基于聚类与分类方法对我国省级政府返乡创业政策工具使用倾向的研究

数 71 王梓涵 2017012136

摘要

2016 年底《国务院办公厅关于支持返乡下乡人员创业创新促进农村一二三产业融合发展的意见》印发后，返乡创业政策在全社会得到广泛响应，各地方政府强化主体责任，出台了大量返乡创业扶持政策。本研究从各省级政府网站收集到的返乡创业政策有效文本 31 份，利用人工标注的方法对每份文件使用政策工具不同类型的数量进行记录，使用 K-means 聚类的方法将全部样本划分为环境导向型政策和供需导向型政策。再收集各省份在政策发布当年部分经济、社会、文化特征变量，利用因子分析的方法进行降维。最后使用 Fisher 判别法、Logistic 回归建立二分类模型，并以此得到变量系数估计的正负及显著性，来解释选择不同类型返乡创业政策倾向的来源。

本研究得到结论：供需导向型政策受人口、教育、失业率的正向影响，环境导向型政策受投资率、西部地区、经济发展水平的正向影响，其中失业率因子与西部地区因子对政策类型的选择倾向有显著影响。

关键词：返乡创业　政策工具　K-means 聚类　因子分析　Fisher 判别法

目录

1 引言：背景、意义、数据、设计

1.1 研究问题

2016 年底《国务院办公厅关于支持返乡下乡人员创业创新促进农村一二三产业融合发展的意见》印发后，返乡创业政策在全社会得到广泛响应，特别是在扶贫攻坚战全面打响、乡村振兴战略深入推进的时代背景中，出现了大量的农民工、青年返乡创业的新热潮。2018 年 1 月 17 日国务院总理李克强主持召开的国务院常务会议指出，进一步支持农民工、高校毕业生和退役士兵等各类人员返乡下乡创业，有利于促进实施乡村振兴战略，推动更多人才、技术、资本等资源要素向农村汇聚，以大众创业、万众创新开辟就业新渠道、培育“三农”发展新动能。各地方政府强化主体责任，出台了大量返乡创业扶持政策，其中部分省份出台了不止一份政策文件。

本文的研究问题是：

1.不同地区、不同批次的返乡创业扶持政策中所使用的政策工具有何差异？

2.地区的哪些因素导致了这种差异？又是如何作用于省级政府返乡创业政策的选择倾向？

笔者正在攻读公共管理学院开设的行政管理学第二学位，近两学期选修了公共政策分析、社会政策概论等多门与公共政策相关的课程，对政策工具的理论非常感兴趣。

1.2 研究意义

党的十九大报告首次提出乡村振兴战略，作为解决我国“三农”问题的重要战略部署，旨在弥合城乡发展差距，推动社会共同进步、促进人民共同富裕。乡村振兴离不开既熟悉农村现实情况、又有城市现代产业部门工作经验和创业经验的返乡劳动力（王轶等，2020）[1]，同时，乡村振兴战略的提出为我国开展精准扶贫工作提供了政策支持，保障了精准扶贫与返乡创业二者之间的联动发展（刘溢海，来晓东，2020）[2]。对不同地区之间返乡创业政策工具使用差异及产生差异的内在机制进行比较研究，将对我国乡村振兴战略的具体实施、以及精准扶贫的实现提供重要的政策参考意义。

根据文献调研，笔者发现学界对于返乡创业政策的定量研究主要解决了本文提出的第 1 个问题，即不同地区政策工具使用的差异，而对第 2 个问题几乎没有相关的研究，这一定程度是由客观原因造成的。因此，本文将为此方向的研究提供一种新的思路，体现在研究设计与统计分析方法方面，具有重要的意义。

此外，在研究进程中，笔者请教了公管学院朱旭峰老师。朱老师对本文的研究设计、政策工具人工标注的标准等方面提出了许多建议。

1.3 数据来源

本研究的数据选取包括两个方面——各省返乡创业政策工具与经济、社会、文化属性等。

在返乡创业政策数据收集方面，本研究以各省级政府 2015 年后出台的返乡创业政策文本为数据来源。政策文本主要来源于各省级政府网站、北大法宝等数据库。在实际数据搜集过程中，笔者以“返乡创业”为关键词系统搜索各省级政府网站政策文本，最终整理 22 个省、自治区、直辖市有效文本 31 份[3]。部分省、自治区、直辖市并未出台返乡创业政策文件，或与本文研究目的不相符，如上海市出台关于鼓励“大众创新，

[1] 王轶，熊文，黄先开.人力资本与劳动力返乡创业[J/OL].东岳论丛，2020(03):14-28+191

[2] 刘溢海，来晓东.农民工返乡创业与精准扶贫共域性研究——基于乡村振兴战略视域[J].技术经济与管理研究，2020(01):119-123.

[3] 具体而言，包括以下省份和时间的政策文本：云南（2015）、陕西（2015）、甘肃（2015）、青海（2015）、海南（2015）、四川（2015）、黑龙江（2015）、吉林（2015）、辽宁（2015）、山西（2015）、河北（2015）、贵州（2015）、江苏（2015）、福建（2015）、广东（2016）、河南（2016）、湖北（2016）、西藏（2016）、福建（2017）、山东（2017）、湖南（2017）、河南（2017）、江西（2017）、湖北（2017）、云南（2017）、陕西（2017）、贵州（2017）、河北（2017）、吉林（2017）、辽宁（2017）、内蒙古（2017），具体政策文号及文件名称见附录部分。

万众创业”的相关政策，与“返乡”无关。

对于将政策文本转化为政策工具的方法，本研究在权衡不同工具分类方式偏好的基础上，选择最为主流的 Rothwell&Zegveld 分类模型，以及赵筱园、苏竣(2007)[4]等学者创建的公共政策三维立体框架分析理论方法，将以大众创新创业政策所涉及的基本政策工具分为供给面政策工具（记为 A）、环境面政策工具（记为 B）和需求面政策工具（记为 C）。其中，供给面政策工具可细分为人才培养、信息技术支持、资源配置、公共服务等；需求面政策工具可细分为政府采购、服务外包、市场壁垒、价格补贴等；环境面政策工具可细分为金融支持、税收优惠、法规管制、策略性措施等。

在各省经济、社会、文化属性的选择方面，本研究选择以下特征作为模型中的解释变量：（1）经济因素：①财政收入（影响对返乡创业的财政支持）；②人均 GDP；③居民人均可支配收入；④资本形成率。（2）社会因素：①常住人口数量；②地域划分：东部、中部、西部；③乡村人口数量（有理论研究证明农民等社会资本和人力资本不足的群体更倾向于独自创业）；④城镇登记失业率（失业人员选择返乡创业再就业）。（3）文化因素：①普通高等学校数；②普通高等学校招生数。这些解释变量的选择参考了朱旭峰老师的意见以及现有的学术研究。

以上数据来源为：（1）中国经济社会大数据研究平台-统计资料-统计年鉴-地区分组[5]；（2）中国经济社会大数据研究平台-数据分析-年度数据分析-选择省市县[6]；（3）国家统计局-数据查询[7]。所有数据的收集工作均由笔者手动完成，工作量较大。

1.4 研究设计

本研究主要分为三个部分：聚类、降维、建模。

首先，笔者阅读收集到的 31 份政策文本，采用人工标注的方法统计每份政策文本使用的每种政策工具类型的数量。对于每一个样本，计算供给面、需求面、环境面政策工具所占的比例，再利用 K-means 等聚类方法对样本进行二分类，获得每个样本的分类标注。

