慢性病的主要着力点，提高中老年人健康预期寿命。

第一章为绪论。本章节在慢性病已经成为人类尤其是中老年人群体健康的一大威胁、各种不同的慢性病之间存在广泛的关联性以及实证数据中的多源异构性不可忽视的背景下，阐述了本文的研究目的及意义、本文的研究内容与章节安排以及创新和不足之处。

第二章为文献综述。对国内外关于“慢性病空间分布关联性”、“慢性病共病模式研究”、“整合分析研究”、“亚组分析研究”等与本论文研究问题密切相关的国内外研究成果进行梳理，并进行简要叙述评论，提出本文的研究相比于以往研究的扩展之处。

第三章为相关模型以及算法介绍。这章主要是介绍与本文提出的 **LRDSA** 方法相关的模型和算法，为下一章介绍本文提出的 **LRDSA** 方法做准备。首先，介绍了 Logistic 模型、Lasso 模型、Fused Lasso 模型，然后介绍了交替乘子算法（ADMM）相关的理论以及计算步骤。

第四章为基于矩阵低秩分解的双重亚组学习及数值模拟。首先介绍了 **LRDSA** 学习方法和求解算法，并给出了详细的实现步骤。然后通过多次不

同情况的数值模拟探讨 **LRDSA** 学习方法对于识别多源数据环境下，两个参数相似对和不相似对的聚类表现以及参数估计的准确性，并将其与同质模型和异质模型进行参数估计、聚类效果对比。

第五章为 CHARLS 实证数据分析。将本文提出的 **LRDSA** 学习方法应用于包含 25 个省份和 10 种慢性病的中老年人健康调查数据，首先介绍模型的解释变量、被解释变量以及控制变量的赋值，其次在描述性分析中老年人各省份慢性病患病情况以及慢性病共病情况的基础上，实证研究中老年人慢性病的空间分布特征以及慢性病的关联性，得出相应的结论。

第六章为本文的研究结论与政策建议。对全文所得出的结论进行总结，并对中老年人慢性病防治提出相应的、合理的政策建议。

1.4 创新和不足之处

1.4.1 论文创新点

（1）在多源数据环境下考虑慢性病关联子群，区别于以往慢性病或者省份单一视角的研究，一定程度上是对慢性病研究的补充。

（2）提出由数据驱动的基于矩阵低秩分解的双重亚组（聚类）学习方法 **LRDSA**，给出了方法的详细步骤。并且针对直接优化目标函数比较困难的问题，详细给出了可行的求解算法。此方法可以同时实现两个维度的聚类，并且和同质、异质模型相比，估计参数更加准确。

（3）将 **LRDSA** 应用到来自多个省份的中老年人慢性病关联分析中，实现多种慢性病和省份的同时聚类，发现不同空间单位中老年人患慢性病的情况以及不同慢性病和不同省份的聚类结构，帮助缩小区域健康差异、综合治疗慢性疾病。

1.4.2 论文不足之处

由于自身能力水平和调查数据等方面的限制，本文对于多源数据环境下中老年人慢性病关联子群的研究并不完善，仍有较多不足之处：

（1）本文将省份视为最小的研究单位，然而，在现实情况下，在各个地区，不同的城市、社区、县、镇，其环境特点和人文特点与中老年人的慢性病患病情况有着更加密切的关系，本文未能对地区进行更加细致的研究。

（2）研究纳入的 10 种慢性病患病信息均为受访者自我报告，自我报告的结果与真实情况可能存在偏差。对危险因素进行干预是慢性病防控的重要内容，对危险因素的空间分布特征进行分析可以解释各因素对不同空间单位疾病发病、患病等的影响，从而为疾病的区域化防控提供指导。但由于慢性病的病因复杂，其危险因素目前尚不明确，因此本研究未能对这部分内容进行分析。

（3）虽然 CHARLS 数据库质量很高，包含的指标已经十分齐全和丰富，为本文的研究带来了很多便利。但是也会存在很多因为涉及受访者隐私而未被纳入的指标。CHARLS 中生物学标志物的数据较少，而这些指标毫无疑问对中老年人慢性病患病情况有显著影响，如果把这些变量纳入到模型中，得到的结论会更加可靠。

（4）本文提出的 **LRDSA** 学习方法，虽然是由数据驱动，不需要提前给出聚类的数目，但是得到的亚组结构其医学内涵还有待进一步的研究，希望以后该方法能够有效的运用于其他更多多源数据中。

2.文献综述

本部分主要从慢性病空间分布关联性研究、慢性病共病模式研究、亚组分析研究、整合分析研究进行国内外文献梳理，并从以前的研究中发掘本文的研究扩展之处。

2.1 慢性病关联性研究

2.1.1 慢性病空间分布关联性研究

80%的流行病学的研究数据都具有空间属性[6]，发现疾病流行规律的前提是认识并且利用好数据的空间属性。地理位置本身暗含丰富的信息，不同区域其气候、自然资源、经济发展、饮食口味、生活习惯等等都有着显著的差异，这些因素都可能是慢性疾病的危险因素，改变慢性病患病率情况。国内外已有大量学者研究慢性病空间分布特征、空间聚集性、影响因素异质性等等[7]，主要采用的方法是疾病地区等级分布制图、空间自相关分析（spatial autocorrelation analysis）、空间回归分析、Logistic 回归模型等。

（1）疾病地区等级分布地图

绘制疾病地区等级分布图是流行病学空间数据可视化的常用手段，通过对某区域以一定地理单位标准化处理后的患病率、死亡率等指标用分级地图表达，能够简单直观的描述疾病的空间分布情况。邓群[8]根据第三次全国抽样调查数据，以省份为单位绘制高血压患病等级分布图，发现内蒙古、天津、北京、河北等西北和东北地区是高血压的高发区。

（2）空间自相关分析（spatial autocorrelation analysis）

空间自相关分析是空间数据分析中的常用统计技术，用于研究空间中某位置的观察值和距离近的观察值是否存在相关性以及相关性程度[9]。空间自

相关分为三种情况：第一种是正相关，即相邻区域的属性值相同。第二种是负相关，即相邻区域的属性值不相同。第三种是无相关，即变量属性值无明显空间分布特征。随着空间统计技术的发展和对慢性病相关地理因素的认识，空间自相关分析广泛运用于慢性病的空间分析中。

Freyssenge 等[10]应用空间自相关分析对法国罗纳省中风率进行了空间分布研究，结果表明地区差异性十分明显，尤其在该省北部地区和西部地区都市圈。谭利明等[11]对 2015 年中国中老年人高血压患病空间分布进行空间自相关分析，局部空间分析研究结果表明，中国东北部和东部是高血压患病的高发地区，应该重点防控东部和东北部。Guo 等[12]对 2015 年中国老年人共病模式进行空间相关分析，“血脂异常、高血压、关节炎或风湿病/心脏病”共病模式存在全局空间正自相关。王浩等[13]对 2018 年中老年人慢性病共病空间分布规律进行空间分析，全局空间自相关分析结果表明，中老年人共病患病率呈正相关，患病率高的地区周围患病风险也较大，患病率低的地区周围患病风险也较小。局部空间自相关结果表明，青海、甘肃两个省份共病患病率高，其周围新疆、四川、内蒙患病率也较高，表现为高—高聚集；福建患病率风险低，周围广东、江西、浙江患病率也较低，表现为低—低聚集。程文炜等[14]研究中老年人糖尿病的空间分布，局部空间分析结果表明，我国北部是糖尿病患病的聚集区域，需要重点防控。

