2013 年 9 月 数理统计与管理 Sep. 2013
第 32 卷 第 5 期 Journal of Applied Statistics and Management Vol. 32 No. 5

**文章编号**: 1002-1566 (2013) 05 - 0761 - 08

# 中国区域创新能力静态分析 —— 基于自适应赋权主成分聚类模型

朱建平 [1,2] 王德青 [1,2] 方匡南 [1,2]

(1. 厦门大学经济学院统计系, 福建 厦门 361000, 2. 厦门大学数据挖掘研究中心, 福建 厦门 361000)

**摘要**: 本文在对经典聚类模型和现有改进聚类模型优点与不足剖析的基础上, 通过定义客观加权主成分距离为分类统计量, 提出了一种自适应赋权的主成分聚类模型。与现有同类方法相比, 新模型克服了指标之间的高度共线性, 能够对指标重要性的客观差异进行自适应赋权, 每一步都有充分的理论保证其必要性、合理性。应用加权主成分聚类对中国区域创新能力进行集团划分, 分类结果的可解释性明显提高, 统计检验效果显著, 所得的结论对了解和推动中国区域创新能力发展具有借鉴意义。

**关键词**: 分类; 自适应赋权; 加权主成分聚类; 区域创新能力

**中图分类号**: C812, O212 **文献标识码**: A

# Static Analysis of China's Regional Innovation Capability — Based on Adaptive Weighting Principal Component Cluster Model

ZHU Jian-ping[1,2] WANG De-qing[1,2] FANG Kuang-nan[1,2]

(1. Department of Statistics, School of Economy, Xiamen University, Fujian Xiamen 361005, China,
2. Data Mining Research Center, Xiamen University, Fujian Xiamen 361005, China)

**Abstract**: Based on summation of advantages and defects of classical cluster models and existed improved cluster models, this paper puts forward an adaptive weighting principal component cluster model by defining objective weighted principal component distance as classification statistic. Compared with existed similar methods, the new model not only solves high colinearity among variables, but also weights adaptively according to their objective importance, so the new model deserves sufficient theoretical basis for its necessity and rationality at every stages. At last this paper applies weighted principal component cluster model on the division of China's regional innovation capability, and gets highly improved explanation results as well as significant statistical test. Meanwhile the conclusion provides significant reference to understand and promote China's regional innovation capability.

**Key words**: classification, adaptive weighting, weighted principal component cluster, regional innovation capability

## 0 引言

随着现代数据存储技术的飞速发展, 海量数据库的内在规律愈加复杂难辨。在对海量数据分类挖掘时, 传统的系统聚类和 K 均值聚类等经典模型面临诸多局限。事实上, 每一经

**收稿日期**: 2012 年 10 月 16 日 **收到修改稿日期**: 2013 年 6 月 13 日

**基金项目**: 国家社会科学基金项目 (编号: 11BTJ001); 国家自然科学基金青年项目 (编号: 710201139); 全国统计科学研究计划重大项目 (编号: 2012LD001)。

典的聚类分析技术都是有针对性地分析数据中的某类规律，如果忽略模型的适用前提和待聚类对象的具体特点，简单地套用传统聚类技术难以取得理想的分类效果。关于如何解决传统聚类分析处理海量数据时凸现的弊端，国内外学者做了许多有益的探讨。Michel Mouchart 和 JeroenV.K.Romouts[1] 将逐步回归方法与聚类分析结合，解决了数据缺失的单指标面板数据聚类问题；Shia Ben-Chang 等 [2] 在 Fisher 典型判别分析基础上，将模糊聚类理论引入判别分析中，提出了一种能够处理模糊现象分类的聚类模型；郑兵云 [3] 详细分析多指标面板数据的数据格式，通过重构面板数据的相似性测度拓展了面板数据的聚类分析方法；任娟 [4]、肖泽磊等 [5] 从多元统计分析理论角度提出多指标面板数据的三维信息融合技术，改进了多指标面板数据的因子分析和系统聚类，弥补了单一分析的片面性和局限性；毛国敏 [6]、孙锐 [7]、袁建新 [8]、陆根尧 [9] 和庞丽 [10] 等通过主成分因子分析将原始指标体系压缩为少数综合指标，以综合指标代替原始指标进行聚类，一定程度上拓展了经典聚类分析的应用。综观国内外近年来关于分类问题的研究文献发现，聚类分析在自然科学领域和社会科学领域均有广泛的应用，但由于经典聚类分析的假定前提比较苛刻，实际应用中存在较大的局限性。因此，如何科学地借鉴其它分类方法的优点以弥补传统聚类模型的缺陷是分类数据挖掘的研究热点。