其次，对于收集到的 9 个数值型解释变量，利用因子分析的方法提取因子，达到降维和增强解释力的目的，并与各省地域因子和政策分布的时间因子结合，组成新的解释变量集。

最后，利用前两步获得的聚类标注与解释变量集，建立二分类模型。分别利用 Fisher 判别法与 Logistic 回归模型建模，观察模型的分类能力与各解释变量（因子）对分类的作用方向，以此来得到每个解释变量对省级政府返乡创业政策工具使用倾向的作用机制。

2 聚类：获得政策类型的标注

在这一节，我们来处理通过阅读政策文本、人工标注得到的各省政策工具使用的数据。

首先我们先来对每种政策工具使用的频数有大致的了解。对于全部样本，我们绘制出政策工具数量的直方图，见下页图 1，其中 A、B、C 分别代表供给面、需求面、环境面政策工具，二级编号为政策工具的二级分类。观察可得，在全部样本中，供给面和环境面政策工具使用的次数较多，而需求面政策工具使用的次数较少。

在进行聚类分析时，如果我们考虑 12 维特征来寻找相似度，效果会非常差，容易受异常值的影响，因此我们选择以政策工具的一级分类作为样本的 3 维特征。对每个样本，取每种政策工具使用的频率代替使用的频数，用以消除不同省份制定政策时，政策文本有相异的条数习惯带来的误差。而当我们将特征转化为

[4] 李江，刘源浩，黄萃，苏竣.用文献计量研究重塑政策文本数据分析——政策文献计量的起源、迁移与方法创新[J].公共管理学报，2015，12(02):138-144+159.

[5] 链接：http://data.cnki.net/Yearbook/Navi?type=type&code=A#

[6] 链接：http://data.cnki.net/YearData/Analysis

[7] 链接：http://data.stats.gov.cn/

每个样本使用不同类工具的频率后，3 个特征的量纲已经相同，在聚类时无需进行特征的标准化。

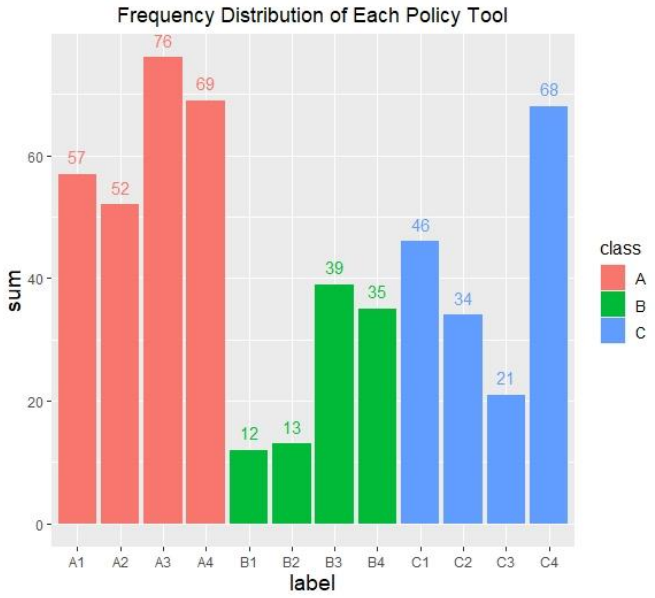

图 1 不同政策工具的频数分布

下面确定聚类的个数。我们计算不同聚类个数聚类后的组内方差，画出崖底碎石图（见图 2），推荐的聚类个数为 4。于是我们利用 K-means 方法来对上述处理后的数据进行聚类，结果见图 3。

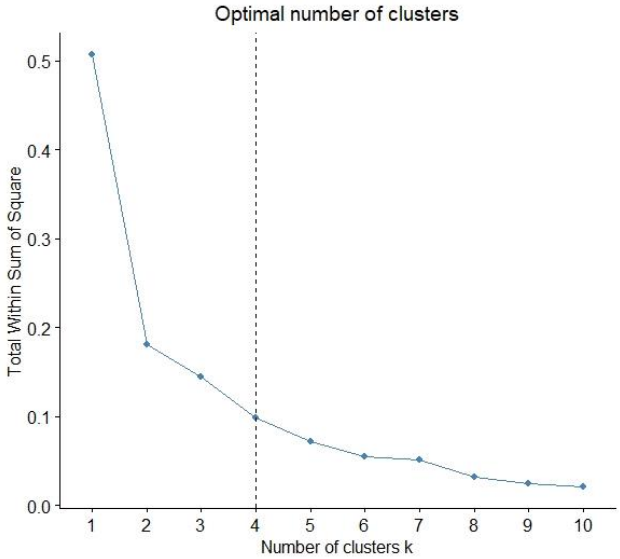

图 2 确定推荐的聚类个数

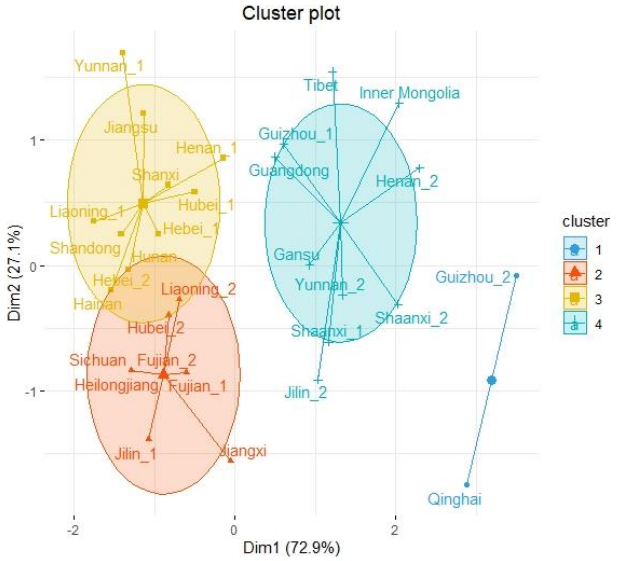

图 3 聚类个数为 4 时 K-means 聚类结果

观察图 3，实际上聚类数为 4 的聚类效果并不好，因为第 1 类的样本个数较少（仅为 2），与其他类严重不均衡，并且第 2 类和第 3 类椭圆重合度较高。基于此我们选择将聚类个数设为 2，聚类的效果和解释力将更好，且便于我们第 4 部分建立二分类模型。

因此，我们设置聚类个数为 2，分别利用 K-means 聚类和层次聚类的方法，聚类结果见下页图 4、图 5。其中，样本点之间的距离函数均取欧式距离，层次聚类的组间距离函数取 Ward's linkage。观察图 4 和图 5，两种聚类方法的结果基本相同，唯一的区别在于对江西省的聚类类别不同，其他样本结果均相同。在后文的分析中，我们取 K-means 方法的聚类结果作为两类返乡创业政策的标注方案，其中第 1 类有 19 个样本，我