（3）空间回归分析

空间数据除了可能具有相关性以外，往往还具有异质性。如果对不同空间数据直接整合，采取全局回归模型来分析，容易得出不科学的结论。全局回归模型暗含不同空间单位具有完全同质性假定，回归系数估计值只能代表变量对整个空间的平均影响。基于局部回归和变参研究，1996 年 Brunsdon 等[15]首次提出了地理加权回归（geographical weighted regression,GWR）。GWR 通过在回归模型中引入空间关系，能够很好的反映解释变量的局部空间关系和空间异质性。

谭利明等[11]首先在省级空间数据对中老年人高血压患病风险因素进行全局回归分析，得出男性、年龄和超重是高血压患病的主要风险因素。并且还对全局分析得到的危险因素进行 GWR 分析，结果表明高龄、男性、超重这些高血压患病的危险因素存在空间分布差异，北部的高龄老年人、中部的男

性、东北部的超重老年人患高血压的风险更大。程文炜等[14]对糖尿病患病风险因素空间分析，结果显示高学历占比、体重超重率对糖尿病患病有显著影响并且存在地理分布差异，应该对不同地区采取不同疾病防治措施。

（4）Logistic 回归模型

Logistic 回归模型的因变量为分类变量，回归模型通俗易懂，由于该方法不要求数据的分布以及自变量的类型，常常被用于医学中患病情况的研究。近些年来，我国老年人高血压患病率不断攀升，陈娇[16]采用多元 Logistic 回归模型分析老年人高血压风险因素，结果表明，东西部地区、城镇化率较高以及 PM2.5 污染水平较高的地区，老年人患病率更高。陈星霖[17]在传统 Logistic 回归模型基础上，加上贝叶斯思想，建立贝叶斯 Logistic 回归模型，系统探讨全国、东、中、西部地区老年人慢性病患病情况。分区域模型回归结果显示，虽然各地区的回归结果作用方向与全国大致类似，但是分区域老年人慢性病患病因素重要性存在异质性。任仲达[18]使用多因素 Logistic 回归模型，分析中国中老年人关节炎空间分布，结果显示，西南地区是中老年人关节炎患病的高风险区域。

2.1.2 慢性病共病模式研究

慢性病之间不是完全独立的，很多慢性病之间会存在共同的代谢异常和遗传背景，而且在多种机制上互相影响、互为因果。控制一种慢性病的发生和恶化往往也能减少与之相关的慢性病暴露的风险，一种慢性病暴露的风险增大往往也会增大与之相关的慢性病恶化的可能性。2008 年 WHO 将同时患有两种或两种以上慢性病的现象定义为共病[19]，慢性病共病现象在中老年人群体里十分普遍，甚至很多中老年人同时患有四种及以上慢性病。目前已有众多关于不同慢性病共病模式的研究，主要是基于横截面研究，从患者病历信息或者自我报告慢性病调查数据中挖掘慢性病共病模式，主要采取的研究方法是患病率描述、关联规则挖掘、Logistic 回归模型、聚类分析、网络分析方法等[20]。

（1）患病率描述

直接统计不同慢性病患病人群比例，识别常见的慢性病共病模式。闫伟

等[21]对 2015 年 60 岁以上老年人共病患病描述性统计分析发现，共病率 23.33%的疾病组合“胃部或消化道系统疾病、关节炎或风湿病”是最常见的二元共病模式，共病率 10.14%的疾病组合“高血压、胃部或消化道系统疾病、关节炎或风湿病”是最常见的三元共病模式。患病率描述方法不需要专业知识、操作简单、通俗易懂，但是当面对大量数据集时分析过程繁琐。

（2）关联规则挖掘

关联规则挖掘算法中一个经典的算法是 Apriori 算法，最早由 1993 年 Agrawal 等[22]针对购物篮问题提出，发现不同商品之间的联系，能够帮助卖家更合理进行商品的库存、摆放等。经典的案例就是尿布和啤酒，在超市购物数据中发现男人们时常买尿布的时候会顺便买一些啤酒，于是超市就把尿布和啤酒进行打包销售来获取更多的利润。后面这种方法也广泛应用于医疗数据，挖掘疾病的风险因素和不同疾病的关联性，为疾病的筛查、预防、综合治疗提供了新的思路和方法。

徐小兵等[23]采取 Apriori 算法分析中国老年人共病患病模式，得出 37 条强关联规则，其中高血压、慢性肺部疾患、消化系统疾病、血脂异常和心脏病和其他慢性疾病存在紧密联系和规则，应该针对这些慢性病发病特点，加强对关联疾病的筛查、预防、治疗。程杨杨等[24]对 14 种常见慢性病利用 Apriori 算法，分析中老年人群共病关联规则，结果显示，心脏病和其他慢性疾病共病模式十分常见，且血脂异常、糖尿病、高血压三种慢性病关联性强。Held 等[25]分析悉尼人群患病共病模式，发现“关节炎、听力障碍”是最常见的二元共病模式，“关节炎、听力障碍、肥胖”是最常见的三元共病模式。关联规则结果解释性较强，但是识别出的慢性病共病模式依赖关联规则的设定。

（3）Logistic 回归模型

Logistic 回归模型除了被广泛应用于疾病和疾病影响因素的研究，同时也可以用来分析疾病和疾病的关联关系。陶然等[26]通过非条件 Logistic 回归分析方法，探讨血脂异常和高血压的关系，结果表明，血脂异常和高血压紧密相关，应该注意两个疾病的综合防治。Lin 等[27]基于主成分分析的 Logistic 回归模型来预测中国汉族高血压人群中患肾脏疾病的风险，研究结果表明，已经患有高血压慢性病的人再患有冠心病、糖尿病、肾损伤的风险更大。

（4）聚类分析方法

聚类是一种重要的数据挖掘和模式识别的技术，属于无监督学习方法。聚类的目的是将一组研究对象分为不同群组，使得同一群组的研究对象之间相对同质，不同群组的研究对象有很大差异性。Zemedikun 等[28]通过层次聚类将 36 种慢性病聚类成三种共病模式，分别为“心肌梗死、心绞痛”、“ 心血管、肌肉骨骼、呼吸和神经退行性疾病等 26 种疾病”、“癌症、高血压、哮喘等 8 种疾病”。Lu 等[29]对山西省老年共病模式基于潜在类别分析，识别出“退行性/消化系统疾病模式”、“代谢性疾病模式”和“心血管疾病模式”三种共病模式。王福琳等[30]通过自组织映射神经网络和 K-Means 相融合的聚类算法，对 2015 年中老年人群体 14 种慢性病共病模式可视化聚类，发现所有共病中老年人可以聚类成四类人群。侯家麒[31]分别利用 K-Means 和 DBSCAN 两种聚类方法对 64 种疾病进行聚类处理，深入挖掘不同慢性病之间的关联规则。陈梦碟[32]使用 K-Means++、平均连接法和离差平方和法三种聚类方法对高血压、糖尿病等 27 种慢性疾病进行聚类分析，发现了各种慢性病并发症，从而建立起疾病网络。Marengoni 等[33]以瑞典一个城市为基础的前瞻性队列数据，应用聚类分析研究老年人慢性病共病模式，结果显示聚类成五组，“血管疾病（循环和心肺疾病）”、“精神疾病和肌肉骨骼疾病”、“糖尿病”、“恶性肿瘤”、“视力损害和贫血”。聚类算法简单快速，但是一些聚类算法需要事先指定聚类数目，并且对参数设置依赖性强。