聚类分析的分类质量取决于两个核心问题：一是用什么指标构建聚类统计量以表征样本之间的相似程度；二是采用何种具体聚类技术对样本进行类属划分 [3,11]。从理论角度看，现有聚类技术的改进研究都是从所研究的具体问题出发，鲜有方法论框架下的聚类技术理论分析，改进的聚类模型缺少理论基础，普遍适应性较差。从应用角度看，由于没有系统、规范的理论指导，聚类过程缺乏科学性，分类结果难以合理解释。基于以上认识，本文对经典聚类模型和改进聚类技术的优势与不足进行梳理，提出一种无先验知识的自适应赋权聚类方法，并应用该方法对 2011 年中国省级区域的创新水平进行静态的集团划分，为更深层次推动区域创新能力发展提供借鉴。

## 1 加权主成分聚类分析

### 1.1 传统聚类分析的不足

传统的聚类分析多是基于样本 (指标) 之间距离 (相关系数) 的亲疏关系进行分类，相似性度量不仅取决于指标之间的亲疏程度，而且依赖于指标重要性的内在差异，因此，用于构建聚类统计量的指标选择至为重要。传统的聚类算法要求描述样本的指标重要性相同并且彼此独立，然而对于复杂的海量数据库，系统层次结构的指标体系中各指标重要性相差悬殊，指标之间不可避免地信息重叠。如果对存有高度共线性的指标不加处理直接聚类，那么聚类统计量将同类指标重复计算，过于放大共线性指标的作用而淹没独立性指标的贡献，导致难以解释的分类结果。应用传统聚类模型处理实际分类问题时，为了克服指标体系的高度共线性往往是定性分析指标之间的机理关系，主观剔除信息重叠的指标以达到聚类指标彼此独立的目的，同时通过专家打分赋予不同指标相应的权重以体现指标重要性的差异。显然，定性地筛选指标和主观赋权需要对每一指标的实际意义有深入的了解，并且要求分析者具有相关的领域知识和客观公正的赋权标准，这在实际应用中难以保证。

### 1.2 主成分聚类分析及其不足

主成分分析是降低数据空间维数的重要方法，其分析结果是将原始错综复杂的指标体系通过线性变换转化为少数相互独立的主成分综合指标，并且要求低维主成分空间能够体现原

始指标体系的绝大部分信息。受主成分分析方法的启发，文献 [6-7] 将主成分分析与聚类分析集成，即应用主成分分析克服原始指标之间的共线性影响，然后用少数主成分代替原始指标进行聚类 (为叙述方便，下文称上述方法为主成分聚类分析)。值得肯定的是，主成分聚类克服了传统聚类模型不能处理指标之间高度共线性的缺点，但应该注意到，不同主成分体现原始指标体系信息的能力 (方差贡献率) 往往相差悬殊。极端情形下，如果忽略不同主成分重要性的客观差异，不加区别地直接将主成分代替原始指标聚类则必然会影响主成分聚类分析的准确性 [12−13]。

为了体现主成分重要性的差异，文献 [8–10] 提出以各个主成分的方差贡献率大小为权重构造主成分综合得分，然后以主成分综合得分代替原始指标聚类分析 (下文称上述方法为主成分综合得分聚类)。主成分综合得分融合了多个主成分的信息含量，重要性的差异也通过方差贡献率的客观赋权得到体现。但这种看似合理的信息融合方法未必能提高聚类的效率，甚至可能会降低聚类质量。