们标注为“1”；第 2 类有 12 个样本，我们标注为“0”。

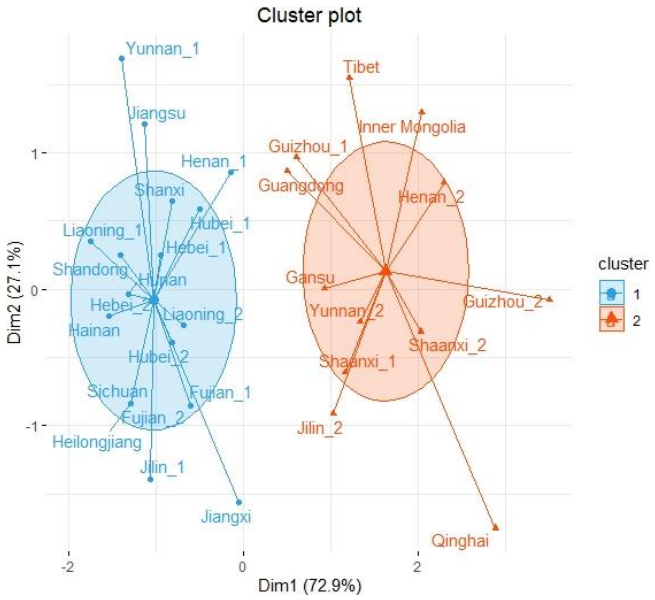

图 4 聚类个数为 2 时 K-means 聚类结果

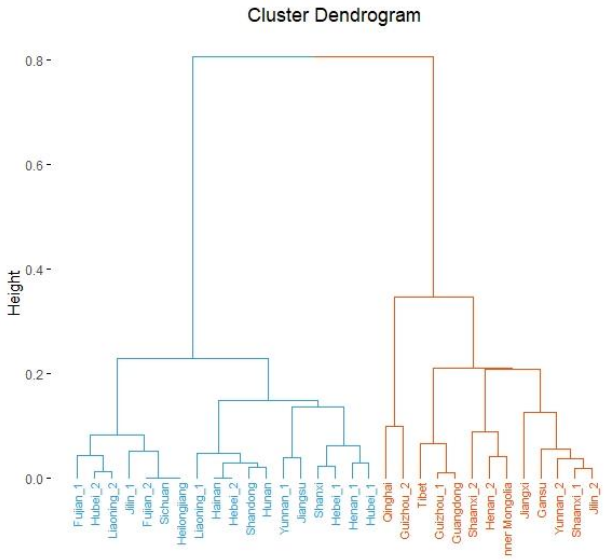

图 5 聚类个数为 4 时层次聚类结果

我们来观察上述聚成的 2 类样本在 3 个维度的特征上有什么特点。我们分别画出“0 类”与“1 类”在供给面政策工具、需求面政策工具、环境面政策工具使用频率数据的密度图，结果见图 6。

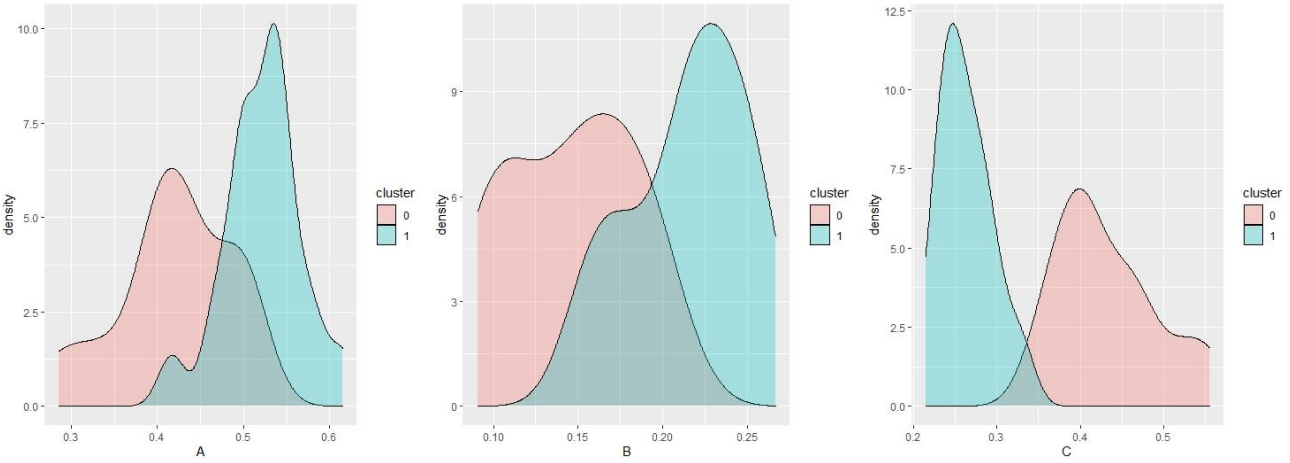

图 6 两类返乡创业政策在 3 个特征维度的密度曲线

由图 6 可以很清晰地观察到，“0 类”返乡创业政策在环境面政策工具的使用频率方面远高于“1 类”政策，而“1 类”返乡创业政策在供给面和需求面政策工具的使用频率更高。因此，我们定义“0 类”返乡创业政策为“环境导向型政策”，“1 类”返乡创业政策为“供需导向型政策”。

具体地，我们可以计算环境导向型政策与供需导向型政策在 3 个特征维度的均值：

频率均值表	环境导向型政策	供需导向型政策
供给面工具	42.06%	52.23%
需求面工具	14.58%	21.48%
环境面工具	43.36%	26.29%

表 1 两类政策使用的 3 种政策工具频率的均值

3 降维：获得维度更低的解释变量

在 1.3 节，我们一共选择了 9 个数值型变量、1 个因子型变量（地区）和 1 个控制变量（发布时间）。下面我们展示这些变量的含义、符号与基本统计量。

符号	含义	均值	方差
revenue	财政收入（元）	32828540.16	33568564.85
gdp	人均 GDP（元）	50186.03	16401.10
income	居民人均可支配收入（元）	20882.30	4879.23
per_pop	常住人口数量（万人）	5011.21	2797.07
rur_pop	农村人口数量（万人）	2361.62	1327.07
unemployment	城镇登记失业率	3.30	0.57
university	普通高等学校数（个）	91.23	40.08
student	普通高等学校招生数（万人）	27.42	15.41
capital_formation	资本形成率	65.40	19.79
符号	**含义**	**分布**	
year	政策发布年份	2015:14 2016:4 2017:13	
region	地域划分	西部 E:11 中部 M:10 东部 E:10	
cluster	政策类型	环境导向型 0:12 供需导向型 1:19	

表 2 每个变量的描述与基本统计量

由于前 9 个数值型解释变量的量纲并不相同，可以对它们分别进行标准化处理。此外，政策发布年份可以转化为 0-1 二值变量，因为在 2016 年发布的政策较少，且对于前后发布 2 个政策文件的省份，第 2 批次的政策均在 2017 年分布，转化为二值变量后还可一定程度代表政策的批次。因此我们将在 2015、2016 年发布的政策变量 year 记为 0，在 2017 年发布的政策变量 year 记为 1。

下面我们来做探索性数据分析。首先考虑政策类型（环境导向型为 0、供需导向型为 1）与部分解释变量之间的关系。分别画出不同地域划分、不同发布年份与政策类型的关系（见图 7），及以城镇登记失业率作为横坐标、资本形成率作为纵坐标得到的散点图（见图 8）。

