

所属类别	2020 年“华数杯”全国大学生数学建模竞赛	参赛编号

基于大数据对脱贫帮扶绩效的评价

摘要

今年是全面建成小康社会目标实现之年，也是全面打赢脱贫攻坚战收官之年，任务艰巨又遭遇疫情影响，给打赢脱贫攻坚战增添了困难。为了更好的激励各帮扶单位提高扶贫效率，我国的研究团队分别在五年前与今年对需要帮扶的贫困村进行了贫困调研，本文基于脱贫帮扶绩效评价相关数据，结合实际情况，对于不同问题建立了不同的数学模型进行求解。

针对问题一：以**典型相关分析法**为主要解题思路，将五年前的各项评价指标与今年的各项评价指标作为两组典型变量，建立**典型相关模型**，利用附件中的指标数据得出五个评价指标即居民收入、产业发展、居住环境、文化教育、基础设施的**典型相关系数**分别为0.5342、0.6399、0.7372、0.6361、0.5945。根据典型相关系数判断出居住环境这一指标的相关性较强，其他四个指标的相关性较弱，但总的来说，没有相关性非常强的指标。为了检验该方法的正确性，进行显著性检验，所有的检验结果都是通过，得出结论并该问各指标没有很强烈的相关规律。

针对问题二：以**主成分分析法**为主要解题思路，我们将五个评分数据与总分数据综合考虑，计算六个数据在2020年相较于2015年的增长率，利用主成分分析法，将6个数据的增长率与2020年的6个原始数据作为相关系数矩阵，得到该矩阵的特征值以及相应的特征向量，并根据各个主成分的**贡献率**选取主成分的个数，建立了基于评分数据以及总分数据增长率的**得分模型**，计算出所有贫困村对应的综合得分，根据得分的高低得出不同类型的帮扶单位绩效的高低，给出了脱贫帮扶绩效前十名的帮扶单位编号分别为58、100、20、97、144、145、79、55、58、58。再综合得分排名前列的帮扶单位统计每种单位类型所占比例，得出帮扶单位类型绩效排名为为：类型3>类型1>类型5>类型2。

针对问题三：我们单独选择每个标号所包含单位对应的村庄，将这些村庄的每个指标的增长率求取平均值，然后依次选择0-159各个标号所包含单位对应的村庄，对这些村庄的每个指标的**增长率**求取平均值，建立基于增长率年平均值的**排序模型**，将各个标号所对应该指标的增长率平均值进行排序，得出了各单项评价指标帮扶业绩明显的帮扶单位。

针对问题四：我们选择 2015 年的六个指标数据以及帮扶单位的编号为输入，2020 年的总分数据为输出，利用**神经网络算法**，建立**预测模型**，预测出该 10 个村庄 2020 年的总分数据，将其与总分数据排名前 10000 名的村庄相比较，判断出编号为 25149 的村庄能评上“脱贫先进村庄”称号。根据一级称号与二级称号的比例判断出编号为 25149 的村庄为“脱贫先进一级村庄”。

针对问题五：我们向国家扶贫办写了一封建议信，阐述了我们对于脱贫攻坚的认识并给出了相关的观点和建议。

关键词：相关系数 主成分分析法 得分模型 排序模型 神经网络算法 预测模型

一、 问题重述

1.1 问题背景

消除贫困、改善民生、逐步实现共同富裕，是社会主义的本质要求，是我们党的重要使命。党的十八大以来，国家把扶贫开发工作纳入“四个全面”战略布局，作为实现第一个百年奋斗目标的重点工作，摆在更加突出的位置。

党的十八大以来，以习近平同志为核心的党中央采取一系列超常规举措，举全党全社会之力打赢脱贫攻坚战，使得我国扶贫工作不断取得新的进展，其中所积累的一系列先进经验值得深刻总结，包括“以人民为中心”的理念贯穿始终、坚持制度变革推动减贫进程、坚持以经济发展推动贫困治理等等。上述基本经验对于探寻我国扶贫工作的基本规律、开展 2020 年后治理相对贫困工作具有重要意义^[1]。

2020 年是精准扶贫的收官之年，如何防止贫困反弹并实现扶贫效果长久化、制度化，成为政府及社会各界必须面对和考虑的战略性问题。研究指出，“后 2020”时期，绝对贫困即将成为历史，相对贫困将会显现，为应对相对贫困，需要调整扶贫标准，克服扶贫过程中暴露出来的一些痼疾顽症^[2]。因此，为了了解我国的脱贫情况，我们有必要利用相关数学算法，结合大数据背景，评价当前脱贫帮扶的绩效如何，从而提出针对性的应对措施。

1.1 问题提出

为了更好地评价脱贫帮扶绩效，我们建立数学模型解决下列问题：

(1) 一般人的理解是，五年前的评分与 2020 年对应的各项评分有着直接的关联，如五年前的居民收入不错，现在的居民收入也会是不错的。你认为本问题有这个规律吗？请分析各个评价指标的对应关系。

(2) 160 个帮扶单位帮扶着基础不同的村庄，帮扶单位帮扶工作的态度、目标、投入、帮扶干部素质等显然是有差异的。仅仅用 2020 年各村庄评分高低显然是无法真正有效的体现一个帮扶单位在脱贫攻坚提升方面所做出的努力。请你运用附件的数据，阐明什么类型的帮扶单位，哪些帮扶单位在脱贫帮扶上面有较高的绩效？请给不同类型的帮扶单位绩效排序，给出脱贫帮扶绩效前十名的帮扶单位编号。