（5）网络分析

随着复杂科学的发展，网络分析方法因其展示直观的优点，已经成为挖掘慢性病共病模式的常用工具。该方法将每一种慢性病作为网络中的一个节点，不同疾病之间的关联作为网络的边，构建慢性病共病网络。通过网络拓扑学指标，此方法不仅能够挖掘慢性病共病模式，还能识别出慢性病的重要性。但是一些网络拓扑学指标实际意义解释、理解的难度大，因此网络分析和关联规则常常结合使用，通过关联规则识别慢性病共病模式，再利用网络图可视化共病模式，识别网络中的核心共病。Hernández 等[34]利用关联规则识别出爱尔兰老年人最常见的“高血压、胆固醇”二元共病模式，再使用网络分析找出其中核心疾病（高血压、胆固醇、关节炎）。

2.2 亚组分析及整合分析相关研究

2.2.1 亚组分析研究

在社会学中，实证研究的方法常常是以变量间的联系为中心的。由于因果推断方法的日益流行，量化社会学的研究已经从注重相互关联的角度，转变为注重因果的角度[35]。传统的平均因果效应从理论和实践两方面都不再适用了。理论方面，众多的社会理论是对特殊群体进行分类，突出了个体之间的差异，在检验和发展这些理论时，学者们格外注意处理效应异质性。现实方面，许多政策相关的研究都关注不同人群之间政策实施效应的差异性[36]。传统的参数异质性处理方法主要是这两种:（1）通过包括研究变量和其他协变量之间的相互作用项来呈现处理效应的异质性[37]，（2）通过将所有有可能引起异质性的变量降维为一个倾向值，研究处理效应如何随着倾向值的改变而改变[38]。

这两种传统的方法都有各自的优缺点，回归模型中增加交互项的分析方法虽然被大量使用，但是也持续存在质疑。主要质疑两个问题：第一，可能带来异质性的因素有很多，怎样确定选择哪些因素出现在模型里。选择过多的因素会导致参数过度化，从而损害一些参数估计的统计性质；第二，选择的因素应该以什么样的形式出现在模型里，因素的形式往往都是研究者的主观设定，缺乏科学性。倾向值的方法，是把所有有可能引起异质性的变量降维为一个值，虽然解决了交互项过度参数化的问题，但是会使得模型缺乏解释性，无法找出异质性的根本来源。并且以上两种方法都很依赖对模型的设定，需要模型设定足够准确，否则可能会得出错误的结论。

现实情况中，每个个体某个处理变量效应完全一致（完全同质）和每个个体某个处理变量效应都不相同（完全异质）这两种极端的情况发生可能性较小，更普遍的情况是不同个体处理变量效应是处于完全异质和完全同质之间。如果斜变量对一些个体的处理效应相似或者相同，那么这些个体属于同一个亚组。同一亚组的个体斜变量的处理效应相似，不同亚组的个体斜变量的处理效应有显著差异。现在大量学者广泛关注的精准医疗[39]就是亚组识别在医学领域上的应用，对不同亚组患者采取差异治疗，对相同亚组患者采取

相似治疗，从而提高医疗的效率和准确性。国内外对数据进行亚组识别的想法由来已久，1960 年末学者们普遍关注数值分析法，聚类分析和判别分析为数值分析的两个分支。数据的亚组识别和聚类分析具有相近的思想，但是亚组分析更多应用于医学数据对混合人群进行分类。Breiman 等[40]提出分类和回归树方法进行亚组识别参数划分，Shen 等[41]提出基于贝叶斯的潜变量模型来识别不同治疗获益率的子群，并应用在随机临床实验中降低死亡率。随着数据量不断增大，一系列非参数方法因其具备更强的通用性和更高的计算效率被广泛应用于亚组识别。Chen 等[42]提出 PRIM 方法搜寻预后特征进行患者分层，Huang 等[43]提出 sequential-batting 方法应用于临床药物开发患者亚组识别，Zeileis 等[44]提出一种基于模型的递归分割算法 MOB。混合模型也是亚组识别和数据聚类的常用方法，Titterington 等[45]提出将数据视为源自自身确定参数值的亚组，使用混合效应模型进行亚组识别。Shen 等[46]基于二项—正态混合模型进行亚组识别。但是基于混合模型的方法往往需要提前设定数据的分布和混合成分的数量，在实际数据中应用难度较大。因此，在没有先验信息的情况下，如何由数据驱动，能够自动识别亚组结构和数目变成亚组分析研究需要解决的问题。惩罚项不仅可以对高维数据进行变量选择，还可以用于亚组分析。Zhu 等[47]在高维回归分析中，为了应对维度灾难，通过非凸惩罚进行分组追踪和特征选择。Ma 等[39]提出对异质回归系数通过增加成对凹函数（SCAD 和 MCP）惩罚把相近的系数分为一组，并用交替方向乘子法对目标函数迭代求解。该方法在稀疏约束下，能够正确识别亚组，在医疗相关研究中得到广泛应用。

2.2.2 整合分析研究

伴随着科技的持续发展，数据的采集和存储都变得越来越方便和快捷。由此，进入了一个数据量爆炸的时代，每时每刻都在不知不觉中产生大量数据。由不同来源、格式或者主体的数据整合在一起形成了大数据，例如源自不同省份的问卷调查数据、不同实验室的医学数据等。合并多个数据集建模的方式经常可见，大数据分析的重要目标就是研究不同子样本之间的关联性和差异性[48]。但是对这种数据整合形成的样本建模会有一些特殊性：一方面，

如果直接对不同来源的样本各自单独建模，会使得待估参数增多，模型过于冗余，也会忽略数据间的相似性；另一方面，如果简单的把所有样本直接合并，会无视不同来源数据的同一变量回归参数的差异性。所以为了研究不同数据源同一变量回归参数的相似性和差异性，整合分析方法（Integrative Analysis）基于合并不同来源数据的目标函数，从统计角度出发，综合考虑数据的同质性和异质性，共同求解多个模型。起源于 1960 年的整合分析方法，是一种对不同来源数据集同时分析的方法，综合考虑多个来源数据集。实际研究过程中，样本来自不同数据集的情况十分常见，不同数据集的样本往往存在差异，从而导致模型不稳定。整合分析方法能够集中更多的信息，同时克服建模不稳定，大幅度提高模型的解释性和预测能力[49]。不同来源数据的整合分析，主要分为两类：一类是横向整合，即变量相同但是样本不同；另一类是纵向整合，即样本相同但是变量不同[50]。

整合分析是解决特征维度多、样本量少的有效手段。它集中多个数据集而大量增加样本，有效解决数据信息量不足问题。同时随着收集的数据爆发式增长，大数据带来的维度灾难问题也十分突出，任何微不足道的噪音都可能被收集起来，如何提取出有效的信息是需要解决的主要问题。惩罚函数思想和整合分析相结合的惩罚方法的整合分析（Penalized Integrative Analysis）是降低数据维度、提取有效信息的常见手段。基于惩罚函数的整合分析不仅可以选择模型中有效变量，还可以确定不同来源数据集之间的关联性和差异性。

惩罚思想是单数据集变量选择中被广泛使用的一种有效方法，通过对参数施加约束进行压缩，它不仅可以选择有效变量来提高模型可解释性，还可以降低参数估计偏差来提高预测精度。基于惩罚思想的变量选择，主要分为以下几类：针对单个变量的变量选择，Lasso、SCAD、MCP、Bridge；针对共线数据的变量选择，弹性网、Mnet 等；针对以组形式的变量选择，Group Lasso、CAP 等；针对组内和组间双重变量选择，Sparse Group Lasso、L_1 Group Bridge 等。