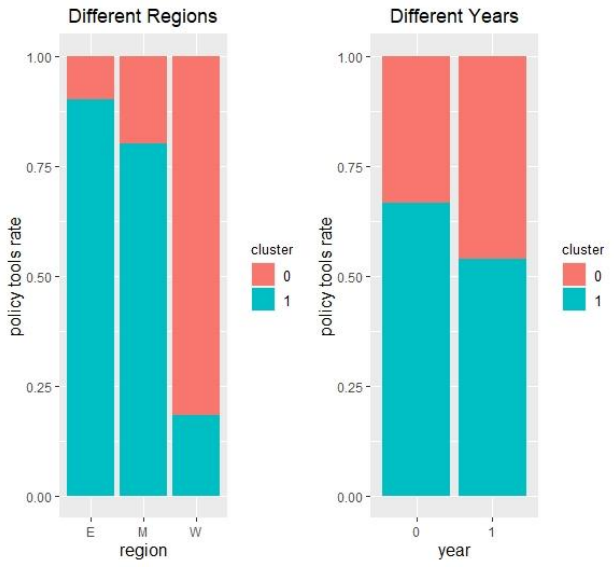

图 7 不同地域、不同年份两种政策的比例

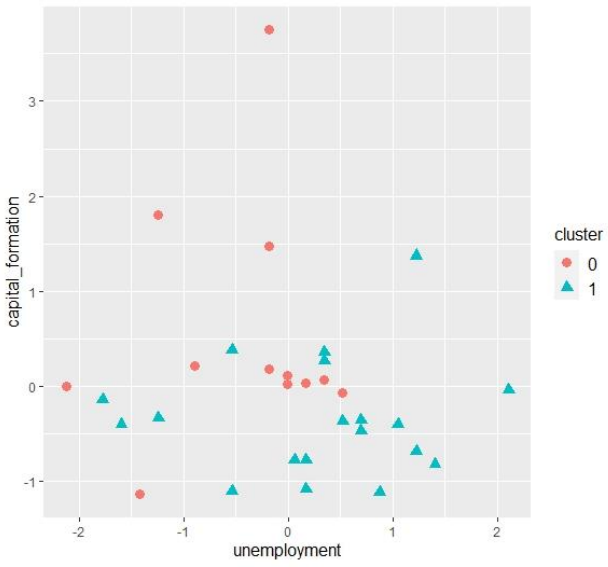

图 8 失业率与资本形成率对政策类型的影响

观察图 7，可以发现不同地域对政策类型的选择存在影响，西部地区更倾向于环境导向型政策，东部、中部地区更倾向于供需导向型政策，而政策发布的年份对政策类型基本没有影响。观察图 8，可以发现失业率对政策类型的选择存在一定影响，失业率越高的地区更倾向选择供需导向型政策，而资本形成率对政策类型似乎没有明显的影响。限于篇幅，其余解释变量与政策类别的探索性数据分析省略。

接下来我们考虑数值型解释变量之间的关系。我们画出 9 个标准化后数值型解释变量的相关系数热力图如下：

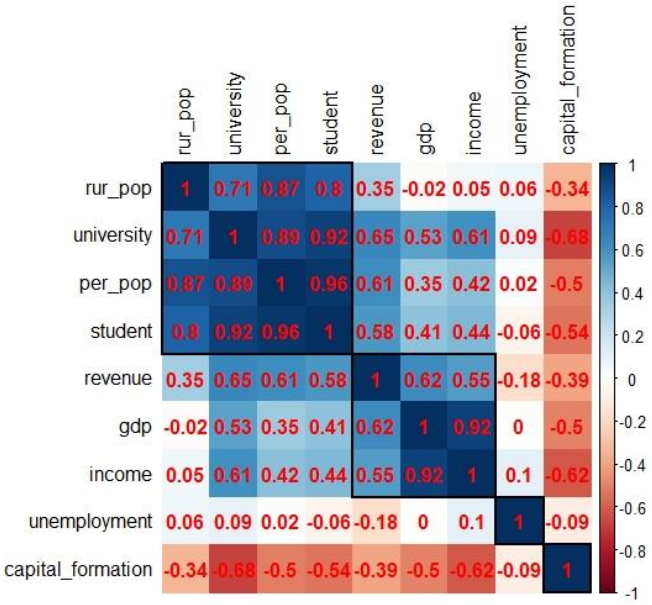

图 9 数值型解释变量相关系数热力图

由上图，容易发现许多解释变量之间的相关系数绝对值较大，有 3 对甚至达到了 0.9 以上。这提示我们需要对这 9 个解释变量进行降维处理，提取其中的潜在因子来做进一步的建模。图 9 其实还有自动的特征聚类，可以看到，解释变量可以初步归为 4 类。

进行因子分析时，我们采用 PCA 的载荷估计方法。事实上笔者曾尝试了 ML 方法与 PCA 方法，其中后者得到的因子解释力更强，对建模帮助更大。我们先画出崖底碎石图（见图 10），发现潜在因子个数应取 4。再利用 PCA 方法来估计因子载荷，设定载荷旋转为 varimax，得到成分分析树状图（见图 11）。

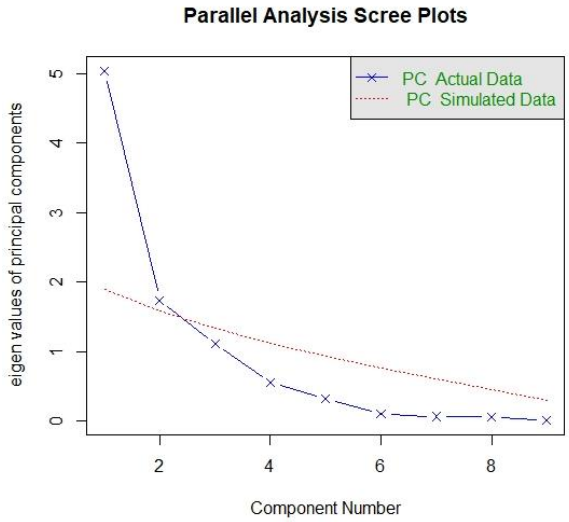

图 10 崖底碎石图

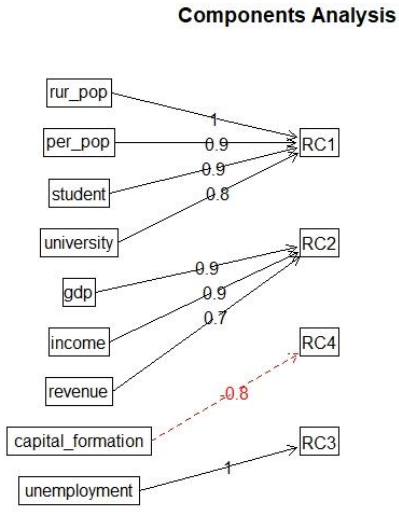

图 11 成分分析树状图

我们这里省略具体的因子载荷数值，图 11 的树状图中体现的信息足以展示因子与原来解释变量之间的关系，其中箭头上标注的数字即为因子载荷值，而对于载荷值小于 0.5 的连线进行了忽略。我们按照从上到下的顺序对每个因子进行分析与定义如下：

（1） 因子 1 主要由农村人口数量（1）、常住人口数量（0.9）、普通高等学校招生数（0.9）和普通高等学校数（0.8）4 个变量贡献，它反映的是一个省份人口数量的信息，从一般意义来讲，一个省份常住人口越多，它境内的高校数量、高校招生数也越多，因此我们将因子 1 定义为“人口因子”；

（2） 因子 2 主要由人均 GDP（0.9）、居民人均可支配收入（0.9）、财政收入（0.7）3 个变量贡献，它反映的是一个省的经济发展水平，因此我们将因子 2 定义为“经济因子”；

（3） 因子 4（RC4）主要由资本形成率（-0.8）贡献，而资本形成率的另一种说法为投资率，因此我们将因子 4 定义为“投资率因子”，同时注意到调用函数产生的投资率因子与原来的资本形成率变量呈负相关（因子载荷取负），我们在建模前将它取负变为原来的方向；

（4） 因子 3（RC3）主要由失业率（1）贡献，因此我们将因子 3 定义为“失业率因子”。

这样，我们就完成了降维的工作，得到了 4 个数值型变量（上述 4 个因子）、1 个因子型变量（地域因子）、1 个控制变量（年份因子）作为新的解释变量集。

4 建模：获得政策倾向差异的解释

在第 3 节，我们利用聚类的方法完成了对政策类别的标注，将所有样本划分为环境导向型政策和供需导向型政策；在第 4 节，我们利用因子分析的方法完成了解释变量的降维处理。下面，我们在以上工作的基础上建立二分类模型，分别考虑 Fisher 判别法与 Logistic 回归模型。