(3) 每个帮扶单位在扶贫上有不同的工作特色，如有些单位在提高居民收入上效果很好，而有些帮扶单位可能在改善基础设施上帮助的效果不错。请问，哪些帮扶单位分别在居民收入、产业发展、居住环境、文化教育、基础设施等评价指标上帮扶业绩明显？请列出各单项评价指标前五名的帮扶单位编号。

(4) 全国计划给予 10000 个村庄“脱贫先进村庄”称号。请问，哪些因素对获得这个荣誉称号有着非常重要的影响？数据表中最后有 10 个村庄的 2020 年的评价分数被删除，请你判断他们能评上“脱贫先进村庄”称号吗？如果称号分为一级和二级（一、二级称号比例为 1:3），这 10 个村庄中谁能评上“脱贫先进一级村庄”称号？

(5) 依据你的研究成果，向国家扶贫办写一封 500 字左右的信，阐述你的观点和建议。

二、问题分析

问题一：该问题要求我们判断各个评价指标五年前的评分与 2020 年对应的评分是否有着直接的关联，并分析各个评价指标的对应关系。这是一个对指标进行相关性判断的问题，我们可以以典型相关分析法为主要思路，将五年前的各项评价指标与今年的各

项评价指标作为两组典型变量，利用附件中的指标数据得出五个评价指标典型相关系数，取某一数值为典型相关系数的临界值，若某一指标的相关系数大于该值，则该指标五年前的评分与 2020 年对应的各项评分有着直接的关联，否则认为该指标五年前的评分与 2020 年对应的各项评分没有直接的关联。为了检验该方法的正确性，可进行显著性检验，若所有的检验结果都是通过，则证明使用典型相关分析法的正确性。为分析各个评价指标的对应关系，可建立典型相关模型，基于典型相关系数的区间范围，将各个评价指标的对应关系进行分类，该相关系数越大，关系越密切，从而判断出各个评价指标的对应关系。

问题二：该问题要求我们运用附件的数据，阐明什么类型的帮扶单位，哪些帮扶单位在脱贫帮扶上面有较高的绩效，并给不同类型的帮扶单位绩效排序，给出脱贫帮扶绩效前十名的帮扶单位编号。我们选择将五个评分数据与总分数据综合考虑，计算六个数据在 2020 年相较于 2015 年的增长率，利用主成分分析法，将 6 个数据的增长率与 2020 年的 6 个原始数据作为相关系数矩阵，并根据各个主成分的贡献率选取主成分的个数，计算出所有贫困村对应的综合得分，根据得分的高低便可得出不同类型的帮扶单位绩效的高低，对于得分高的帮扶单位即为在脱贫帮扶上面有较高绩效的帮扶单位。再综合得分排名前列的帮扶单位统计每种单位类型所占比例，数量比例高的即为在脱贫帮扶上面有较高绩效的帮扶单位类型。

问题三：问题三要求我们列出各单项评价指标前五名的帮扶单位编号，由于被帮扶的村庄划分为160个集合，每个集合指定帮扶单位（标记为0-159）进行帮扶，即每个标号对应一个帮扶单位，该帮扶单位分别帮扶其对应的贫困村庄。我们可以单独选择每个标号所包含单位对应的村庄，将这些村庄的每个指标的增长率求取平均值，然后依次选择0-159各个标号所包含单位对应的村庄，对这些村庄的每个指标的增长率求取平均值。这样一来，就某一单项评价指标而言，将各个标号所对应该指标的增长率平均值进行排序，从而得出针对某一单项评价指标帮扶业绩明显的帮扶单位。

问题四：问题四首先要求我们给出影响“脱贫先进村庄”这一称号的因素，为了判断最后10个村庄能否评上该称号，我们可以选择2015年的六个指标数据以及帮扶单位的编号为输入，2020年的总分数据为输出，利用神经网络算法，建立预测模型，通过预测出的该10个村庄2020年的总分数据，与总分数据排名前10000名的村庄相比较，从而判断他们是否能评上该称号。由于一级称号与二级称号的比例为1:3，选择总分数据前2500名的村庄获得一级称号，总分数据后7500名的村庄获得二级称号。那么这10个村庄中总分数据位于前2500名则能评上“脱贫先进一级村庄”的称号。

问题五：问题五要求我们根据给国家扶贫办写一封信，阐述我们团队对于当前我国脱贫攻坚的认识，并给出一些观点和建议。结合前四个问题的研究结果，并结合我国的实际情况，向国家扶贫办给出一些对于扶贫开发工作的观点和建议。

三、基本假设

- 1、假设对于脱贫帮扶绩效评价的标杆，模型所确立的指标，能够准确地代表脱贫帮扶绩效的高低；
- 2、假设题目附件所给的科研团队调研数据真实有效；
- 3、假设科研团队在进行贫困调研时，对各个村庄的评价公平公正；
- 4、假设不考虑贫困村的人口流动等因素的影响，对于脱贫帮扶绩效评价，除去本文提到的指标因素，对其他因素所造成的影响忽略不计；
- 5、假设在不考虑其他因素的情况下，某村庄的总分数据能够反映该村庄目前的贫困情

况;

6、在打分过程中，各因子的贡献率能够代表其在实际生活所占权重。

四、符号说明

符号	说明
U_k, V_k	某一指标的第 k 对典型相关变量
ρ_k	第 k 对典型相关变量的典型相关系数
e_k, f_k	正交单位特征向量
a	居民收入的增长率
b	产业发展的增长率
c	居住环境的增长率
d	文化教育的增长率
e	基础设施的增长率
f	总分的增长率
b_j	信息贡献率
α_p	累积贡献率
Z	综合得分