整合分析的变量选择仍然可以运用惩罚思想，不同之处在于整合分析的变量选择面临两个问题：第一，如何选择有效变量；第二，如何确定这些变量在不同来源数据集的显著性。Ma 等[51]在传统 Logistic 模型中加入 L_2 Group

Bridge 惩罚函数，考虑组内和组间复合惩罚，从理论上证明了此惩罚方法在整合分析中的显著优势。马双鸽等[49]通过多次模拟，对各种惩罚方法进行比较，结果表明 L_1 Group Bridge 惩罚函数表现最好，并应用于具有地区差异的新农合家庭医疗支出以及来源于 3 个不同研究机构的肺癌基因数据筛选分析。惩罚方法的整合分析除了能在不同数据集筛选出变量的显著关系，也可以研究同一个变量在不同来源数据集中回归参数之间的关联性。

整合分析和亚组分析具有紧密的联系，整合分析的每个回归参数的含义需要从两方面来理解：一方面是变量层面，这与单个数据集的模型含义一致；另一个方面是反映不同数据集之间的关联性，有 M 个不同来源的数据集，同一个解释变量会有 M 个回归参数，M 个不同数据集之间的联系恰巧是通过这 M 个回归参数连接。因此亚组分析方法被广泛应用于跨数据集的研究中，挖掘出同一变量回归参数在不同来源中的关联性。段谦等[52]将 MCP 惩罚函数与二值分割算法相结合，提出一种新的高维纵向数据亚组识别方法，并将此方法应用于全国 31 个省份 13 年的产业结构数据集。Shi 等[53]提出对回归参数差值进行惩罚的 Contrast 惩罚函数，能够挖掘出同一变量回归参数在不同来源中的关联性。Shen 等[54]未考虑斜变量的顺序，通过惩罚所有成对系数的差值，提取基因数据的分组结构，该方法计算量大且只针对同一来源数据。为了减少计算量，Wang 等[55]针对合并多个研究数据，提出一种基于自适应参数排序的 Fused Lasso（FLAPO）方法，只考虑相邻对的参数差异。FLAPO 能够显著降低计算量，并且和所有成对系数差异约束模型相比，参数估计的误差更小。Tang 等[56]提出基于 Fused Lasso 惩罚函数的方法 FLARCC 来实现不同数据集中回归参数的聚类，FLARCC 方法考虑了参数的顺序，在不损失参数估计的统计性质的情况下，大幅度提升了计算效率。

2.3 文献评述

通过从多种角度的文献搜集、梳理，可以看出已有研究在慢性病空间分布关联性、慢性病共病模式研究、亚组分析、整合分析方面都十分丰富，这也给本文的研究给予了很多指导和灵感。然而，目前研究还存在一些不足和可以深入的地方，现有文献对中老年人慢性病空间分布关联性研究中，大多

都只考虑是否患慢性病或者某个单独的慢性病，很少对多种不同慢性病进行探讨。同时，研究大多都是慢性病关联性或者慢性病空间分布的单一视角，综合考虑慢性病关联性和空间分布的文献很少。另外，很少有慢性病关联性研究是基于多源数据环境进行分析的。对于参数的亚组分析方法，现有文献大多都是针对单个数据源，涉及多源数据的参数融合的研究很有限。并且在有限的多源数据参数聚类研究中，很少有文献提出参数分解的双重亚组分析方法。

因此，本文将在现有的异质性分析基础上，提出一种基于矩阵低秩分解的双重亚组（聚类）学习方法 **LRDSA**，将此方法应用于多源实证数据中老年人慢性病关联研究中，同时探索慢性病和省份两个维度特征向量的亚组结构。

3.相关模型及算法介绍

本章主要介绍关于 **LRDSA** 学习方法的相关预备知识，简要阐述了 Logistic 模型、Lasso 模型、Fused Lasso 模型以及 ADMM 求解算法。

3.1 Logistic 模型

逻辑回归模型是一种常见的定性变量分析的统计方法，根据分类型或者连续型自变量来预测分类型因变量。Logistic 回归分析方法不但可以从大量自变量中选择出对被解释变量有显著影响的变量，还可以给出被解释变量的预测情况。因变量取值为 0、1 的称为二元逻辑回归，因变量取值多于两个类别的称为多元逻辑回归。逻辑回归模型不需要满足很多古典假定，例如不需要自变量满足正态分布和自变量协方差矩阵相等，因此该方法被广泛应用。

二分类逻辑回归，因变量取值为 0、1，常常用于医学中慢性病患病与否的诊断。大量学者们都用逻辑回归模型[57]或者逻辑回归模型基础上改进的模型[58]来研究各种慢性病的患病风险因素，从而进行慢性病患病诊断和预测。

二分类任务，希望输出 $y\in\{0,1\}$，但是传统线性模型输出的预测值如(3-1)所示：

$$z=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b, \tag{3-1}$$

利用 Sigmoid 函数使得输入的每一组数据都能映射到 0～1 之间的数，Sigmoid 函数形式如下：

$$y=\frac{1}{1+e^{-z}}, \tag{3-2}$$

将 Sigmoid 函数带入传统线性模型，得到下式：

$$y=\frac{1}{1+e^{-\left(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b\right)}}, \tag{3-3}$$

对于向量 $\boldsymbol{x}$，属于类别 1、0 的概率如下所示：

$$\begin{aligned} p(y=1\mid\boldsymbol{x})&=\frac{e^{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b}}{1+e^{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b}}=\sigma\left(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b\right),\\ p(y=0\mid\boldsymbol{x})&=\frac{1}{1+e^{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b}}=1-\sigma\left(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b\right), \end{aligned} \tag{3-4}$$

参数估计通过极大似然估计方法得到 $\boldsymbol{\omega},b$ 相应估计。

给定测试数据集 $\left\{\left(\boldsymbol{x}_i,y_i\right)\right\}_{i=1}^{n}$，模型似然函数为：

$$L(\boldsymbol{\omega},b)=\sum_{i=1}^{n}\ln p\left(y_i\mid\boldsymbol{x}_i;\boldsymbol{\omega},b\right), \tag{3-5}$$

通过将 $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b$ 写成 $\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}$，式(3-5)改写成式子(3-6)：

$$L(\boldsymbol{\beta})=\sum_{i=1}^{n}\ln\left(y_i p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)+\left(1-y_i\right)p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right). \tag{3-6}$$

其中 $p_1(\hat{\boldsymbol{x}};\boldsymbol{\beta})=p(y=1\mid\hat{\boldsymbol{x}};\boldsymbol{\beta}),p_0(\hat{\boldsymbol{x}};\boldsymbol{\beta})=p(y=0\mid\hat{\boldsymbol{x}};\boldsymbol{\beta})=1-p_1(\hat{\boldsymbol{x}};\boldsymbol{\beta})$，式(3-6)是关于 $\boldsymbol{\beta}$ 的高阶可导连续凸函数，可以用经典的数值优化算法求解，$\boldsymbol{\beta}^{*}=\arg\min_{n}L(\boldsymbol{\beta})$。

3.2 Lasso 模型

如今由于数据的收集、储存都十分方便，人们面对的数据维度和容量在不断提升。传统的统计理论已经不能解决高维统计的维度灾难问题，最小二乘法在变量个数大于样本个数时，不能得到唯一的估计结果，并且得到的大部分结果都会产生过拟合的问题。受到 Breiman[59]提出的非负铰链模型的启发，Tibshirani[60]提出了 Lasso 模型。Lasso 模型通过在模型中加入 L_1 惩罚项，能够筛选出对模型比较重要的变量，将对模型不太重要的变量压缩为 0，从而降低数据的维度。假设在线性回归中存在 n 组观测样本 $\left(\boldsymbol{x_i},y_i\right)$，$i=1,\ldots,n$，$\boldsymbol{x}_i=\left(x_{i1},\ldots,x_{ip}\right)^{T}$ 为 p 维自变量，$\boldsymbol{\beta}=\left(\beta_1,\ldots,\beta_p\right)^{T}$ 为系数向量，y_i 是相应的响