4.1 Fisher 判别法

我们先利用 Fisher 判别法对上述处理后得到的数据进行判别分析。Fisher 判别法的思想是将多维解释变量找一个方向进行投影，使两个类别尽可能在投影方向分开。我们计算得到这样的投影方向为：

$$\vec{a} = (0.15, -0.18, -0.25, 0.72, -1.06, -0.44, -2.80)'$$

分别对应（人口因子，经济因子，投资率因子，失业率因子，年份因子，中部地区因子，西部地区因子）的系数。当一个解释变量对应的投影方向系数取正，说明增大此解释变量将提高样本是“类 1”中的可能性；系数取负，说明增大此解释变量将提高样本是“类 0”中的可能性。

我们还可以画出上述 Fisher 法对每个样本的分类结果：

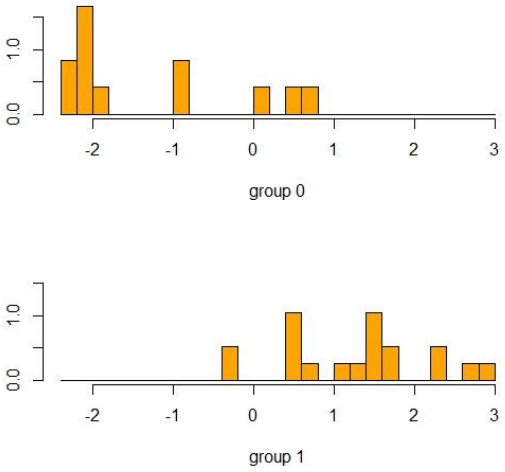

图 12 Fisher 判别法分类数值结果

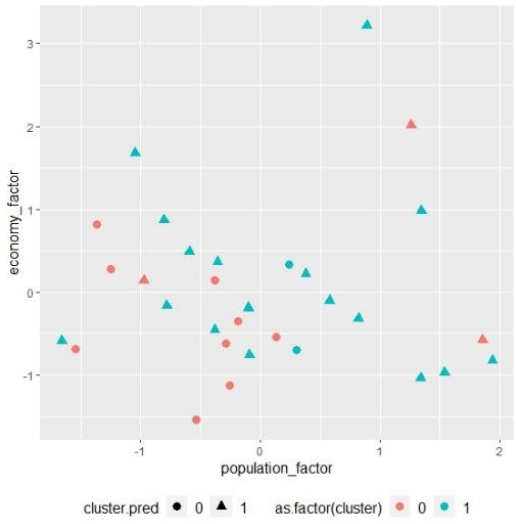

图 13 Fisher 判别法分类结果散点图

由图 12，真实类别为环境导向型政策的 12 个样本正确分类了 9 个，真实类别为供需导向型政策的 19 个样本正确分类了 17 个，表征错误率（APER）为 16.13%。图 13 通过点的颜色和形状分别代表真实类别与预测类别，可以直接观察样本的类别预测正确与否。

这里需要特别注意的是，我们没有利用交叉验证的方法来计算表征错误率以检验 Fisher 判别法的预测能力，这是因为本研究的主要目的并不在于对将来新的观测值进行预测，而在于每个解释变量对政策类别倾向的作用方向与程度；我们不要求一个预测能力很强的模型，但至少需要一个分类效果和解释能力较好的模型。此外，我们使用的数据量太小（只有 31 个样本），如果使用交叉验证则训练数据更少，模型受异常值影响大，难以建立稳定的模型，对 APER 的计算不一定准确。

根据上面我们计算得到的投影方向，每个解释变量对地方政府返乡创业政策类型选择倾向的作用方向可以做出如下简单的解释：当一个省份的经济因子越大、投资率因子越大、年份因子取 1、地域分布越偏向中西部，越倾向于制定环境导向型政策；而人口因子越大、失业率因子越大，越倾向于制定供需导向型政策。至于具体的作用机制，笔者将在第 5 节讨论。

4.2 Logistic 回归

上面我们使用了 Fisher 判别法对样本进行了判别分析，得出了每个解释变量作用方向的初步结论，但我们还希望研究解释变量对政策选择倾向的显著性。因此，下面我们建立 Logistic 回归模型来做二分类。

首先我们将所有的解释变量加入模型中，得到 Logistic 模型 1，它的细节信息表如下：

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          5.1894     2.4119   2.152   0.0314 *
population_factor    0.1944     0.7588   0.256   0.7978
economy_factor      -0.6253     0.9120  -0.686   0.4929
capital_factor      -0.6108     1.0729  -0.569   0.5691
unemployment_factor  2.2707     1.1485   1.977   0.0480 *
year                -3.8069     2.4380  -1.561   0.1184
regionM             -0.4980     2.2022  -0.226   0.8211
regionW             -6.8590     3.5046  -1.957   0.0503 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.381  on 30  degrees of freedom
Residual deviance: 15.225  on 23  degrees of freedom
AIC: 31.225
```

图 14 Logistic 模型 1 的细节信息表

类似 4.1 节，我们还可以画出模型 1 具体的分类结果：

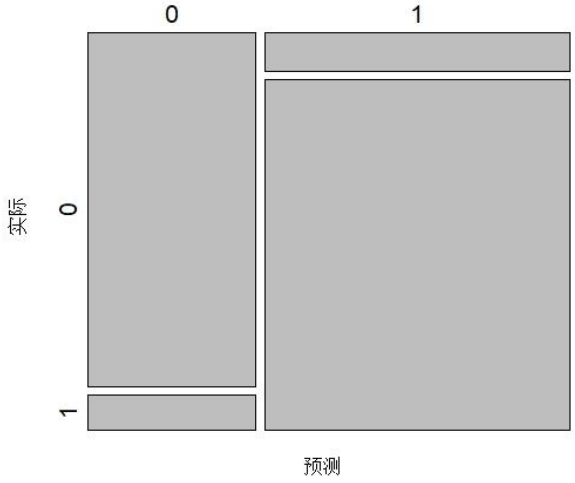

图 15 模型 1 分类实际标签与预测结果的列联表图

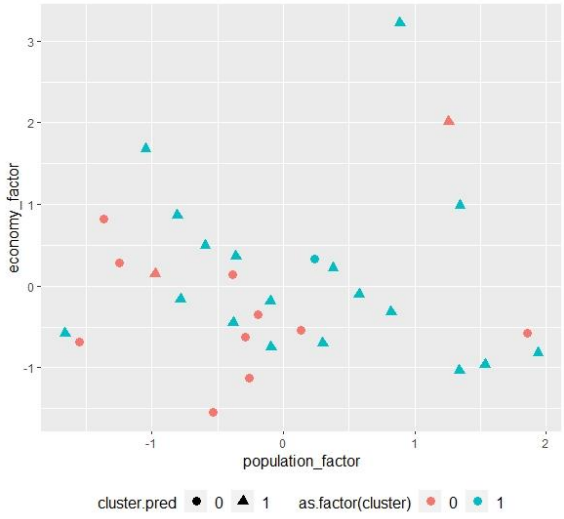

图 16 模型 1 分类结果散点图

根据真值与预测值的列联表可以得到，真实类别为环境导向型政策的 12 个样本正确分类了 10 个，真实类别为供需导向型政策的 19 个样本正确分类了 18 个，表征错误率（APER）仅为 9.68%，相比 Fisher 判别法有所降低。同样地，笔者在这里不进行交叉验证，我们更关注模型的解释，希望建立的模型能最大程度挖掘数据中的信息，也即拟合能力，而并非预测能力。

根据图 14 中模型 1 给出的各项系数及显著性，我们发现绝大部分解释变量都是不显著的，这提示我们应该对模型做变量选择。笔者这里使用逐步回归法（step-wise）来选择模型的解释变量，得到 Logistic 模型 2，它的细节信息表如下：

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.3384     1.7437   2.488   0.0128 *
unemployment_factor  2.1132     0.9690   2.181   0.0292 *
year                -3.2461     2.0233  -1.604   0.1086
regionM              0.1819     1.7181   0.106   0.9157
regionW             -5.9746     2.3353  -2.558   0.0105 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.381  on 30  degrees of freedom
Residual deviance: 16.780  on 26  degrees of freedom
AIC: 26.78
```