五、数据的预处理

5.1 数据的筛选和删除

通过查询网站数据并结合附件所给数据分析，可以将一些无关项删除，保留有用项。如通信网络、生产条件、社会保障等指标，由于这些指标数据对研究项目影响很小或者没有，所以本文仅选择附件中的数据将其保留。

5.2 数据的标准化

理论上，标准化适用于服从正态分布的数据，目前很多工程都依赖大数据，所以在样本足够多的情况下，往往直接使用标准化对数据进行无量纲化预处理，在深度学习中，将数据标准化能够保证有更好的收敛。如果不进行数据标准化，有些特征将会对损失函数影响很大，使得其他值比较小的特征重要性降低。

为了便于比较和研究，本文已将所有数据都进行了标准化处理，故不再作处理。

六、模型的建立与求解

6.1 问题一模型的建立与求解

6.1.1 问题一模型的建立

问题一要求我们分析各个评价指标的对应关系，我们将五年前的各项评价指标与今年的各项评价指标作为两组典型变量，利用附件中的指标数据得出五个评价指标典型相关系数，建立典型相关模型，取0.5为典型相关系数的临界值，若某一指标的相关系数大于该值，则该指标五年前的评分与2020年对应的各项评分有着直接的关联，否则认为该指标五年前的评分与2020年对应的各项评分没有直接的关联，从而得出各个评价指标的对应关系。

• 模型的准备

通过查阅文献可知^[3]，对于两组随机变量 $(x_1, x_2, \dots, x_p)$ 和 $(y_1, y_2, \dots, y_q)$ ，考虑 $(x_1, x_2, \dots, x_p)$ 的一个线性组合 U 及 $(y_1, y_2, \dots, y_q)$ 的一个线性组合 V ，希望找到 U 和 V 之间的最大可能的相关系数，来充分反映两组变量之间的关系。这样就可以把研究两组随机变量间相关关系的问题转化为研究两个随机变量之间的相关关系。如果说一对变量 (U, V) 还不能完全刻画两组变量间的相关关系，可以继续找第二对变量，但是要保证第二对的变量与第一对的变量不相关，直至找不到相关变量时为止。

(1) 确定总体典型相关变量

设有两组随机向量 $x = (x_1, x_2, \dots, x_p)^T$ ， $y = (y_1, y_2, \dots, y_q)^T$ ($p \leq q$)，将两组合并成一组向量 $(x^T, y^T) = (x_1, x_2, \dots, x_p, y_1, y_2, \dots, y_q)^T$ ，其协方差矩阵为：

$$\Sigma = \begin{bmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{bmatrix} \quad (6-1-1)$$

其中 $\sum_{11} = \text{cov}(x)$ ， $\sum_{22} = \text{cov}(y)$ ， $\sum_{12} = \sum_{21}^T = \text{cov}(x, y)$ 。

根据典型相关思想，问题是要寻找 $x = (x_1, x_2, \dots, x_p)^T$ ， $y = (y_1, y_2, \dots, y_q)^T$ ($p \leq q$)的线性组合：

$$U_1 = a_1^T x = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p$$

$$V_1 = b_1^T y = b_{11}y_1 + b_{12}y_2 + \dots + b_{1q}y_q$$

要使得 U_1, V_1 的相关系数 $\rho(U_1, V_1)$ 达到最大，这里

$$a_1^T = (a_{11}, a_{12}, \dots, a_{1p})$$

$$b_1^T = (b_{11}, b_{12}, \dots, b_{1q})$$

由(6-1-1)式， $\text{var}(U_1) = a_1^T \sum_{11} a_1$ ， $\text{var}(V_1) = b_1^T \sum_{22} b_1$ ， $\text{cov}(U_1, V_1) = a_1^T \sum_{12} b_1$ ，所以 U_1, V_1 的相关系数为：

$$\rho(U_1, V_1) = \frac{a_1^T \sum_{12} b_1}{\sqrt{a_1^T \sum_{11} a_1} \sqrt{b_1^T \sum_{22} b_1}} \quad (6-1-2)$$

又由于相关系数与量纲无关，因此可设定约束条件：

$$a_1^T \sum_{11} a_1 = b_1^T \sum_{22} b_1 = 1 \quad (6-1-3)$$

满足约束条件 (6-1-3) 的相关系数 $\rho(U_1, V_1)$ 的最大值称为第一典型相关系数, U_1, V_1 称为典型相关变量。

(2) 计算典型相关系数

将矩阵 $[X^T, Y^T]^T$ 的协方差矩阵或相关系数矩阵表示如下:

$$\Sigma = \begin{bmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{bmatrix}, \quad R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

令 $A = (\sum_{11})^{-1/2} \sum_{12} (\sum_{22})^{-1} \sum_{21} (\sum_{11})^{-1/2}$, $B = (\sum_{22})^{-1/2} \sum_{21} (\sum_{22})^{-1} \sum_{12} (\sum_{22})^{-1/2}$, 求 A, B 的特征值 $\rho_1^2, \rho_2^2, \dots, \rho_p^2$ 以及对应的正交单位特征向量 $e_k, f_k (k=1, 2, \dots, p)$ 。

则 x, y 的第 k 对典型相关变量为:

$$\begin{cases} U_k = a_k^T x = e_k^T \sum_{11}^{-0.5} x \\ V_k = b_k^T y = f_k^T \sum_{22}^{-0.5} y \end{cases}, k=1, 2, \dots, p$$