事实上，设 $F_1, F_2, \cdots, F_m$ $(m \le p)$ 为由 $p$ 维指标向量 $X = (X_1, X_2, \cdots, X_p)^T$ 提取的主成分，$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$ 为主成分 $F_i$ $(i = 1, 2, \cdots, p)$ 对应的特征值，定义 $\alpha_i = \lambda_i / \sum_{j=1}^{p} \lambda_j$ 为主成分 $F_i$ 的方差贡献率，则主成分综合得分为：

$$F = \alpha_1 F_1 + \alpha_2 F_2 + \cdots + \alpha_m F_m. \tag{1}$$

由于 $F_i$ 与 $F_j$ $(i \ne j; i, j = 1, 2, \cdots, m; m \le p)$ 不相关，所以有

$$\mathrm{Var}(F) = \sum_{i=1}^{m} \alpha_i^2 \mathrm{Var}(F_i) = \sum_{i=1}^{m} \alpha_i^2 \lambda_i \le \lambda_1 \sum_{i=1}^{m} \frac{\lambda_i^2}{(\sum_j^p \lambda_j)^2} = \lambda_1 \frac{\sum_1^m \lambda_i^2}{(\sum_j^p \lambda_j)^2} < \lambda_1 = \mathrm{Var}(F_1),$$

也即综合得分的信息含量小于第一主成分的信息含量，可见 (1) 式主成分综合得分的聚类效率未必优于第一主成分的聚类效率。

### 1.3 加权主成分聚类分析

影响经典聚类分析分类质量的因素多种多样，其中指标之间的高度共线性 (信息重叠) 和不同指标重要性存在的客观差异是典型的两个影响因素。现有关于经典聚类分析拓展的研究文献多数只解决了上述问题之一，改进聚类模型的分类质量是否一定优于经典聚类分析需要理论分析和实践应用的双重检验。对经典聚类分析的拓展应该综合考虑共线性影响和重要性差异两个方面，基于上述分析，借鉴主成分聚类的思想，本文按照方差贡献率大小对不同主成分进行自适应赋权以体现其聚类效率的差异。

**定义** 设 $F_i, \lambda_i, \alpha_i$ $(i = 1, 2, \cdots, m; m \le p)$ 分别为主成分向量及其对应的特征根和方差贡献率，令 $\beta_k = \alpha_k / \sum_i^m \alpha_i$ $(k = 1, 2, \cdots, m)$ 为主成分 $F_k$ 的距离权重。称

$$d_{ij}(q) = \Big( \sum_{k=1}^{m} \big(\beta_k |F_{ik} - F_{jk}|\big)^q \Big)^{\frac{1}{q}}, \tag{2}$$

为样本 $i, j$ 之间的加权主成分距离。

聚类距离的定义需要满足正定性、对称性和三角不等式 [14]，不难证明公式 (2) 满足上述三条性质。与现有聚类分析改进的研究成果相比，加权主成分聚类的核心优势在于同时克服了经典聚类分析存在的两个典型缺陷：(1) 通过主成分的特征提取剔除了原始指标体系高度的重叠信息；(2) 每一主成分的距离权重 $\beta_k$ 来源于原始指标数据，体现了不同主成分聚类效率

的差异，赋权准则客观合理。应该注意到，当不同主成分信息含量相差不大时，加权主成分聚类等同于普通主成分聚类，也即文献 [6–10] 中的改进方法为本文的特例。本文拓展的自适应赋权主成分聚类分析每一步都有充分的理论保证其必要性、合理性，具体步骤如下：

**步骤 1** 计算原始指标体系的相关系数矩阵，对比 KMO 值与显著性水平的临界值以判断主成分分析的可行性；

**步骤 2** 比较原始指标取值数量级的差异程度，决定分析对象是相关系数矩阵或是协方差矩阵；

**步骤 3** 进行主成分分析，提取方差贡献率并按 $\beta_k$ 的定义式计算距离权重，以 (2) 式为相似性测度进行聚类，结合实际情况确定最终的分类结果。

主成分本质上是原始指标信息核心特征的体现。应用加权主成分聚类模型解决实际分类问题时，为了排除次要信息干扰以简化数据运算复杂程度，通常提取方差贡献最大的前 $m$ 个主成分即可。但当待分类样本的相似度较高，为了提高分类的准确性，需要选取全部主成分进行聚类分析。