图 17 Logistic 模型 2 的细节信息表

我们发现，保留下面的解释变量只剩失业率因子、年份因子和地域因子，模型的 AIC 值由 31.225 下降到 26.78，但对于现有的 31 个样本，预测能力是一样的，表征错误率仍为 9.68%。对于二分类问题，除了 APER，我们还可以计算模型的 AUC 值，并绘制 ROC 曲线。

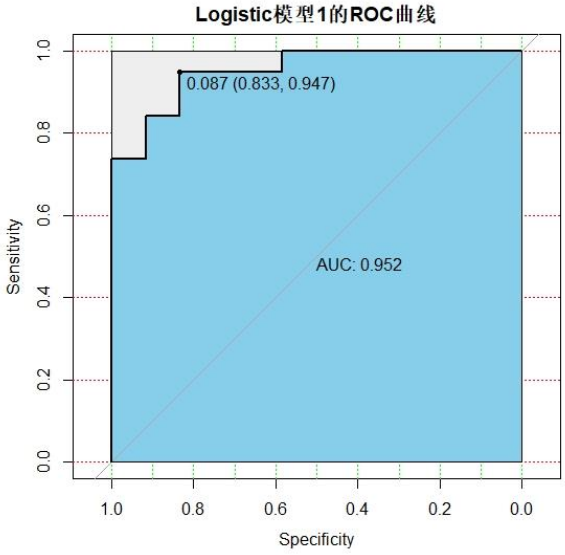

图 18 模型 1 的 ROC 曲线

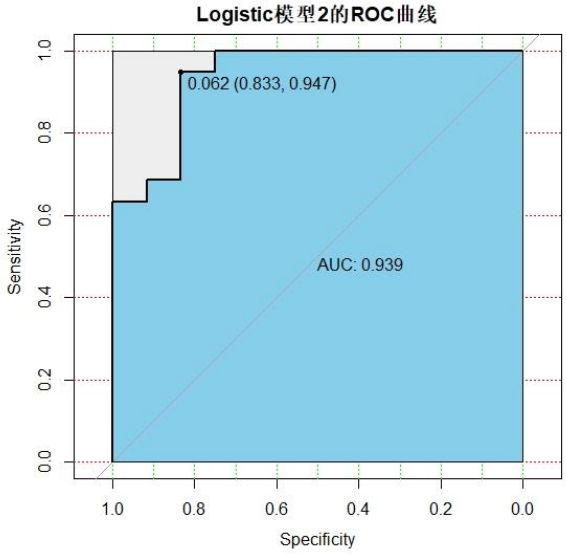

图 19 模型 2 的 ROC 曲线

可以看到模型 2 的 AUC 值相比模型 1 并未下降太多，说明预测（拟合）能力相差不大，但模型 2 的解释变量个数只有 3 个，相比之下模型 1 的 6 个解释变量有些冗余。

综上，我们完成了二分类模型的建立与初步分析的工作，其中 Fisher 判别法与 Logistic 回归模型对于每个解释变量对政策选择倾向的作用方向是一致的，其中 Logistic 模型拟合能力更好；此外，我们使用变量选择的方法得到了优化后的 Logistic 模型，利用变量系数的显著性利于进一步开展理论分析。

5 结语：结论、反思与致谢

5.1 研究结论

上述我们得到的模型 1 中许多解释变量是不显著的，这可能包括两方面原因：其一，此特征可能确实对地方政府制定返乡创业政策的使用的政策工具选择倾向基本没有影响；其二，由于笔者使用的数据样本量确实过小，可能会导致变量系数估计的标准误偏大，导致不显著。我们在分析结论是，先不考虑解释变量在模型中是否显著，只分析它们对政策类型选择的作用方向。

在这个设定下，我们对不同省份政策工具使用的差异可能的来源做出如下解释：

（1）供需导向型政策受人口、教育、失业率的正向影响。无论是需求面政策工具还是供给面政策工具，其适用对象都是以创业者为核心的“人”。我国当前返乡创业需求最大的群体包括二次就业的失业人口、高校毕业大学生、新生代返乡农民工等，而供给面政策工具的核心目标也是消除创业壁垒、吸引创业人才回流、培育本地创业人才。因此，农村人口规模较大、高等院校及学生数量较多、城镇失业率较大的省份往往更倾向于制定供需导向型政策以实现满足返乡人才创业需求、增加返乡创业人才供给的目标。

（2）环境导向型政策受投资率、西部地区因子、经济发展水平的正向影响。投资率决定了返乡创业的市场投资环境。随着国家发展政策的调整，我国资本形成率较高的地区相对来说是经济发展水平较低、基础设施建设落后的地区。因此，投资率高的地区一般是市场环境较差的地区，这些地区往往选择环境导向型政策来促进创业市场环境的改善。同理，环境导向型政策受区域位置影响较大，其根本原因在于我国东部地区经过长期发展，已经形成较为完善的创业金融、市场、政策、文化环境，而西部地区则在这些方面有待完善。

如果考虑变量的显著性，则认为失业率对选择供需导向型政策有显著正向作用，地域在我国西部在选择环境导向型政策有显著正向作用，这种差异的来源解释同上。

5.2 研究不足和改进方向

本研究取得了一定的成果，但仍存在以下较为明显的不足：

（1）最重要的一点，数据的样本量太小。这在一定程度由客观原因决定，因为笔者在研究设计中以省级政府作为研究单位，而省级政策文件的总数很有限，符合本文研究目的的政策文本范围较窄。样本量过小导致研究的结论可能不准确，容易受异常值的影响。改进的方法可以考虑以县级政府为单位，目前国家已设立了 3 批共计三百余个返乡创业试点县，将来的研究可以考虑以此为对象扩大样本量至 300 左右。

（2）本研究在对政策工具进行标注时使用阅读文本并手动记录的方法，这样可能会导致研究结论受个人因素影响，可能对少许政策条目的分类有时会模棱两可，不同研究者有不同的选择。改进的方法是使用社会科学研究中文本分析的方法，比如词频分析来替代，寻找可以量化的方法，这样也可以提高数据采集的效率，对于数百个样本用人工阅读的方法不太现实。但限于篇幅本文就不赘述了，未来的研究可以考虑。

（3）本研究忽略是一个重要的假设，相同省份在不同批次分布的政策并非完全独立的，这是一个类似时间序列的问题，不同年份发布的两份文件一定程度中是有关联的，笔者对于年份的处理过于简单；此外，对于 31 个样本的数据而言，我们选择候选的解释变量实际上有些多，模型拟合能力强不代表解释力强。

（4）笔者有意在本文基础上，考虑更多的改进方向，将来进一步完善研究成为二学位的毕业设计。

5.3 致谢

非常感谢邓老师一学期以来的教导，让我对多元统计的学习产生了浓厚的兴趣，您辛苦啦！很遗憾下学期没有您开设的课程，希望大四下我能跟您学习贝叶斯统计。同时感谢杨萱铃、孙爽两位助教为我们批改作业和答疑。在本文完成的过程中，公管学院朱旭峰老师给了我很多指导，新闻学院的张艺璇同学、社科学院的李朵同学、毛旺同学、周枢阁同学都给予了我很多帮助，弥补了我在政策文本分析和公共政策理论方面的知识缺乏，在此表示由衷的感谢。