其中 $\sum_{11}^{-0.5}$, $\sum_{22}^{-0.5}$ 分别为 $\sum_{11}$, $\sum_{22}$ 的平方根矩阵的逆矩阵。

x, y 的第 k 对典型相关变量的相关系数为:

$$\rho_k = a_k^T \sum_{12} b_k (k=1, 2, \dots, p)$$

• 模型的建立

由典型相关分析可知, 将2015年的居民收入 (SR)、产业发展 (CY)、居住环境 (HJ)、文化教育 (WJ)、基础设施 (SS) 五个评价指标分别与2020年的居民收入、产业发展、居住环境、文化教育、基础设施作为五组典型相关变量, 建立典型相关模型, 画出模型结构如下:

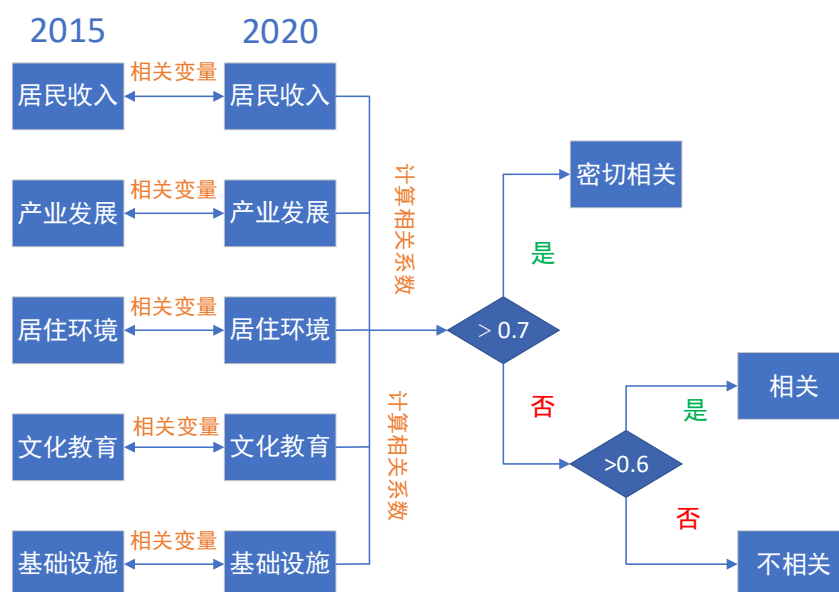


图1: 典型相关模型结构图

由上图可知各个评价指标最终的对应关系是由典型相关系数来判断的, 相关系数越

大，说明该指标在五年前与今年的关联度越高，该系数大小的确定需要根据题目中所给出的数据进行求解，求解过程及结果在下面模型求解中给出。至此，我们建立出典型相关模型如下：

$$\begin{cases} U_k = a_k^T x = e_k^T \sum_{11}^{-0.5} x \\ V_k = b_k^T y = f_k^T \sum_{22}^{-0.5} y \end{cases} (k=1, 2, \dots, 32165) \quad (6-1-4)$$

$$\rho_k = a_k^T \sum_{12} b_k (k=1, 2, \dots, 32165) \quad (6-1-5)$$

其中 U_k ， V_k 为某一指标的第 k 对典型相关变量， ρ_k 第 k 对典型相关变量的典型相关系数。

相关系数的意义如下表所示：

表1：相关系数的意义

相关系数绝对值	大体划分	细分
0.9-1.0	相关	相关性非常强
0.7-0.9	相关	相关性比较强
0.5-0.7	相关	相关性有点弱
<0.5	不相关	相关性非常弱

6.1.2 问题一模型的求解

该问题的整体求解思路为计算出各个评价指标的典型相关系数，再对结果进行显著性检验，检验该典型相关分析法是否正确。那么模型的求解可大致分为以下 3 步：

Step1：求解典型相关系数

根据题目附件所给的五个评价指标的相关数据，利用协方差矩阵和相关系数矩阵求得各指标的典型相关系数如下表：

表 2：各指标的典型相关系数

居民收入 (SR)	产业发展 (CY)	居住环境 (HJ)	文化教育 (WJ)	基础设施 (SS)	总分
0.5342	0.6399	0.7372	0.6361	0.5945	0.7845

Step2：显著性检验

在本模型中我们使用的检验方法是 P 值检验与统计量检验，P 值即为(零假设的)拒绝域的面积或概率。用 P 值做判断与检验统计量做判断得出的结论是完全一样的，在检验中有两种方法，第一种，检验统计量是否落在拒绝域中，第二种，P 值是否小于显著性水平。

其实两种判断是等价的，拒绝域是根据显著性水平计算出来的，然后看检验统计量与拒绝域比较；而 P 值是根据检验统计量算出来的，然后看 P 值与显著性水平比较。在

分布图上，P 值所在位置就是检验统计量所在位置。

表 3：显著性检验表

wilks	df1	df2	F	pF	Chisq	pChisq	dfe	p
0.7146	1	32153	1.2842e+04	0	1.0805e+04	0	1	0
0.5905	1	32153	2.2301e+04	0	1.6939e+04	0	1	0
0.4566	1	32153	3.8267e+04	0	2.5206e+04	0	1	0
0.5954	1	32153	2.1853e+04	0	1.6674e+04	0	1	0
0.6466	1	32153	1.7574e+04	0	1.4020e+04	0	1	0
0.3846	1	32153	5.1442e+04	0	3.0721e+04	0	1	0

其中，wilks 为似然比统计量，df1 为 F 统计量的分子自由度（卡方统计的自由度），df2 为 F 统计量的分母自由度，F 为近似 F 统计量，pF 为 F 的右尾显著性水平，chisq 为近似卡方统计，pchisq 为 chisq 的右尾显著性水平。

Step3：得出各个评价指标的对应关系

由表 2 可知，第三个指标（居住环境）的典型相关系数位于区间 $[0.7, 0.9]$ ，表明五年前的评分与 2020 年对应的各项评分相关，且相关性较强；第一个指标（居民收入）、第二个指标（产业发展）、第四个指标（文化教育）与第五个指标（基础设施）的典型相关系数位于区间 $[0.5, 0.7]$ ，表明五年前的评分与 2020 年对应的各项评分相关，且相关性较弱。