客观公正地评判模型的分类质量是困难而复杂的问题，目前没有评判所有聚类模型有效性的统一标准。在众多的评判标准中，比较客观的是将聚类模型的分类结果与预先已知的本来类属进行对比，以错分率为标准判断不同聚类模型的优劣。为验证拓展聚类模型的有效性，本文选用 Fisher R A 在 1936 年分类研究使用的鸢尾花数据为标准测试数据 [15]，三类鸢尾花的属性特征由花瓣长度、花瓣宽度、萼片长度、萼片宽度四个指标刻画。计算原始数据的 KMO 值为 0.599，表明原始指标之间的信息高度重叠。提取信息含量最大的前两个主成分方差贡献率分别为 91.89% 和 5.57%，分别用不同的聚类模型进行分类，对比结果如表 1 所示。

**表 1** 聚类模型分类效率对比

| 聚类方法 | 经典聚类模型 | | | 主成分聚类模型 | | | 加权主成分聚类模型 | | |
|---|---|---|---|---|---|---|---|---|---|
| 分类结果 | 植物品种 | | | 植物品种 | | | 植物品种 | | |
| | 刚毛 | 变色 | 弗吉尼亚 | 刚毛 | 变色 | 弗吉尼亚 | 刚毛 | 变色 | 弗吉尼亚 |
| 1 | 50 | 0 | 0 | 49 | 0 | 0 | 50 | 0 | 0 |
| 2 | 0 | 35 | 1 | 1 | 36 | 19 | 0 | 43 | 3 |
| 3 | 0 | 15 | 49 | 0 | 14 | 31 | 0 | 7 | 47 |
| 合计 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| 错分率 % | 10.67 | | | 22.7 | | | 6.67 | | |

注：由于鸢尾花类属预先已知，本文统一采用 $q = 2$ 的 K-means 聚类准则。

从表 1 的结果对比可知，三种聚类模型的分类质量存在明显差异。以错分率为评判模型分类质量优劣的标准，依次是加权主成分聚类模型、经典聚类模型和主成分聚类模型，印证了本文拓展模型的优效性。需要注意的是，主成分聚类模型分类质量明显劣于本文拓展的加权主成分聚类模型，甚至显著低于传统的经典聚类模型。结合主成分分析结果，究其原因在于两个主成分的方差贡献率相差悬殊，即二者体现原始数据的信息能力存在显著差异，等同地直接以主成分对鸢尾花进行聚类分析，则削弱了第一主成分的分类效率而导致不理想的分类结果。

## 2 基于加权主成分聚类的区域创新能力再评价

### 2.1 实证对象的确定及数据来源

在经济全球化的背景下，区域创新能力日益成为国际竞争优势的决定性因素。中国的区域创新能力相差较大 [16]，从区域角度分析比较省级创新能力的异同具有重要的现实意义 [17]。

习惯上的东部、中部和西部的定性划分带有主观任意性，缺乏科学的定量分析基础。如果就每个省 (区) 市进行研究，结果只能反映单个省 (区) 市的个例特征，难从总体上把握区域之间创新发展的不平衡分布状态，同时也忽略了相关省 (区) 市之间创新能力的有机联系。所以有必要对省 (区) 市的创新能力进行定量的集团聚类，综合比较分析以加深对各类区域技术创新能力的认识，为整体上把握技术创新活动、科学制定创新政策提供依据。基于上述考虑，本文以《中国区域创新能力报告 2011》(以下简称《报告》) 综合指标数据为基础，综合运用新提出的加权主成分聚类及其它多种聚类模型，深层次挖掘隐藏在现有研究成果背后的潜在信息和内在机理。

### 2.2 数据处理及结果分析

《报告》从知识创造 ($X_1$)、知识获取 ($X_2$)、企业创新 ($X_3$)、创新环境 ($X_4$) 和创新绩效 ($X_5$) 五个方面表征地区的整体创新能力，数据分析发现五个指标的相关系数介于 0.7–0.82，KMO 值为 0.832，表明《报告》中的五个指标之间存有高度的共线性，而且上述五个方面对形成区域整体创新能力的贡献程度也存在差异 [17]。破坏了经典聚类模型有效应用的前提假设。依照本文拓展聚类模型的具体步骤，主成分分析的结果汇总如表 2 所示。