应变量，对普通最小二乘估计法加上正则化，Lasso 回归模型如(3-7)所示：

$$f(\boldsymbol{\beta})=\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|. \tag{3-7}$$

其中，第一项为残差平方和，第二项为L_1惩罚项，$\lambda>0$，Lasso 罚模型通过L_1惩罚项，能够压缩模型系数，可以把对模型贡献度较小的特征系数压缩到零，有效的进行了变量选择和防止模型过拟合。但是同时 Lasso 也有很多的不足之处，对于不同的系数 Lasso 采取同样的压缩，这可能会导致最终模型特别稀疏。此外，Lasso 只考虑到单个变量，没有考虑到很多特征之间的相似性。因此在 Lasso 模型的基础上，很多学者对于惩罚项进行变化，产生了很多扩展的 Lasso 模型。

3.3 Fused Lasso 模型

Tibshirani 和 Saunders[61]于 2005 年在 Lasso 模型的基础上提出了 Fused Lasso 模型。Fused Lasso 模型在传统 Lasso 模型的基础上加上了系数差分的绝对值约束，它认为系数之间其实是有一定顺序的，并且往往系数和它的邻近系数差异较小。于是 Fused Lasso 模型不仅能够筛选出对模型不重要的变量，还使得相邻系数之间的差异尽可能小，邻近系数之间能够更加平滑。并且和不考虑系数顺序的成对融合模型相比，可以大幅减少计算量。

假设存在 n 组观测样本$(\boldsymbol{x_i},y_i)$，$i=1,\ldots,n$，$\boldsymbol{x}_i=\left(x_{i1},\ldots,x_{ip}\right)^T$为 p 维自变量，$\boldsymbol{\beta}=\left(\beta_1,\ldots,\beta_p\right)^T$为系数向量，$y_i$是相应的响应变量，Fused Lasso 模型的定义式可以表示为：

$$f(\beta)=\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda_1\sum_{j=1}^{p}\left|\beta_j\right|+\lambda_2\sum_{j=2}^{p}\left|\beta_j-\beta_{j-1}\right|. \tag{3-8}$$

其中$\lambda_1>0,\lambda_2>0$，模型有两个约束，第一个约束就是 Lasso 的L_1惩罚项，是对于单个系数希望尽可能接近 0；第二个约束是 Fusion 惩罚项，对于邻近系数的差异希望尽可能稀疏。Fused Lasso 具有唯一解，它是严格凸并且能准确捕捉一个时间间隔内跳跃点的位置。

Fused Lasso 模型应用广泛，被提出用于识别各研究对象或者数据集的回归系数的异质性模式，可以用来实现回归系数的聚类。并且结合 Fused Lasso 函数的聚类，不需要事先指定聚类的数目，是完全由数据驱动的，减少很多人为设定的主观性，在实际数据中具有更强的实用性。现实中由不同来源整合的多源数据集，斜变量效应完全异质和完全同质的情况都比较特殊，更常见的情况是多个数据集的回归系数会形成关联子群，在同一个子群里斜变量对因变量的效应是相同的，在不同子群里斜变量对因变量的效应是有显著差异的。

对于 Fused Lasso 模型的参数估计求解问题，采用的是交替方向乘子法（ADMM）。

3.4 ADMM 算法

3.4.1 软阈值和硬阈值函数

最早 Donoho 和 Johnstone[62]于 1994 年提出了软阈值(soft thresholding)函数和硬阈值(hard thresholding)函数，硬阈值函数的定义表达式如下：

$$\eta_H(w,\lambda)=wI\,|w|>\lambda, \tag{3-9}$$

软阈值函数的定义表达式如下：

$$\eta_s(w,\lambda)=sgn(w)(|w|-\lambda)_+, \tag{3-10}$$

其中 $I\,|w|>\lambda$ 为示性函数，sgn 为符号函数，当 $|w|>\lambda$ 时，$(|w|-\lambda)_+=|w|-\lambda$，当 $|w|\leq\lambda$ 时，$(|w|-\lambda)_+=0$，因此硬阈值函数的定义表达式等同如下：

$$\eta_H(w,\lambda)=\begin{cases} w, & |w|>\lambda \\ 0, & |w|\leq\lambda \end{cases}, \tag{3-11}$$

软阈值函数的定义表达式等同如下：

$$\eta_s(w,\lambda)=\begin{cases} w-\lambda, & w>\lambda \\ 0, & |w|\leq\lambda \\ w+\lambda, & w<-\lambda \end{cases}. \tag{3-12}$$

3.4.2 交替方向乘子法求解一般约束问题

面对大量数据，如何快速得到 Lasso 模型等凸优化求解是一个十分重要的问题。2010 年 Boyed[63]等人对交替方向乘子法(ADMM)进行详细介绍，指出此方法最早是 1976 年由 Gabay 和 Mercier[64]提出的。交替方向乘子法广泛被应用于解决可分离凸规划问题，通过在原目标函数的基础上，添加约束，构造增广拉格朗日函数，每次固定其他参数，求解一个参数，不断交替方向求解，使得复杂的高维优化问题转化为多个简单的低维问题。并且 ADMM 算法得到的最优解，不仅收敛速度快、而且统计性质良好，所以该算法被广泛应用于各个领域。

交替方向乘子法一般求解如下带有等式约束的凸优化问题：

$$\begin{aligned} &\min F(\beta)+G(\gamma) \\ &s.t.\ X\beta+Y\gamma=b, \beta\in\Pi, \gamma\in\Gamma. \end{aligned} \tag{3-13}$$

在上式(3-13)中，$\beta\in\mathbb{R}^n$ 和 $\gamma\in\mathbb{R}^m$ 是待优化的目标变量，$\Pi\subseteq\mathbb{R}^n$ 和 $\Gamma\subseteq\mathbb{R}^m$ 是已知的非空闭式凸集，$F:\Pi\to\mathbb{R}$ 和 $G:\Gamma\to\mathbb{R}$ 是符合条件的闭式凸函数。$X\in\mathbb{R}^{l\times n}$ 和 $Y\in\mathbb{R}^{l\times m}$ 是已知的样本矩阵，$b\in\mathbb{R}^l$ 是确定的约束向量。设非空 Ω^* 为上式(3-13)的解集。

上式(3-13)构造增广拉格朗日函数如下：

$$\begin{aligned} L_\mu(\beta,\gamma,\alpha)=F(\beta)+G(\gamma)+\alpha^T(X\beta+Y\gamma-b) \\ +\frac{\mu}{2}\left\|X\beta+Y\gamma-b\right\|_2^2, \end{aligned} \tag{3-14}$$

其中，$\mu>0$ 是一个可以改变大小的惩罚参数，$\alpha\in\mathbb{R}^l$ 既是拉格朗日乘子，同时也是对偶变量。首先运用增广拉格朗日乘子法来对上式(3-14)进行求解。在给定 α^0 初始值后，式(3-14)的优化求解迭代步骤如下[65]：

$$\left\{\begin{array}{l} (\beta^{k+1},\gamma^{k+1})=\arg\min\left\{L_\mu\left(\beta,\gamma,\alpha^k\right)\right\} \\ \alpha^{k+1}=\alpha^k+\mu\left(X\beta^{k+1}+Y\gamma^{k+1}-b\right) \end{array}\right\}. \tag{3-15}$$