6 附录：R 代码与政策文本条目

6.1 R 代码

注：以下的代码排版经过额外处理后比较紧凑，可读性不强，是因为易读的代码加入 6.2 节政策文本条目会超过 5 页，在这里呈现仅为证明我严格遵守了作业要求。
如果需要可读性更强的代码，可以见作业文件夹中附上的.R 文件。望您海涵，谢谢！

```
library(readxl); library(cluster); library(factoextra); library(psych); library(corrplot);
library(pROC); library(ggplot2); library(gridExtra); library(MASS)
## 载入数据
data <- read_excel("F:/学习资料/第六学期/多元统计分析/大作业/data.xlsx"); data <- as.data.frame(data)

###      聚类部分      ###
n <- nrow(data)
mydata <- data[,c(3:14)]
sum_mydata <- c()
for(i in 1:12){ sum_mydata[i] <- sum(mydata[,i]) }
lab <- c('A1', 'A2', 'A3', 'A4', 'B1', 'B2', 'B3', 'B4', 'C1', 'C2', 'C3', 'C4')
myclass <- c(rep('A',4), rep('B',4), rep('C',4))
sum_mydata <- data.frame(sum_mydata, myclass, lab)
names(sum_mydata) <- c('sum', 'class', 'label')
ggplot(data = sum_mydata, mapping = aes(x = label, y = sum, fill = class)) +
   geom_bar(stat="identity") + labs(title = 'Frequency Distribution of Each Policy Tool') +
   geom_text(aes(label = sum, vjust = -0.8, hjust = 0.5, color = class), show.legend = TRUE) +
   theme(axis.title.x = element_text(size=14), axis.title.y = element_text(size=14),
         plot.title = element_text(hjust = 0.5), legend.title = element_text(size = 12),
         legend.text = element_text(size = 10))
a <- mydata$a1 + mydata$a2 + mydata$a3 + mydata$a4
b <- mydata$b1 + mydata$b2 + mydata$b3 + mydata$b4
c <- mydata$c1 + mydata$c2 + mydata$c3 + mydata$c4
mydata <- cbind(a, b, c)
for(i in 1:n){ mydata[i,] <- mydata[i,]/(sum(mydata[i,]))}
rownames(mydata) <- data$province; df <- mydata
## 聚类数确定
fviz_nbclust(df, kmeans, method = "wss") + geom_vline(xintercept = 4, linetype = 2) + theme(plot.title = element_text(hjust = 0.5))
## K-means 聚类
set.seed(2017012136)
km_result <- kmeans(df, 4, nstart = 24); print(km_result)
fviz_cluster(km_result, data = df, palette = c("#2E9FDF", "#FC4E07", "#E7B800", "#00AFBB"),
             ellipse.type = "euclid", star.plot = TRUE,   repel = TRUE, ggtheme = theme_minimal()
) + theme(plot.title = element_text(hjust = 0.5))
```

```
## 聚成 2 类
set.seed(2017012136)
km_result <- kmeans(df, 2, nstart = 24); print(km_result)
fviz_cluster(km_result, data = df, palette = c("#2E9FDF", "#FC4E07"),
                    ellipse.type = "euclid", star.plot = TRUE,    repel = TRUE, ggtheme = theme_minimal()
) + theme(plot.title = element_text(hjust = 0.5))
## 层次聚类
result <- dist(df, method = "euclidean")
result_hc <- hclust(d = result, method = "ward.D2")
fviz_dend(result_hc, k = 2, cex = 0.6,
                k_colors = c("#2E9FDF", "#FC4E07"), color_labels_by_k = TRUE
) + theme(plot.title = element_text(hjust = 0.5))
## 观察两类的区别
mydata <- cbind(mydata, cluster = km_result$cluster); mydata <- as.data.frame(mydata)
mydata[which(mydata$cluster == 2), 'cluster'] <- 0
mean(mydata[which(mydata$cluster == 1),'a']);mean(mydata[which(mydata$cluster == 0),'a'])
mean(mydata[which(mydata$cluster == 1),'b']);mean(mydata[which(mydata$cluster == 0),'b'])
mean(mydata[which(mydata$cluster == 1),'c']);mean(mydata[which(mydata$cluster == 0),'c'])
data <- cbind(data, cluster = km_result$cluster)
data[which(data$cluster == 2), 'cluster'] <- 0
mydata$cluster <- as.factor(mydata$cluster)
names(mydata) <- c('A','B','C','cluster')
g1 <- ggplot(data = mydata, aes(x = A, fill = cluster)) + geom_density(alpha = 0.3)
g2 <- ggplot(data = mydata, aes(x = B, fill = cluster)) + geom_density(alpha = 0.3)
g3 <- ggplot(data = mydata, aes(x = C, fill = cluster)) + geom_density(alpha = 0.3)
grid.arrange(g1,g2,g3,ncol=3)

###      降维部分      ###
## 先对数据进行整理
data <- data[c(1, 15:25)]
data[which(data$year <= 2016), 'year'] <- 0
data[which(data$year == 2017), 'year'] <- 1
## 描述性统计
p <- ncol(data); describe(data[,c(1:(p-2))])
## 标准化
data[,c(2:10)] <- scale(data[,c(2:10)])
## 探索类别与部分解释变量的关系
data$cluster <- as.factor(data$cluster)
## 与区域分布的关系
g4 <- ggplot(data, aes(x=as.factor(region), fill=cluster)) + geom_bar(position = "fill") + labs(y = 'policy tools rate',
x = 'region') + labs(title = 'Different Regions') + theme(axis.title.x = element_text(size=12), axis.title.y =
element_text(size=12), plot.title = element_text(hjust = 0.5), legend.title = element_text(size = 10), legend.text =
```

```
element_text(size = 10))
g5 <- ggplot(data, aes(x=as.factor(year), fill=cluster)) + geom_bar(position = "fill") + labs(y = 'policy tools rate', x = 'year') + labs(title = 'Different Years') + theme(axis.title.x = element_text(size=12), axis.title.y = element_text(size=12), plot.title = element_text(hjust = 0.5), legend.title = element_text(size = 10), legend.text = element_text(size = 10))
grid.arrange(g4,g5,ncol=2)
## 与失业率和资本形成率的关系
ggplot(data = data, aes(x = unemployment, y = capital_formation, color = cluster, shape = cluster)) +
  geom_point(size = 3) + theme(axis.title.x = element_text(size=12), axis.title.y = element_text(size=12),
        plot.title = element_text(hjust = 0.5), legend.title = element_text(size = 12),
        legend.text = element_text(size = 12))
data$cluster <- as.numeric(data$cluster)
data[which(data$cluster == 1), 'cluster'] <- 0
data[which(data$cluster == 2), 'cluster'] <- 1
## 下面对数值型变量做探索性数据分析
cor_data <- cor(data[,c(2:10)])
corrplot(corr = cor_data, method = "color",order = "hclust", tl.col="black",addrect=4,addCoef.col = 98)
## 因子分析降维
tempdata <- data[,c(2:10)]
n <- nrow(tempdata)
fa.parallel(tempdata, n.obs = n, fa = "pc", n.iter = 100)
pc_model <- principal(tempdata, nfactors = 4, rotate='varimax'); pc_model
fa.diagram(pc_model)
findata <- cbind(pc_model$scores, data$year, data$cluster)
findata <- as.data.frame(findata); findata <- cbind(findata, data$region)
names(findata) <- c('population_factor', 'economy_factor', 'capital_factor', 'unemployment_factor', 'year', 'cluster', 'region')
findata$capital_factor <- -findata$capital_factor

###      建模部分      ###
## Fisher 判别法
L <- lda(cluster~., data = findata); L
plot(L, col = 'orange', main = "Fisher's Approach")
yhat <- predict(L, findata)$class; findata$cluster.pred <- yhat
tab1 <- table(y_true = findata$cluster, yhat, dnn=c("实际","预测")); tab1
ggplot(findata, aes(x = population_factor, y = economy_factor, col = as.factor(cluster), shape=cluster.pred)) +
  geom_point(size = 3) + theme(legend.position='bottom') + theme(axis.title.x = element_text(size=12), axis.title.y = element_text(size=12), plot.title = element_text(hjust = 0.5), legend.title = element_text(size = 12), legend.text = element_text(size = 12))
## logistic 模型
findata <- subset(findata, select = -8)
log.fit <- glm(cluster~., data = findata, family = binomial(link = logit))
```

```
summary(log.fit)
pred_fit = predict(log.fit, newdata = findata)
roc_fit <- roc(findata$cluster, pred_fit)
plot(roc_fit, print.auc=TRUE, auc.polygon=TRUE, grid=c(0.1, 0.2),grid.col=c("green", "red"), max.auc.polygon=TRUE,auc.polygon.col="skyblue", print.thres=TRUE,main='Logistic 模型 1 的 ROC 曲线')
pred_fit_in <- as.numeric(predict.glm(log.fit, type="response")>.5)
tab2 <- table(y_true = findata$cluster, pred_fit_in, dnn=c("实际","预测")); tab2
mosaicplot(t(tab2), cex=1.3, main = '', legend = TRUE)
findata$cluster.pred <- as.factor(pred_fit_in)
ggplot(findata, aes(x = population_factor, y = economy_factor, col = as.factor(cluster), shape = cluster.pred)) + geom_point(size = 3) + theme(legend.position='bottom') + theme(axis.title.x = element_text(size=12), axis.title.y = element_text(size=12), plot.title = element_text(hjust = 0.5), legend.title = element_text(size = 12), legend.text = element_text(size = 12))
## 模型选择-逐步回归法
logit.step <- step(log.fit, direction = "both")
summary(logit.step)
pred_step = predict(logit.step, newdata = findata)
roc_step <- roc(findata$cluster, pred_step)
plot(roc_step, print.auc=TRUE, auc.polygon=TRUE, grid=c(0.1, 0.2),grid.col=c("green", "red"), max.auc.polygon=TRUE,auc.polygon.col="skyblue", print.thres=TRUE,main='Logistic 模型 2 的 ROC 曲线')
pred_step_in <- as.numeric(predict.glm(logit.step, type="response")>.5)
table(y_true = findata$cluster, pred_step_in, dnn=c("实际","预测"))
```