**表 2** 主成分因子分析结果

| 主成分 | 特征根 | 方差贡献 (%) | 权重 | 因子载荷 | | | | | 因子命名 |
|---|---|---|---|---|---|---|---|---|---|
| | $\lambda_i$ | $\alpha_i$ | $\beta_i$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | |
| $F_1$ | 4.065 | 81.306 | 0.9242 | 0.395 | 0.604 | 0.836 | 0.796 | 0.855 | 创新因子 |
| $F_2$ | 0.333 | 6.668 | 0.0758 | 0.895 | 0.679 | 0.427 | 0.466 | 0.391 | 知识因子 |

由表 2 可知，前两个主成分的方差贡献率已达 87.97%，基本体现了原始指标数据的核心信息。结合因子载荷矩阵发现，主成分 $F_1$ 在企业创新、创新环境和创新绩效三个指标上的载荷较大，主要反映创新能力形成的基础及背景，将其命名为创新因子；$F_2$ 在知识创造、知识获取指标上的载荷较大，主要反映知识因素对创新能力的贡献，将其命名为知识因子。应该注意到，两个主成分的信息含量相差 12 倍多，如果忽略主成分之间客观存在的悬殊差异，直接以主成分代替原始指标聚类，分类结果的准确率有待商榷。仿照《报告》将省 (区) 市分为 5 类的研究思路，本文统一以 $q = 2$ 的欧氏距离为样本相似性测度，采用离差平方和法将 31 个省 (区) 市分为 5 类，不同聚类模型的分类结果对比如表 3 所示。

由于没有预先定义的类别标准来表明数据集中哪种期望关系是有效的，评价聚类模型的有效性必须定量分析和定性分析综合考虑。“可解释性” 是评判分类效率的重要依据，聚类模型的优劣首先表现在能否对聚类结果做出合理的解释。表 3 的分类结果显示，六种聚类模型基本都能将江苏、广东、北京、上海、浙江、山东与其它省 (区) 市分开，原因在于上述六省 (区) 市各项创新能力指标值都远远领先其它省 (区) 市，类别界限明显。但其余 25 个省 (区) 市的创新指标取值相差不大，聚类空间狭窄，六种聚类模型的分类结果存在较大差异，具体表现为第四、五类的样本数目较多，各样本归属于哪一类的规律性不明显。若以每一类中样本数目的均匀性来看，加权主成分聚类分析优于其它聚类模型。特别的是，第一主成分聚类与加权主成分聚类的分类结果完全一致，究其原因在于第一主成分的权重远远高于第二主成分，按 (2) 式计算的聚类统计量中第一主成分占绝对的主导作用，在此极端情况下，加权主成分聚类等同于第一主成分聚类。尤为引起注意的是，普通主成分聚类将浙江、山东、天津与江苏、广东归类为创新水平高于北京、上海的领先集团。结合《报告》中的原始数据发现，除知识创造

指标外，江苏和广东的其它指标都领先上海、北京，但浙江、山东和天津的多数指标落后于北京、上海，三个省市的综合排名也落后于北京、上海，所以将浙江、山东和天津划分为领先于北京、上海的集团难以解释。出现上述极端聚类结果的原因在于，普通主成分聚类未曾区分不同主成分因子的聚类效率，导致难以解释的分类结果。