从上式优化求解迭代步骤可以看出，(β,γ) 两个变量每次是同时更新的，在每次迭代求解中要将 (β,γ) 作为一个整体，一起迭代求解。这样的求解方式对函数 F,G 是有要求的，如果函数 F,G 具有一些特殊的性质，整体一起求

解(β,γ)得到的结果将不能保证是有效的。于是(β,γ)不再作为一个整体更新，而是每次先固定一个变量，更新另一个变量，再更新对偶变量，不断交替方向，迭代求解。具体求解步骤如下式(3-16)所示:

$$\begin{cases}\beta^{k+1}=\arg\min\left\{L_{\mu}\left(\beta,\gamma^{k},\alpha^{k}\right)\right\}\\ \gamma^{k+1}=\arg\min\left\{L_{\mu}\left(\beta^{k+1},\gamma,\alpha^{k}\right)\right\}\\ \alpha^{k+1}=\alpha^{k}+\mu\left(X\beta^{k+1}+Y\gamma^{k+1}-b\right)\end{cases}. \tag{3-16}$$

3.4.3 交替方向乘子法求解 Fused Lasso 模型

这一小节重点介绍交替方向乘子法(ADMM)求解 Fused Lasso 模型的具体步骤。

Fused Lasso 模型可以由式子(3-8)重新写成下面式子：

$$\min_{\beta\in\mathbb{R}^{p}}\left\{\frac{1}{2}\left\|y-X\beta\right\|_{2}^{2}+\lambda_{1}\left\|\beta\right\|_{1}+\lambda_{2}\left\|R\beta\right\|_{1}\right\}, \tag{3-17}$$

其中，矩阵$R\in\mathbb{R}^{(p-1)\times p}$定义如下：

$$R=\begin{bmatrix}-1 & 1 & 0 & \cdots & 0\\ 0 & -1 & 1 & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & -1 & 1\end{bmatrix}, \tag{3-18}$$

为了应用 ADMM 算法求解式子(3-17)，引入辅助变量$\gamma=R\beta(\gamma\in\mathbb{R}^{p-1})$，于是 Fused Lasso 模型变为下面式子：

$$\begin{aligned}&\min_{\beta\in\mathbb{R}^{p}}\left\{\frac{1}{2}\left\|y-X\beta\right\|_{2}^{2}+\lambda_{1}\left\|\beta\right\|_{1}+\lambda_{2}\left\|\gamma\right\|_{1}\right\}\\ &s.t.\ \ R\beta=\gamma.\end{aligned} \tag{3-19}$$

增广拉格朗日函数为：

$$\begin{aligned}L(\beta,\gamma,\alpha)=&\frac{1}{2}\left\|y-X\beta\right\|_{2}^{2}+\lambda_{1}\left\|\beta\right\|_{1}+\lambda_{2}\left\|\gamma\right\|_{1}\\ &+\alpha^{T}(R\beta-\gamma)+\frac{\mu}{2}\left\|R\beta-\gamma\right\|_{2}^{2},\end{aligned} \tag{3-20}$$

于是式子(3-20)交替方向乘子法的求解步骤是：

$$\begin{cases} \beta^{k+1} = \arg\min\limits_{\beta\in\mathbb{R}^p}\left\{L_\mu\left(\beta,\gamma^k,\alpha^k\right)\right\} \\ \gamma^{k+1} = \arg\min\limits_{\gamma\in\mathbb{R}^{p-1}}\left\{L_\mu\left(\beta^{k+1},\gamma,\alpha^k\right)\right\}. \\ \alpha^{k+1} = \alpha^k + \mu\left(R\beta^{k+1} - \gamma^{k+1}\right) \end{cases} \tag{3-21}$$

第一，对于 β 的求解问题，因为 L_μ 对 β 是可导的，所以直接令 $\frac{\partial L}{\partial \beta}$，可以得到 β 的解。

$$\beta^{k+1} = \arg\min_{\beta\in\mathbb{R}^p} \frac{\partial L_\mu\left(\beta,\gamma^k,\alpha^k\right)}{\partial\beta}. \tag{3-22}$$

第二，对于 γ 的求解问题，γ 的解通过下式得到：

$$\begin{aligned} \gamma^{k+1} &= \arg\min_{\gamma\in\mathbb{R}^{p-1}}\left\{\lambda_2\|\gamma\|_1 + \frac{\mu}{2}\left\|\gamma - R\beta^{k+1} - \frac{\alpha^k}{\mu}\right\|_2^2\right\} \\ &= \eta_s\left(R\beta^{k+1} + \frac{\alpha^k}{\mu}, \frac{\lambda_2}{\mu}\right). \end{aligned} \tag{3-23}$$

其中 γ 的求解问题用到软阈值函数，其计算见式子(3-12)。

综上，求解 Fused Lasso 模型的交替方向乘子法（ADMM）具体步骤如下表 3-1 所示：

表 3-1　ADMM 算法求解 Fused Lasso 问题

算法：

输入：自变量 X，响应变量 y，惩罚参数 λ_1,λ_2,μ，停止条件 η;

输出：$\hat{\beta},\hat{\gamma},\hat{\alpha}$;

目标：多次迭代求解获得 β,γ,α.

初始化 $(\beta^0,\gamma^0,\alpha^0)=(0,0,0), k=0, \mu=1, \eta=0.001$.

While $k \geq 0$, do

通过式子(3-22)计算 β^{k+1}；

通过式子(3-23)计算 γ^{k+1}；

通过 $\alpha^{k+1}=\alpha^k+\mu\left(R\beta^{k+1}-\gamma^{k+1}\right)$ 计算 α^{k+1}.

If $r^{k+1}=R\beta^{k+1}-\gamma^{k+1}, \left\|r^{k+1}\right\|<\eta$

then

$(\hat{\beta},\hat{\gamma},\hat{\alpha})=\left(\beta^{k+1},\gamma^{k+1},\alpha^{k+1}\right)$;

Break;

Else

$k=k+1$;

End

End

4.基于矩阵低秩分解的双重亚组学习及数值模拟

小样本估计时会由于信息量不足带来模型估计效果不佳的问题，为了增加样本量和改善模型估计参数的能力，对多个来源的数据进行整合是一种常用的增加信息的方法。整合多源数据分析带来的好处是毋庸置疑的，但同时也给建模带来很多挑战，如何识别不同子样本之间的关联性和差异性是迫切需要解决的问题。目前大量学者在整合分析的基础上加上各种惩罚项，来进行变量选择和分析不同来源数据集的同构以及异构模型。

实际上，不同数据集模型完全同质和完全异质的可能性都比较小，更多的是不同数据集会形成关联子群，同一个子群里的数据源变量的效应是相似的，不同子群的数据源变量效应是有显著差异的。传统的亚组分析方法大多都是针对单个数据集，很少有考虑到多个数据集，现有的多个数据集亚组识别的方法几乎很少考虑到单个参数进一步分解成多个参数的双重亚组分析。因此，本文基于多源数据异质性不可避免的现实情况，提出基于矩阵低秩分解的双重亚组（聚类）学习方法 **LRDSA**。该方法将二维参数矩阵采取低秩分解，提取出两个维度的特征向量，实现多源数据的整合和两个维度分别的关联性分析。与同质和异质模型相比，该学习方法具有更高的参数估计准确性以及聚类准确度。此外，本文采用基于数据驱动的 Fused Lasso 融合聚类，不需要提前给定聚类数目，在实际数据中具有更强的实用性。