由表 3 可知，五对典型变量均通过了显著性检验，表明相应典型变量之间相关关系显著。

因此，我们得出结论：对于本问中所提到的五年前的评分与 2020 年对应的各项评分有着直接的关联这一规律，居住环境具有这种规律，关联度较高；居民收入、产业发展、文化与基础设施也具有这种规律，但关联度较低。总的来说，没有相关性非常强的指标。

6.2 问题二模型的建立与求解

6.2.1 问题二模型的建立

问题二要求我们运用附件的数据，阐明什么类型的帮扶单位，哪些帮扶单位在脱贫帮扶上面有较高的绩效，并给不同类型的帮扶单位绩效排序，给出脱贫帮扶绩效前十名的帮扶单位编号。我们将五个评分数据与总分数据综合考虑，计算六个数据在 2020 年相较于 2015 年的增长率，利用主成分分析法，将 6 个数据的增长率与 2020 年的 6 个原始数据作为相关系数矩阵，计算出该矩阵的特征值，并根据各个主成分的贡献率选取主成分的个数，建立基于评分数据及总分数据增长率的得分模型。根据得分的高低便可得出不同类型的帮扶单位绩效的高低，对于得分高的帮扶单位即为在脱贫帮扶上面有较高绩效的帮扶单位。

• 模型的准备

主成分分析法是最常用的线性降维方法，它的目标是通过某种线性投影，将高维的数据映射到低维的空间中，并期望在所投影的维度上数据的信息量最大（方差最大），以此使用较少的数据维度，同时保留住较多的原数据点的特性。

假设有 M 个样本 $\{x_1, x_2, x_3, \dots, x_m\}$ ，每个样本有 N 维特征 $x^i = (x_1^i, x_2^i, x_3^i, \dots, x_N^i)^T$ ，每一个特征 x_j 都有各自的特征值。

(1) 对特征进行中心化

求每一个特征的平均值，然后对于所有的样本，每一个特征都减去自身的均值。特征 x_j 的平均值为：

$$\bar{x}_j = \frac{1}{M} \sum_{i=1}^M x_j^i \quad (6-2-1)$$

经过去均值处理之后，原始特征的值就变成了新的值。在此基础上，进行后面的操作。

(2) 求协方差矩阵 C

$$C = \begin{bmatrix} \text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) \end{bmatrix}$$

上述矩阵中，对角线上分别是特征 x_1 和 x_2 的方差，非对角线上是协方差。协方差大于 0 表示 x_1 和 x_2 若有一个增，另一个也增；小于 0 表示一个增，一个减；协方差为 0 时，两者独立。协方差绝对值越大，两者对彼此的影响越大，反之越小。其中， $\text{cov}(x_1, x_2)$ 的求解公式如下：

$$\text{cov}(x_1, x_2) = \frac{\sum_{i=1}^M (x_1^i - \bar{x}_1)(x_2^i - \bar{x}_2)}{M - 1} \quad (6-2-2)$$

其他可类似进行求解，根据上面的协方差计算公式我们就得到了这 M 个样本在这 N 维特征下的协方差矩阵 C 。

(3) 求协方差矩阵 C 的特征值和特征向量

利用矩阵的知识，求协方差矩阵 C 的特征值 λ 和相对应的特征向量 u ：

$$Cu = \lambda u$$

特征值 λ 会有 N 个，每一个 λ_i 对应一个特征向量 u_i ，将特征值 λ 按照从大到小的顺序排序，选择最大的前 k 个，并将其相对应的 k 个特征向量拿出来，我们会得到一组 $\{(\lambda_1, u_1), (\lambda_2, u_2), \dots, (\lambda_k, u_k)\}$ 。

(4) 得到降维后的新 k 维特征

这个选取最大的前 k 个特征值和相对应的特征向量，并进行投影的过程，就是降维的过程。对于每一个样本，原来的特征是 $(x_1^i, x_2^i, x_3^i, \dots, x_N^i)^T$ ，投影之后的新特征是 $(y_1^i, y_2^i, y_3^i, \dots, y_k^i)^T$ ，新特征的计算公式如下：

$$\begin{bmatrix} y_1^i \\ y_2^i \\ \cdot \\ \cdot \\ \cdot \\ y_k^i \end{bmatrix} = \begin{bmatrix} u_1^T \cdot (x_1^i, x_2^i, \dots, x_n^i)^T \\ u_2^T \cdot (x_1^i, x_2^i, \dots, x_n^i)^T \\ \dots \\ \dots \\ \dots \\ u_k^T \cdot (x_1^i, x_2^i, \dots, x_n^i)^T \end{bmatrix}$$

•模型的建立

我们将五个评分数据与总分数据综合考虑，计算六个数据在 2020 年相较于 2015 年的增长率，利用主成分分析^[4]法，将 6 个数据的增长率与 2020 年的 6 个原始数据作为相关系数矩阵，计算出该矩阵的特征值，并根据各个主成分的贡献率选取主成分的个数，建立基于评分数据及总分数据增长率的得分模型。根据各个帮扶单位得分的高低便可得出不同类型的帮扶单位绩效的高低，对于得分高的帮扶单位即为在脱贫帮扶上面有较高绩效的帮扶单位。

