PCA分析中国国债的总结

通过Python爬虫，从中债网上爬取了2002年到2017年全年度每日债券收益率详细数据，存储在16个Excel表格当中，共有数据大概75000余条，但是这些数据年度跨度较大，利率期限结构发生了变化，从最初的19个标准期限（2002年 ~ 2009年），变化到了21个标准期限（2010年 ~ 2015年），然后变化到了16个标准期限（2016年 ~ 2017年）。为了统一维度，我选择了16年当中共有的14个标准期限，这些期限包括0年, 0.08年, 0.17年, 0.25年, 0.5年, 0.75年, 1年, 3年, 5年, 7年, 10年, 15年, 20年, 30年。通过Python程序对这16年的数据进行筛选，共筛选出3996条标准期限数据，这些数据包含前文所提的14个维度。

通过SPSS软件和Python程序的两次分析，确保得到了正确的分析结果，结果包括提取出的所有主因子、各维度相关系数矩阵、总方差解释（贡献度）、碎石图、成分矩阵以及各指标对应系数。

因为主要目的是主成分分析，所以最关心的应该是贡献度详细（总方差解释），下面即为详细：



查看初始特征值的累计项，可以看到第一个主因子解释了期限结构77.875%的变化，前两个主因子解释了期限结构91.803%的变化，根据数据挖掘主成分分析法的原则，超过85%的贡献率即可提取主因子，但是为了尽可能的逼近事实，解释足够多的期限结构变化，我选择前三个主因子，前三个主因子的累计贡献率达到96.071%，即可以认为前三个主因子可以解释期限结构96.071%的变化。

前三个主因子初始特征值及贡献度：

|  |  |  |
| --- | --- | --- |
| 初始特征值 | 贡献度（%） | 累计贡献度（%） |
| 5.074 | 77.875 | 77.875 |
| 0.907 | 13.928 | 91.803 |
| 0.278 | 4.268 | 96.071 |

不难发现，第一主因子对于利率期限结构影响最大，第二主因子次之，第三主因子影响最小。

根据SPSS软件的分析结果，得到成分矩阵：



根据主成分分析理论，我们可以得出各指标对应系数等于成分矩阵对应项除以初始特征值的平方根，即：

但是通过Python程序，可以很快的得到各项指标对应系数：



由于数据量比较大，可能这样的表还不太直观，下方展示两张Excel绘制的图标，能够更加直观的展示三大主因子对于利率期限结构的影响，

柱状图：

折线图：

图中各点代表的是各指标对应系数，所以当系数大于0时，对指标是正影响；当系数小于0时，对指标是负影响。通过上方的图表，不难发现第一主因子对于14期标准期限的收益率而言都是正值，并且全部为了0.3左右，相当稳定，所以第一主因子对于所有期限收益都有同向的影响，可以理解为影响收益率曲线的平移；第二主因子对于早期期限收益而言，是负影响，而对于长期收益而言，是正影响，所以第二主因子影响了收益率曲线的斜率；第三主因子对于早期和长期收益而言，均为负影响，对于中期收益却是正影响，它的变化使得早期和长期收益朝同向发展，而中期收益朝相反方向发展，所以第三主因子影响了收益率曲线的曲度。

以上研究与国外的研究结果基本一致，根据国外研究，以上三个主因子分别可称为水平因子、斜率因子和曲率因子。

根据贡献度的数据，我们可以认为水平因子对收益率曲线走势的影响最大对收益率曲线的走势起主导性作用，斜率因子对收益率曲线走势的影响次之，曲率因子对于收益率曲线走势的影响最小。但是对于不同的期限，三个因子的影响也有所不一，短期债券只对于水平因子敏感，长期债券对斜率因子、曲率因子都敏感。