4.1 LRDSA 方法

假设观测样本 $y_{i,s,d}$ 表示来自第 d 个省份、第 i 个个体慢性病 s 的患病情况，$y_{i,s,d}$ 的取值为 $\{0,1\}$ 的二分类变量，其取值为 1 时表示来自 d 省份的第 i

个个体患有s慢性病，其取值为 0 时表示来自d 省份的第i个个体不患有s慢性病。 $x_i=(x_{i1},\ldots,x_{ip})^T$ 是 p 维控制变量， $\theta=(\theta_1,\ldots,\theta_p)^T\in\mathbb{R}^P$ 表示相应的控制变量回归系数。$\gamma_{s,d}$ 表示来自第d 个省份、第s种慢性病时模型的截距项，也就是连续变量取平均值，虚拟变量取对照组的基准水平。设省份的数量为 D，慢性病数量为 S ，来自第s个慢性病、第d 个省份的样本数量为$n_{s,d}$，总样本数量为$n=\sum_{s=1}^{S}\sum_{d=1}^{D}n_{s,d}$。由于被解释变量为二元变量，所以对每一个省份每个慢性病的患病情况分别建立二元 Logistic 回归模型如下：

$$P\left(y_{i,s,d}=1\right)=\frac{1}{1+e^{-\left(\gamma_{s,d}+x_i^T\theta\right)}},i=1,\ldots,n,s=1,\ldots,S,d=1,\ldots,D. \tag{4-1}$$

为了借用不同慢性病和省份的信息，得到模型截距项在不同慢性病s和不同省份d 回归参数的异同结构，整合 S 个慢性病和D个省份的全部数据集，建立如下目标函数：

$$L(\alpha,\beta,\theta)=\sum_{i=1}^{n_{S,D}}\sum_{s=1}^{S}\sum_{d=1}^{D}\left(y_{i,s,d}\log\left(P\left(y_{i,s,d}=1\right)\right)+\left(1-y_{i,s,d}\right)\log\left(1-P\left(y_{i,s,d}=1\right)\right)\right). \tag{4-2}$$

截距项$\gamma_{s,d}$随省份d 和慢性病s的取值变化而变化，截距项$\gamma_{s,d}$异质性表示不同省份不同慢性病的平均水平，参数矩阵γ形式如下所示：

$$\gamma=\begin{bmatrix}\gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,D}\\ \gamma_{2,1} & \gamma_{2,2} & \cdots & \gamma_{2,D}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{S,1} & \gamma_{S,2} & \cdots & \gamma_{S,D}\end{bmatrix}. \tag{4-3}$$

为了有效识别来自多个省份数据的异质结构，整合同质数据推断慢性病间的关联性，进一步定位异质性来源，采取矩阵低秩分解。矩阵γ可以表示为若干个向量外积之和，若矩阵γ最少可以由R个向量的外积之和表示，如式(4-4)所示：

$$\gamma=\alpha_1\otimes\beta_1+,\ldots,+\alpha_R\otimes\beta_R, \tag{4-4}$$

其中， α_r 为S 维向量， β_r为D维向量， $r=1,\ldots,R$ ，矩阵γ的秩为R，二维矩阵低秩分解实现过程如图 4-1 所示：

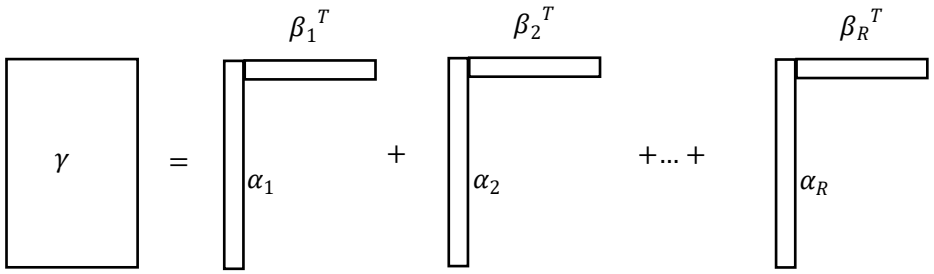

图 4-1 二维矩阵低秩分解

这里假定 $R=1$，使得模型简单且易于解释，此时 α 表示慢性病维度的特征向量，β 表示省份维度的特征向量。

不同慢性病和省份之间存在关联也存在一定差异，为了估计模型的参数 $\{\alpha,\beta,\theta\}$ 以及同时识别两个维度 α,β 的同构、异构结构，在原似然函数的基础上加上惩罚项 $P_{\lambda_1,\lambda_2}(\alpha,\beta)$, 目标函数由式子(4-2)变为式子(4-5)。

$$-L(a,\beta,\theta)+P_{\lambda_1,\lambda_2}(\alpha,\beta). \tag{4-5}$$

$$P_{\lambda_1,\lambda_2}(\alpha,\beta)=\lambda_1\left(\sum_{s=1}^{S-1}\sum_{s'>s}^{S}\mathbf{1}\left\{\left|u_s-u_{s'}\right|=1\right\}\left|\alpha_s-\alpha_{s'}\right|\right)+\lambda_2\left(\sum_{d=1}^{D-1}\sum_{d'>d}^{D}\mathbf{1}\left\{\left|v_d-v_{d'}\right|=1\right\}\left|\beta_d-\beta_{d'}\right|\right)$$

$u=\left(u_1,u_2,\ldots,u_S\right)^T$，$v=\left(v_1,v_2,\ldots,v_D\right)^T$，$u_s=\sum_{s'=1}^{S}\mathbf{1}\left\{\alpha_{s'}\le\alpha_s\right\}$，$v_d=\sum_{d'=1}^{D}\mathbf{1}\left\{\beta_{d'}\le\beta_d\right\}$，

为了减少计算量，这里不对所有的参数差进行惩罚，而是通过相邻两对参数的差来识别亚组，若参数差等于 0，则属于同一亚组。由于参数的顺序是未知的，u,v 分别表示 α,β 的顺序，采用 Tang 等[56]提出的无惩罚项目标函数得到的一致估计, 即最大化式子(4-2)得到的解来定义参数的顺序。λ_1,λ_2 为惩罚参数，$\lambda>0$，关于 λ_1,λ_2 的设定，是基于 BIC 准则来实现。通过 BIC 准则来寻找最优的聚类结果，使得该方法不需要提前给定聚类数目，完全由数据来驱动，减少人为设定的主观性，BIC 定义表达式如下：

$$BIC=-2L\left(\hat{\alpha}\left(\lambda_1\right),\hat{\beta}\left(\lambda_2\right)\right)+\left(df\left(\hat{\alpha}\left(\lambda_1\right)\right)+df\left(\hat{\beta}\left(\lambda_2\right)\right)\right)\log(n), \tag{4-6}$$

其中 $L(\widehat{\alpha}(\lambda_1),\widehat{\beta}(\lambda_2))$ 为式子(4-2)计算出的值，$\hat{\alpha}\left(\lambda_1\right)$ 是 λ_1 取值下参数 α 的估计，$\hat{\beta}\left(\lambda_2\right)$ 是 λ_2 取值下参数 β 的估计，$df\left(\hat{\alpha}\left(\lambda_1\right)\right)=\sum_{s=1}^{S}df(\hat{\alpha}_s,\lambda_1)$ 是 $\hat{\alpha}\left(\lambda_1\right)$