**表 3** 省 (区) 市创新能力聚类分析结果

| 聚类模型 | 第一类 | 第二类 | 第三类 | 第四类 | 第五类 |
|---|---|---|---|---|---|
| 原始指标 | 江苏 | 北京 | 浙江 | 辽宁 四川 重庆 | 河南 江西 河北 |
| | | | 山东 | 湖南 陕西 湖北 | 黑龙江 山西 内蒙古 |
| 聚类分析 | 广东 | 上海 | 天津 | 福建 安徽 吉林 | 海南 甘肃 贵州 云南 |
| | | | | 广西 | 新疆 宁夏 青海 西藏 |
| 普通主成分 | 江苏 广东 | 北京 | 辽宁 四川 重庆 | 福建 安徽 | 山西 甘肃 贵州 |
| | 浙江 | | 湖南 陕西 湖北 | 吉林 西藏 | 云南 新疆 宁夏 |
| 聚类分析 | 山东 天津 | 上海 | 河南 江西 河北 | 内蒙古 | 青海 海南 |
| | | | 广西 | | 黑龙江 |
| 第一主成分 | 江苏 | 浙江 | 辽宁 四川 | 安徽 吉林 | 广西 山西 甘肃 |
| | 广东 | 山东 | 重庆 湖南 | 河南 江西 | 贵州 云南 新疆 |
| | 北京 | 天津 | 陕西 湖北 | 河北 黑龙江 | 宁夏 青海 海南 |
| 聚类分析 | 上海 | | 福建 | 内蒙古 | 西藏 |
| 主成分综合 | 江苏 | 浙江 | 辽宁 四川 | 安徽 吉林 河南 江西 | 宁夏 |
| | 广东 | 山东 | 重庆 湖南 | 河北 黑龙江 广西 | 青海 |
| | 北京 | 天津 | 陕西 湖北 | 山西 甘肃 贵州 云南 | 西藏 |
| 得分聚类分析 | 上海 | | 福建 | 新疆 海南 内蒙古 | |
| 加权主成分 | 江苏 | 浙江 | 辽宁 四川 | 安徽 吉林 | 广西 山西 甘肃 |
| | 广东 | 山东 | 重庆 湖南 | 河南 江西 | 贵州 云南 新疆 |
| | 北京 | 天津 | 陕西 湖北 | 河北 黑龙江 | 宁夏 青海 海南 |
| 聚类分析 | 上海 | | 福建 | 内蒙古 | 西藏 |
| 《创新报告》 | 江苏 | 北京 | 浙江 | 天津 辽宁 四川 | 江西 黑龙江 广西 |
| | | | | 重庆 湖南 陕西 | 内蒙古 山西 甘肃 |
| | | | | 湖北 福建 安徽 | 贵州 云南 新疆 宁夏 |
| 聚类分析 | 广东 | 上海 | 山东 | 吉林 河南 河北 | 青海 西藏 海南 |

特定问题的结构影响特定聚类算法的性能，聚类模型的分类质量除了能科学合理地解释外，还必须通过定量分析的统计检验。显然地，如果模型分类效果显著，则同一类内样本的指标离差较小，而类与类之间样本的指标离差较大。为了比较各种聚类模型的相对优越性和稳定性，本文进一步做方差分析，结果如表 4 所示。

表 4 中的对比结果显示，若以 F 值大小为标准：(1) 普通主成分聚类的分类效果明显劣于其它聚类模型，特别的，在主成分信息含量相差悬殊的极端情况下，普通主成分聚类的分类质量甚至低于经典聚类分析，这再次说明忽略不同主成分重要性的差异，等权地以主成分代替原始指标直接聚类未必一定提高聚类质量。(2) 第一主成分聚类结果的各项指标 F 值均大于主成分综合得分的聚类结果，印证了第一主成分的聚类效率优于主成分综合得分的聚类效率；(3) 加权主成分聚类结果的指标 F 值大于普通主成分聚类和主成分综合得分聚类结果 F 值，说明加权主成分聚类模型较现有改进聚类模型的分类效果明显提高。(4) 相比《报告》的分类方法，加权主成分聚类通过降维不仅使复杂的分类问题简化，而且以方差贡献率对参与聚类的主成分客观赋权，避免了《报告》中专家赋权的主观任意性，依此所做的结论建议更客观、可信。

综合表 3 的分类结果和表 4 的统计检验发现，本文提出的加权主成分聚类对样本的区分度更高，所分类别之间差异的统计检验效果更显著。因此，本文的研究结论可以作为《报告》的有益参考和补充，有助于更深层次对区域创新能力的分布做出综合判断。