6.2 政策文本

这一部分补充材料我们展示在各省、各批次政策工具使用数量获取了调查、阅读的政策文本。由于作业要求提供数据，整理所有的政策文本较繁琐，笔者在这里列出每份政策文本的文号、文件名称、发文时间和发布省份，以供参考。事实上，在人工标注政策工具这方面，前期政策收集的工作量是比较大的。

表 3 各级政府返乡创业政策文本统计表

序号	文号	政策文件名称	发文时间	发布部门
1	云政办发〔2015〕60 号	云南省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2015	云南省
2	陕政办发〔2015〕88 号	陕西省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2015	陕西省
3	甘政办发〔2015〕132 号	甘肃省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2015	甘肃省
4	青政办发〔2015〕241 号	青海省人民政府办公厅关于做好农民工等人员返乡创业工作的实施意见	2015	青海省
5	琼府办〔2015〕254 号	海南省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2015	海南省
6	川办发〔2015〕73 号	四川省人民政府办公厅关于支持农民工和农民企业家返乡创业的实施意见	2015	四川省
7	黑政办发〔2015〕72 号	黑龙江省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2015	黑龙江省
8	吉政办发〔2015〕70 号	吉林省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2015	吉林省
9	辽政办发〔2015〕92 号	辽宁省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2015	辽宁省
10	晋政办发〔2015〕112 号	山西省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2015	山西省

11	冀政办发〔2015〕38 号	河北省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2015	河北省
12	黔府办发〔2015〕31 号	省人民政府办公厅关于印发“雁归兴贵”促进农民工返乡创业就业行动计划的通知	2015	贵州省
13	苏政办发〔2015〕94 号	省政府办公厅关于支持农民工等人员返乡创业的实施意见	2015	江苏省
14	闽政办〔2015〕149 号	福建省人民政府办公厅关于支持农民工等人员返乡创业十二条措施的通知	2015	福建省
15	粤府办〔2016〕68 号	广东省人民政府办公厅关于进一步支持异地务工人员等人员返乡创业的通知	2016	广东省
16	豫政办〔2016〕135 号	河南省人民政府办公厅关于支持农民工返乡创业的实施意见	2016	河南省
17	鄂政办发〔2016〕10 号	省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2016	湖北省
18	藏政办发〔2016〕106 号	西藏自治区人民政府办公厅关于支持农牧民工等人员返乡创业的实施意见	2016	西藏自治区
19	闽人社文〔2017〕325 号	福建省人力资源和社会保障厅等五部门关于实施农民工等人员返乡创业培训五年行动计划（2016-2020 年）的通知	2017	福建省
20	鲁政办发〔2017〕72 号	山东省人民政府办公厅关于支持返乡下乡人员创业创新促进农村一二三产业融合发展的实施意见	2017	山东省
21	湘政办发〔2017〕113 号	湖南省人民政府办公厅关于支持农民工等人员返乡创业的实施意见	2017	湖南省
22	豫政办〔2017〕21 号	河南省人民政府办公厅关于支持返乡下乡人员创业创新促进农村一二三产业融合发展的实施意见	2017	河南省
23	赣府厅发〔2017〕18 号	江西省人民政府办公厅关于进一步支持返乡下乡人员创业创新促进农村一二三产融合发展的实施意见	2017	江西省
24	鄂政办发〔2017〕73 号	湖北省人民政府办公厅关于大力支持返乡下乡人员创业创新促进农村一二三产业融合发展的实施意见	2017	湖北省
25	云政办发〔2017〕45 号	云南省人民政府办公厅关于支持返乡下乡人员创业创新促进农村一二三产业融合发展的实施意见	2017	云南省
26	陕政办发〔2017〕46 号	陕西省人民政府办公厅关于支持返乡下乡人员创业创新促进农村一二三产业融合发展的实施意见	2017	陕西省
27	黔府办发〔2017〕37 号	省人民政府办公厅关于进一步支持返乡下乡人员创业创新促进农村一二三产业的实施意见	2017	贵州省
28	冀政办字〔2017〕131 号	河北省人民政府办公厅关于支持返乡下乡人员创业创新促进农村一二三产业融合发展的实施意见	2017	河北省
29	吉政办发〔2017〕36 号	吉林省人民政府办公厅关于启动农民工等人员返乡创业工程促进农民增收的实施意见	2017	吉林省
30	辽政办发〔2017〕41 号	辽宁省人民政府办公厅关于支持返乡下乡人员创业创新促进农村一二三产业融合发展的实施意见	2017	辽宁省
31	内政办发〔2017〕124 号	内蒙古自治区人民政府办公厅关于支持返乡下乡人员创业创新促进农村牧区一二三产业融合发展的实施意见	2017	内蒙古

资料来源：各省级政府网站、北大法宝等数据库