中总共不相等的参数个数，$df\left(\hat{\beta}\left(\lambda_2\right)\right)=\sum_{d=1}^{D} df(\hat{\beta}_d,\lambda_2)$ 是 $\hat{\beta}\left(\lambda_2\right)$ 中总共不相等的参数个数，n 为样本个数。为了保证参数的可识别性，这里设 $\beta_1=1$。**LRDSA** 方法示意图如图 4-2 所示。

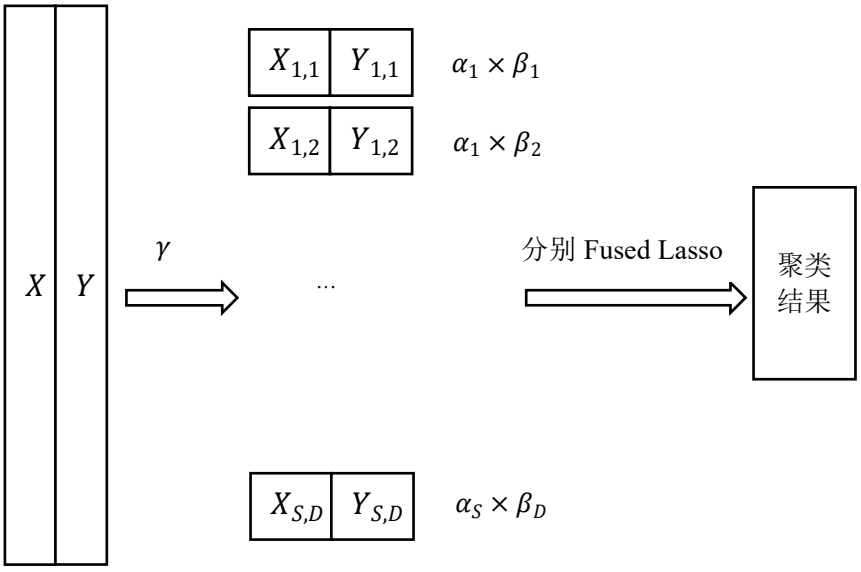

图 4-2 LRDSA 方法示意图

4.2 LRDSA 求解算法

由于目标函数(4-5)中损失函数部分虽然可导，但是 Fused Lasso 惩罚函数部分不可导，直接优化(4-5)十分困难，传统梯度下降法、牛顿法已经不再适用，需要一些特殊算法求解。ADMM 算法是求解 Lasso 以及 Lasso 扩展形式的经典算法，所以本文提出的 **LRDSA** 方法也用 ADMM 算法来求解，首先引入辅助变量 $w_{s,s'},g_{d,d'}$，优化目标函数(4-5)等价于最小化下列函数问题：

$$\min_{\alpha,\beta,\theta}\begin{pmatrix} -L(a,\beta,\theta)+\lambda_1\left(\sum_{s=1}^{S-1}\sum_{s'>s}^{S}\mathbf{1}\{|u_s-u_{s'}|=1\}|\alpha_s-\alpha_{s'}|\right) \\ +\lambda_2\left(\sum_{d=1}^{D-1}\sum_{d'>d}^{D}\mathbf{1}\{|v_d-v_{d'}|=1\}|\beta_d-\beta_{d'}|\right) \end{pmatrix} \tag{4-7}$$
$$s.t\ \begin{cases}\alpha_s-\alpha_{s'}=w_{s,s'} \\ \beta_d-\beta_{d'}=g_{d,d'}\end{cases}.$$

构造增广拉格朗日函数：

$$\begin{aligned} Q(\alpha,\beta,w,g,\varphi,\phi)=&-L(a,\beta,\theta)+\lambda_1\left(\sum_{s=1}^{S-1}\sum_{s'>s}^{S}\mathbf{1}\{|u_s-u_{s'}|=1\}|w_{s,s'}|\right) \\ &+\lambda_2\left(\sum_{d=1}^{D-1}\sum_{d'>d}^{D}\mathbf{1}\{|v_d-v_{d'}|=1\}|g_{d,d'}|\right) \\ &+\sum_{s=1}^{S-1}\sum_{s'>s}^{S}\varphi_{s,s'}\mathbf{1}\{|u_s-u_{s'}|=1\}(\alpha_s-\alpha_{s'}-w_{s,s'}) \\ &+\sum_{d=1}^{D-1}\sum_{d'>d}^{D}\phi_{d,d'}\mathbf{1}\{|v_d-v_{d'}|=1\}(\beta_d-\beta_{d'}-g_{d,d'}) \\ &+\frac{\rho}{2}\sum_{s=1}^{S-1}\sum_{s'>s}^{S}\mathbf{1}\{|u_s-u_{s'}|=1\}(\alpha_s-\alpha_{s'}-w_{s,s'})^2 \\ &+\frac{\rho}{2}\sum_{d=1}^{D-1}\sum_{d'>d}^{D}\mathbf{1}\{|v_d-v_{d'}|=1\}(\beta_d-\beta_{d'}-g_{d,d'})^2. \end{aligned} \tag{4-8}$$

其中 $\varphi_{s,s'},\phi_{d,d'}$ 是对偶变量，$\rho>0$ 是惩罚参数，通过交替方向乘子法（ADMM）来迭代求解 $(\alpha_s,\beta_d,w_{s,s'},g_{d,d'},\varphi_{s,s'},\phi_{d,d'})$ 的估计。

首先求 (α_s,β_d) 的更新值，$Q(\alpha_s,\beta_d,w_{s,s'},g_{d,d'},\varphi_{s,s'},\phi_{d,d'})$ 对 (α_s,β_d) 是可导的，直接令 $\frac{\partial Q}{\partial\alpha_s,\beta_d}=0$，可以求解出 (α_s,β_d)。

然后求 $(w_{s,s'},g_{d,d'})$ 的更新值，

$$\begin{cases} w_{s,s'}=\eta_s\left(\frac{\varphi_{s,s'}}{\rho}+\alpha_s-\alpha_{s'},\frac{\lambda_1}{\rho}\right) \\ g_{d,d'}=\eta_s\left(\frac{\phi_{d,d'}}{\rho}+\beta_d-\beta_{d'},\frac{\lambda_2}{\rho}\right)\end{cases}. \tag{4-9}$$

其中用到了软阈值算子，$\frac{\varphi_{s,s'}}{\rho}+\alpha_s-\alpha_{s'}$为第$s,s'$个分量，$\frac{\lambda_1}{\rho}$为阈值。

$$\eta_s\left(\frac{\varphi_{s,s'}}{\rho}+\alpha_s-\alpha_{s'},\frac{\lambda_1}{\rho}\right)=\begin{cases}\frac{\varphi_{s,s'}}{\rho}+\alpha_s-\alpha_{s'}-\frac{\lambda_1}{\rho} & ,\frac{\varphi_{s,s'}}{\rho}+\alpha_s-\alpha_{s'}>\frac{\lambda_1}{\rho}\\ 0 & ,\left|\frac{\varphi_{s,s'}}{\rho}+\alpha_s-\alpha_{s'}\right|\le\frac{\lambda_1}{\rho}\\ \frac{\varphi_{s,s'}}{\rho}+\alpha_s-\alpha_{s'}+\frac{\lambda_1}{\rho} & ,\frac{\varphi_{s,s'}}{\rho}+\alpha_s-\alpha_{s'}<-\frac{\lambda_1}{\rho}\end{cases}. \quad (4\text{-}10)$$

最后再求$\left(\varphi_{s,s'},\phi_{d,d'}\right)$的更新值，一直不断重复上述步骤，直到达到收敛条件。

综上，下表 4-1 详细给出了本文提出的 **LRDSA** 学习方法总的更新策略：