下面给出该得分模型的流程图：

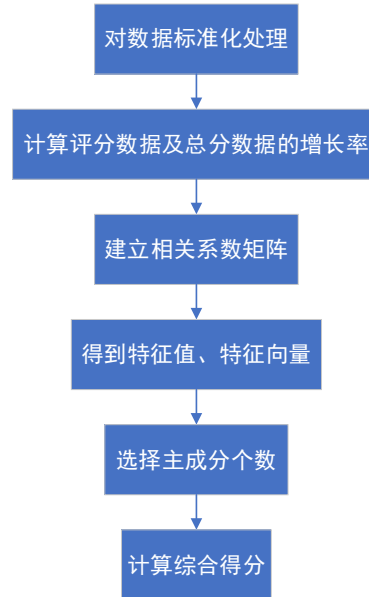


图 2：得分模型流程图

我们对所有贫困村基于其评分数据与总分数据在 2020 年相较于 2015 年的增长率计算出各个村庄的综合得分，得分高的贫困村所对应的帮扶单位在脱贫帮扶上面有较高的绩效。将所有贫困村的得分进行排序，则得到了帮扶单位的绩效排序。我们选取综合得分位于前一百名的贫困村所对应的帮扶单位，认为他们在脱贫帮扶上面有较高的绩效；选取综合得分位于某一范围内的贫困村所对应的帮扶单位，计算不同类型帮扶单位在该范围内的数量比例，认为比例较高的帮扶单位类型在脱贫帮扶上面有较高的绩效。

至此，我们建立基于评分数据及总分数据增长率的得分模型如下：

（1）求取评分数据及总分数据的增长率

令居民收入、产业发展、居住环境、文化教育、基础设施评分以及总分的增长率分别为 a、b、c、d、e、f，表示如下：

$$\left\{ \begin{array}{l} a = \frac{SR(2020) - SR(2015)}{SR(2015)} \\ b = \frac{CY(2020) - CY(2015)}{CY(2015)} \\ c = \frac{HJ(2020) - HJ(2015)}{HJ(2015)} \\ d = \frac{WJ(2020) - WJ(2015)}{WJ(2015)} \\ e = \frac{SS(2020) - SS(2015)}{SS(2015)} \\ f = \frac{\text{总分}(2020) - \text{总分}(2015)}{\text{总分}(2015)} \end{array} \right. \quad (6-2-3)$$

(2) 建立相关系数矩阵 R

根据我们选取的数据分析可知，相关系数矩阵 $R=(r_{ij})_{32165 \times 12}$ 。

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdot & \cdot & r_{(1)(12)} \\ r_{21} & r_{22} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{(32165)(1)} & \cdot & \cdot & \cdot & r_{(32165)(12)} \end{bmatrix}$$

其中，当 $j \leq 6$ 时， r_{ij} 为第 i 个贫困村第 j 个数据的增长率；当 $j > 6$ 时， r_{ij} 为第 i 个贫困村 2020 年第 $(j-6)$ 个数据的原始值。

(3) 计算特征值和特征向量

计算相关系数矩阵 R 的特征值 $\lambda_j = (\lambda_1, \lambda_2, \dots, \lambda_{12})$ 以及对应的标准化特征向量 $u_j = (u_1, u_2, \dots, u_{12})^T$ ，由特征向量组成 12 个新的指标变量：

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_{12} \end{bmatrix} = \begin{bmatrix} u_1 \cdot (x_1, x_2, \dots, x_{12}) \\ u_2 \cdot (x_1, x_2, \dots, x_{12}) \\ \dots \\ \dots \\ \dots \\ u_{12} \cdot (x_1, x_2, \dots, x_{12}) \end{bmatrix}$$

(4) 选择主成分个数

通过计算特征值 $\lambda_j (j=1, 2, \dots, 12)$ 的信息贡献率 b_j 与累积贡献率 α_p ，两者的计算方法如下：

$$\begin{cases} b_j = \frac{\lambda_j}{\sum_{k=1}^{12} \lambda_k}, j=1,2,\dots,12 \\ \alpha_p = \frac{\sum_{k=1}^p \lambda_k}{\sum_{k=1}^{12} \lambda_k} \end{cases} \quad (6-2-4)$$

当 α_p 接近于1时,则选择前 p 个指标变量 $y_1, y_2, \dots, y_p$ 作为 p 个主成分,代替原来的指标变量,从而可对 p 个主成分进行综合分析。

(5) 计算综合得分 Z

最后得出的综合得分值作为我们评判帮扶单位绩效高低的标准,得分越高,代表该贫困村脱贫情况越好,即该贫困村所对应的帮扶单位绩效越高。

$$Z = \sum_{j=1}^p b_j y_j \quad (6-2-5)$$

其中, b_j 为第 j 个主成分的信息贡献率,根据综合得分值就可进行相应排序与评价。

6.2.2 问题二模型的求解

我们利用计算机软件对32165个贫困村的评分数据以及总分数据算出各自指标的增长率,通过构建相关系数矩阵得到特征值如下表所示:

表 4: 矩阵特征值

	1	2	3	4	5	6
特征值	4.7110	1.0283	1.0216	0.9960	0.9835	0.9697
	7	8	9	10	11	12
特征值	0.9666	0.5128	0.3168	0.2545	0.2308	0.0079

由于通过计算机软件的特征向量数据过于冗杂,故这里展示部分结果,完整结果详见附件。

表 5: 部分特征向量值

1	2	3	4
0.0372	0.6433	0.2117	0.4825
0.0670	0.2869	0.1815	0.0480
0.0410	0.0459	0.6817	0.2636
0.0476	0.2664	0.2631	0.8044

由式(6-2-4)分别计算出12个特征值的贡献率如下表所示:

表 6: 特征值的贡献率

	1	2	3	4	5	6
贡献率	39.2584	8.5692	8.5139	8.3005	8.1964	8.0813
	7	8	9	10	11	12

贡献率	8.0550	4.2735	2.6404	2.1214	1.9238	0.0662
-----	--------	--------	--------	--------	--------	--------

最终，我们计算出所有贫困村基于评分数据以及总分数据的综合得分，这里展示得分前五名的贫困村排名，完整排名详见附件“sj2.xlsx”。

表 7：部分贫困村综合得分排名

村庄编号	帮扶单位 (0-159)	帮扶单位类型 (0-5)	综合得分
24032	58	1	75.40644823
28745	100	5	66.5621569
15344	20	2	60.97766683
28103	97	3	56.14412667
50096	144	3	53.93653142
50399	145	3	47.74866269
34386	79	3	45.76284147
23490	55	2	45.19598216
24041	58	1	44.80786136
23945	58	1	43.75222852
29998	104	3	39.99029972
51595	148	3	38.48604738
28100	97	3	37.41162991
25823	91	3	34.63453939
25965	91	3	33.70653591

由上表可知，脱贫帮扶绩效前十名的帮扶单位编号分别为58、100、20、97、144、145、79、55、58、58。

对于阐明哪些帮扶单位在脱贫帮扶上面有较高的绩效这一问题，我们根据附件“sj2.xlsx”选取综合得分位于前一百名的贫困村所对应的帮扶单位，认为他们在脱贫帮扶上面有较高的绩效。

对于阐明什么类型的帮扶单位在脱贫帮扶上面有较高的绩效这一问题，我们选取综合得分位于某一范围内的贫困村所对应的帮扶单位，计算不同类型帮扶单位在该范围内的数量比例，认为比例较高的帮扶单位类型在脱贫帮扶上面有较高的绩效。

我们选取了两种情况进行评价分析：

① 对得分排名前15的贫困村分析

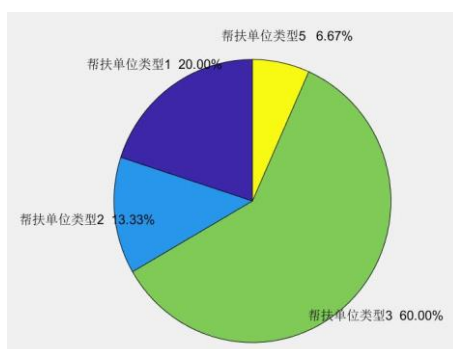


图3：前15名不同帮扶单位类型数量比例

由图3可知，帮扶单位类型3所占比例最大，按数量比例从大到小的顺序排列，得分位于前15名的贫困村所对应的帮扶单位类型数量为：类型3>类型1>类型5>类型2。

② 对得分排名前50的贫困村分析

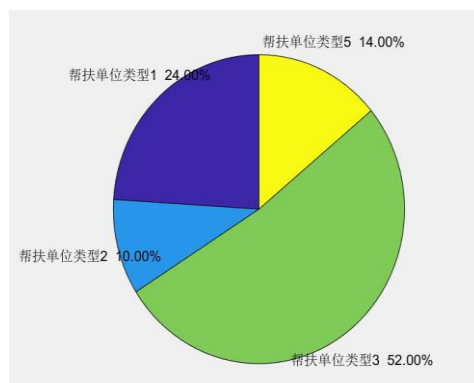


图4：前50名不同帮扶单位类型数量比例

由图4可知，依然是帮扶单位类型3所占比例最大，按数量比例从大到小的顺序排列，得分位于前50名的贫困村所对应的帮扶单位类型数量为：类型3>类型1>类型5>类型2。

综合以上结论，我们认为帮扶单位类型3在脱贫帮扶上面有较高的绩效。

6.3 问题三模型的建立与求解

6.3.1 问题三模型的建立

问题三要求我们列出各单项评价指标前五名的帮扶单位编号，由于被帮扶的村庄划分为160个集合，每个集合指定帮扶单位（标记为0-159）进行帮扶，即每个标号对应一个帮扶单位，该帮扶单位分别帮扶其对应的贫困村庄。通过附件中的数据，单独选择每个标号所包含单位对应的村庄，将这些村庄的每个指标的增长率求取平均值，同理，依次选择0-159各个标号所包含单位对应的村庄，对这些村庄的每个指标的增长率求取平均值。这样一来，就某一单项评价指标而言，将各个标号所对应该指标的增长率平均值进行排序，建立基于单项指标增长率平均值的排序模型，从而得出针对某一单项评价指标帮扶业绩明显的帮扶单位。

该排序模型流程图如下：

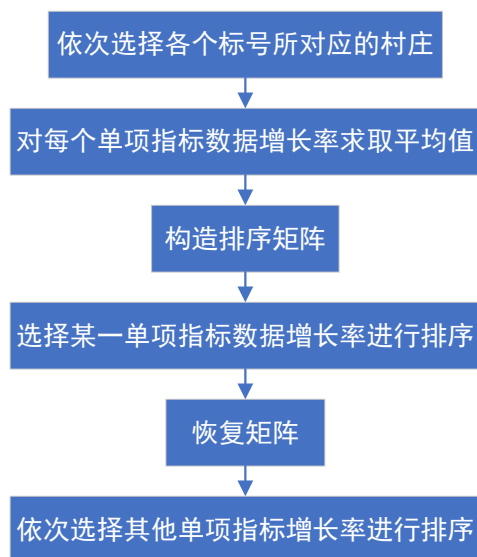


图5：排序模型流程图

至此，我们建立出建立基于单项指标增长率平均值的排序模型如下：

（1）构造排序矩阵

设标号0-159所对应村庄的居民收入(SR)增长率平均值分别为 $\alpha_{SR0}, \alpha_{SR1}, \dots, \alpha_{SR159}$ ，产业发展(CY)增长率平均值分别为 $\alpha_{CY0}, \alpha_{CY1}, \dots, \alpha_{CY159}$ ，居住环境(HJ)增长率平均值分别为 $\alpha_{HJ0}, \alpha_{HJ1}, \dots, \alpha_{HJ159}$ 、文化教育(WJ)增长率平均值分别为 $\alpha_{WJ0}, \alpha_{WJ1}, \dots, \alpha_{WJ159}$ 、基础设施(SS)增长率平均值分别为 $\alpha_{SS0}, \alpha_{SS1}, \dots, \alpha_{SS159}$ 。