表 4 六种聚类模型方差分析对比

| 聚类模型 | 创新指标 | 平均类间离差平方和 | 平均类内离差平方和 | 平均总离差平方和 | F 值 |
|---|---|---|---|---|---|
| 原始指标聚类分析 | $X_1$ | 1071.251 | 39.639 | 177.187 | 27.025 |
| | $X_2$ | 887.484 | 42.242 | 156.141 | 21.009 |
| | $X_3$ | 1511.958 | 21.410 | 220.149 | 70.621 |
| | $X_4$ | 390.362 | 17.690 | 67.380 | 22.066 |
| | $X_5$ | 659.556 | 23.554 | 108.355 | 28.001 |
| 普通主成分聚类分析 | $X_1$ | 1036.905 | 44.923 | 177.187 | 23.082 |
| | $X_2$ | 798.539 | 55.926 | 156.141 | 14.278 |
| | $X_3$ | 1344.709 | 47.140 | 220.149 | 28.526 |
| | $X_4$ | 372.556 | 20.430 | 67.380 | 18.236 |
| | $X_5$ | 613.266 | 30.676 | 108.355 | 19.992 |
| 第一主成分聚类分析 | $X_1$ | 1111.850 | 33.393 | 177.187 | 33.296 |
| | $X_2$ | 941.134 | 33.989 | 156.141 | 27.690 |
| | $X_3$ | 1405.074 | 37.853 | 220.149 | 37.119 |
| | $X_4$ | 413.369 | 14.151 | 67.380 | 29.212 |
| | $X_5$ | 675.412 | 21.115 | 108.355 | 31.987 |
| 主成分综合得分聚类分析 | $X_1$ | 1091.525 | 36.520 | 177.187 | 29.888 |
| | $X_2$ | 913.927 | 38.174 | 156.141 | 23.941 |
| | $X_3$ | 1384.406 | 41.033 | 220.149 | 33.739 |
| | $X_4$ | 389.676 | 17.796 | 67.380 | 21.897 |
| | $X_5$ | 661.544 | 23.249 | 108.355 | 28.455 |
| 加权主成分聚类分析 | $X_1$ | 1111.850 | 33.393 | 177.187 | 33.296 |
| | $X_2$ | 941.134 | 33.989 | 156.141 | 27.690 |
| | $X_3$ | 1405.074 | 37.853 | 220.149 | 37.119 |
| | $X_4$ | 413.369 | 14.151 | 67.380 | 29.212 |
| | $X_5$ | 675.412 | 21.115 | 108.355 | 31.987 |
| 《创新报告》聚类分析 | $X_1$ | 1091.833 | 36.473 | 177.187 | 29.936 |
| | $X_2$ | 845.83 | 48.651 | 156.141 | 17.386 |
| | $X_3$ | 1428.640 | 34.228 | 220.149 | 41.739 |
| | $X_4$ | 432.610 | 11.191 | 67.380 | 38.658 |
| | $X_5$ | 666.904 | 22.424 | 108.355 | 29.740 |

注：表中的 F 值为经自由度调整之后的组间方差与组内方差之比，F 值越大，分类效果越好。

## 3 结论与启示

统计方法和统计模型的层出不穷为学术研究提供了广阔的方法论选择空间，但是如果对经典统计方法的理论基础、适用性前提以及存在的问题缺乏深入理解，盲目地追求对经典统计方法的拓展则可能陷入统计方法的研究误区．本文针对经典聚类分析实际应用中的诸多局限性展开讨论，系统分析了现有改进聚类模型的优点与不足；借鉴现有主成分聚类分析的思想，提出了基于方差贡献率的自适应赋权主成分聚类模型．理论分析表明新模型有机集成了多个理论和方法的长处，每一步都有充分的理论基础保证其必要性、合理性．实践检验表明新方法

能够有效解决现有聚类模型极端情形下的失效问题，有着复杂分类问题下的普适性。

应用新提出的加权主成分聚类对 2011 年全国 31 个省 (区) 市创新能力静态分类发现，区域创新能力呈现长江三角洲、珠江三角洲和京津地区多元化的竞争格局。整体来看，我国目前的区域创新能力仍未摆脱从东部沿海向西部内陆由高到低阶梯分布的格局，表明创新能力与经济发展存在显著的正向反馈机制。针对上述情况，要增强和提升中国的创新能力，必须分集团而不是整齐划一地制定和实施增强区域自主创新能力的对策建议。一方面，要采取适宜对策强化创新能力强的领先集团，把提高原始性创新能力作为该类地区科技创新的重点，使其成为研发核心技术的集中源发地；另一方面，要加大对落后区域的创新投资，构建和增强区域集成创新能力和消化吸收再创新能力。由于 $F_1$(创新因子) 的方差贡献最大，体现了创新能力水平的主要方面，在经济能力受限的条件下，落后区域应该集中精力优先发展企业创新、创新环境和创新绩效三个方面，带动其它两个方面的发展，这也是抓住了工作的重心。