结合式（6-2-3）增长率的计算公式，将标号与各个指标的增长率平均值构造成一个160x6的排序矩阵：

$$\begin{bmatrix}
 0 & \alpha_{SR0} & \alpha_{CY0} & \dots & \alpha_{SS0} \\
 1 & \alpha_{SR1} & \alpha_{CY1} & \dots & \alpha_{SS1} \\
 2 & \alpha_{SR2} & \dots & \dots & \dots \\
 \dots & \dots & \dots & \alpha_{WJ158} & \alpha_{SS158} \\
 159 & \dots & \dots & \alpha_{WJ159} & \alpha_{SS159}
 \end{bmatrix}$$

（2）选取单项指标进行排序

对每个单项评价指标而言，从排序矩阵中单独选取该指标所对应得那一列，利用计算机软件对该列的增长率平均值 α 进行降序排序，根据排序结果即可得出各单项评价指标帮扶业绩明显的帮扶单位。

6.3.2 问题三模型的求解

该问题在前两问的基础上，进行了更加细致的评价，选择了单项指标对帮扶单位的绩效进行评价，对于该问题的求解过程大致分为以下2步：

Step1：得出排序矩阵

由于数据过于繁多，部分矩阵如下表所示，完整矩阵详见附件：“jiegua3.xlsx”。

表8：部分排序矩阵

0	-0.16477	0.016326	0.329099	0.055403	0.234771	0.023215
1	0.017693	-0.02951	0.15246	-0.13817	-0.10018	-0.11234
2	-0.23659	-0.27092	-0.21418	-0.26374	-0.27721	-0.31582
3	-1.1616	0.666466	-0.53598	-0.60321	-0.47542	-0.4829
4	-0.26677	-0.92658	-1.2465	-1.00183	-1.30357	-0.94967
5	0.121446	0.483491	0.568274	-0.94436	-0.89895	-0.82682
6	-0.16144	-1.16191	-4.43454	-1.10761	-5.36079	-2.81736
7	-0.74315	1.156745	0.200018	-1.24971	-0.78407	-0.222
...	...	...	...	...	...	...
159	-0.77246	-1.1631	0.679493	0.02203	-0.31819	-0.0055

Step2：选择单项指标进行排序

我们用2020年的指标值减去2015年的指标值，得到每一个指标的差，差值越大说明该类帮扶单位在对应的指标上绩效高，得到差值图示如下：

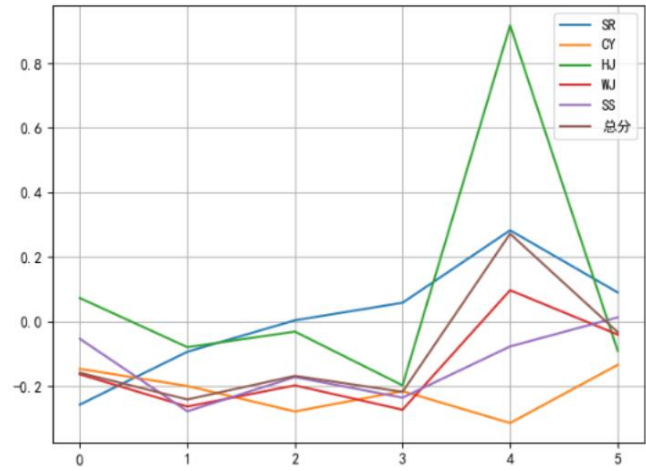


图 6：指标差值

其次我们对每一指标进行具体排序分析，依次列出每项指标的帮扶单位排序结果如下：

① 居民收入

通过计算机软件对居民收入这一指标的增长率平均值进行排序，部分排序结果如下：

表9：居民收入的部分单位排序

20	3.812692
106	3.086956
19	3.034033
55	2.900876
16	2.619129
41	2.384573
29	2.012194
101	1.859347

由上表可知，居民收入这一单项评价指标前五名的帮扶单位编号为20、106、19、55、16。

② 产业发展

通过计算机软件对居民收入这一指标的增长率平均值进行排序，部分排序结果如下：

表10：产业发展的部分单位排序

80	3.534779
157	2.521601
118	2.085688
85	1.904304
102	1.737978
19	1.603025
27	1.370579
45	1.348389

由上表可知，产业发展这一单项评价指标前五名的帮扶单位编号为80、157、118、85、102。

③ 居住环境

通过计算机软件对居民环境这一指标的增长率平均值进行排序，部分排序结果如下：

表11：居住环境的部分单位排序

18	1.368631
154	1.298602
73	1.055104
59	0.704038
65	0.703499
159	0.679493
82	0.602389
5	0.568274

由上表可知，居住环境这一单项评价指标前五名的帮扶单位编号为18、154、73、59、65。

④ 文化教育

通过计算机软件对文化教育这一指标的增长率平均值进行排序，部分排序结果如下：

表12：文化教育的部分单位排序

8	3.843607
72	2.590973
155	1.207743

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/056123141003010104>