## [ 参考文献 ]


[1] Michel Mouchart, Jeroen V K Romouts. Clustered panel data model: An efficient approach for now casting from poor data [J]. International Journal of Forecasting, 2005, 21(5): 577–594.

[2] Shia B C, Zhu J P, Fang K N, Ma S G. Fuzzy canonical discriminant analysis: theory and practice [J]. Communications in Statistics-Simulation and Computation, 2011, 40(4): 1526–1539.

[3] 郑冰云. 多指标面板数据的聚类分析及其应用 [J]. 数理统计与管理, 2008, 27(2): 265–270.

[4] 任娟. 多指标面板数据融合聚类分析 [J]. 数理统计与管理, 2013, 32(1): 57–67.

[5] 肖泽磊, 李帮义, 刘思峰. 基于多维面板数据的聚类方法探析及实证研究 [J]. 数理统计与管理, 2009, 28(5): 831–838.

[6] 毛国敏, 顾建华, 吴新燕. 地震灾害的分类和分级方法研究 [J]. 地震学报, 2007, 19(4): 426–436.

[7] 孙锐, 石金涛. 基于因子和聚类分析的区域创新能力再评价 [J]. 科学学研究, 2006, 24(6): 985–990.

[8] 袁建新, 刘幸赟. 技术引进促进经济增长作用省际差异性影响因素分析 [J]. 中国工业经济, 2010, (5): 78–87.

[9] 陆根尧, 盛龙, 唐晨华. 中国产业生态化水平的静态与动态分析 [J]. 中国工业经济, 2012, (3): 147–159.

[10] 庞丽, 李显君. 我国汽车产业竞争力区域差异的实证研究 [J]. 数理统计与管理, 2011, 30(6): 951–959.

[11] 李因果, 何晓群. 面板数据聚类方法及应用 [J]. 统计研究, 2010, 27(9): 73–79.

[12] 王德青, 朱建平, 谢邦昌. 主成分聚类分析有效性的思考 [J]. 统计研究, 2012, 29(11): 84–87.

[13] 王德青. 主成分聚类分析在矿井安全评价应用中的思考 [J]. 中国矿业, 2011, 20(1): 51–57.

[14] 朱建平. 应用多元统计分析 [M]. 北京: 科学出版社, 2013.

[15] Fisher R A. The use of multiple measurement in taxonomic problems [J]. Annals of Eugenics, 1936, (7): 179–188.

[16] 中国科技发展战略研究小组. 中国区域创新能力报告 2011 [M]. 北京: 科学出版社, 2012.

[17] 周立, 吴玉鸣. 中国区域创新能力: 因素分析与聚类研究 [J]. 中国软科学, 2006, (8): 96–103.

[18] 袁永生等. 黄河下游复杂水位有效拟合新方法研究 [J]. 中国科学 (技术科学), 2009, 39(11): 1875–1880.

[19] 王德青等. 基于主成分的改进雷达图及其在综合评价中的应用 [J]. 数理统计与管理, 2010, 29(5): 881–889.

[20] 徐雅静, 汪远征. 主成分分析应用方法的改进 [J]. 数学的实践与认识, 2006, 36(6): 68–75.

[21] 朱建平. 数据挖掘的统计方法及实践 [M]. 北京: 中国统计出版社, 2005.

[22] 王德青. 统计分类方法的比较 [J]. 中国统计, 2008, (9): 45–46.

[23] 朱建平, 陈民恳. 面板数据聚类分析及其应用 [J]. 统计研究, 2007, 24(4): 11–14.

[24] 朱建平, 方匡南. 有序秩聚类及对地震活跃期的分析 [J]. 统计研究, 2009, 26(1): 83–87.