## 本阶段成果

上一阶段中，已经完成了开发环境的搭建，这一阶段，主要要完成对债券收益率数据的PCA分析，以及使用蒙特卡洛模拟法计算债券的在险价值。下面对这些成果进行阐述。

**通过对中国国债数据的PCA分析得到以下结论：**

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。

**通过研究中国国债和美国国债，可以得出以下结论：**

1. 中国国债和美国国债都可以通过三个主成分解释，并且可以解释收益率曲线绝大部分的变化。三个主成分可以解释中国国债96.07%的变化，解释美国国债99.78%的变化。这三个主成分分别可以归纳为水平因子、斜率因子和曲率因子；其中，水平因子主要影响收益率曲线的平移，斜率因子主要影响收益率曲线倾斜度的变化，曲率因子主要影响收益率曲线的曲度。



中国国债PCA碎石图



美国国债PCA碎石图

1. 在对美国国债的研究中，第一主成分（水平因子）可以解释92.75%的收益率曲线变化，但是中国国债的第一主成分（水平因子）只能够解释77.87%的收益率曲线变化。通过后续对SHIBOR以及LIBOR的研究，可以发现这可能是由于中国的基准利率规范不够市场化，国债购买者通过国债收益率曲线获取到的信息不足，所以导致资产定价不够稳定。
2. 中国国债的第二、第三主成分对收益率曲线走势的影响远大于美国国债，导致收益率曲线的变化更加复杂，造成这一情况的原因可能是由于中国债券市场还不够成熟，可供购买者选择的产品不多，所以市场上存在许多炒作和投机现象，使得收益率曲线的变化更具有复杂性。
3. 中国国债的第四个主成分还有着2.12%的贡献率，美国国债的第四个主成分贡献度却只有0.12%，这说明相对于美国债券市场，中国国债未被分解出的成分更多，这使得之后的预测过程相对于美国国债而言会出现更多的未知因素，中国债券市场的波动性也会更大。





上面的两张图是中国国债以及美国国债的不同到期期限债券的日收益率，美国国债有3061个数据点（3061天的日收益率数据），中国国债包含有3900余条日收益率数据。从上面的两张图我们可以发现美国国债多年（2006~2018年）以来收益率一直维持在一个小的范围内，变化相比于中国国债而言比较简单而且少，中国国债在这段时间内（2002~2017年）收益率数据变化次数比较多，并且变化比较复杂。此外我们可以发现美国国债收益率曲线中的多条曲线（即多个维度）的走势具有高度相似性，可以理解为形状大致相同，所以多个维度之间的相关系数会比较大，提取出的主成分中第一主成分贡献度会比较高，这两个结论符合我们之前所做的推断。

**通过研究SHIBOR和LIBOR，可以得出以下结论：**

1. SHIBOR与中美国债类似，也可以由主成分分析法（PCA）提取出3个主成分，这三个主成分累计贡献度可以达到97.89%，第一主成分贡献度就达到了84.71%，可以说SHIBOR的主成分构成有点类似美国国债了，因为影响的因素比较少，所以SHIBOR的曲线相对而言会比较平稳，变化不会十分复杂。



SHIBOR收益率曲线

1. 对LIBOR进行主成分分析可以发现其第一主成分贡献度达到了惊人的98.85%，第二主成分就只有0.871%的贡献度了，即通过一个主成分就可以比较完善地解释LIBOR收益率曲线的变化了。



LIBOR收益率曲线

通过上面这张LIBOR收益率曲线图，我们可以发现LIBOR收益率数据的多个维度之间具有高度的相关性，走势几乎是完全一致的，只不过维度之间具有垂直方向的差异，即不同期限债券收益不同，这几乎完全是由水平因子导致的，所以LIBOR的第一主成分（水平因子）贡献度会达到惊人的98.85%。



上表是LIBOR收益率数据的相关系数矩阵，其中的每一项都代表了LIBOR收益率数据两个维度之间的相关系数，可以看到在这个相关系数矩阵中，几乎每一项都十分接近于1，所以LIBOR收益率数据的各个维度具有高度一致性。

**此外，通过对债券数据进行蒙特卡洛模拟计算，得到VaR的过程如下：**

假设现有的回报数据是服从正态分布的，那么由于回报数据服从正态分布，所以可以使用德尔塔-正态法简化计算。德尔塔-正态法需要3个参数，置信度对应分位数、正态分布标准差以及持有期t。通过对历史数据进行计算，可以求得标准差和平均值，然后通过这两个参数可以使用随机数模拟出一个新的正太分布，然后可以作为计算在险价值的新数据。

具体可以通过如下步骤进行计算：

1. 求历史数据的标准差以及平均值。
2. 使用随机数模拟至少10000个数据，服从标准差为以及平均值为的正态分布。
3. 求出指定置信度对应的分位数，以及新数据的标准差。
4. 根据德尔塔-正态法的公式计算在险价值。

**进行实证分析，结果如下：**

查询中债网VaR数据，可以得到2018年4月20日代码为1282337的债券的最新在险价值数据。

具体如下：



通过在Wind财经软件上查询，可以得到3月1日到4月20日1282337号债券最新的财经数据。

大致如下：



其中包含中债网估计的收益率数据，可以利用该数据进行在险价值预测。通过Python程序计算，这组收益率数据的平均值为mu = 5.28，标准差为0.2。利用Python的随机数生成模块，可以得到一万组模拟收益率数据，数据如下：



通过Python程序的matplotlib模块，可以描绘出这一万组数据的分布情况，如下图：



所以通过这一万组数据，以及它的正态分布模型，可以进一步利用德尔塔-正态法计算在险价值VaR。

公式如下：

所以通过Python程序的计算，可以得到置信水平为95%，持有期为1时，在险价值Var = 0.3342，截图如下：



通过Python程序的计算，可以得到置信水平为99%，持有期为1时，在险价值VaR = 0.4727，截图如下：



通过对比，可以发现与中债网所登记的数据基本吻合，由于数据量有限，以及随机模拟存在误差，所以对比时还是有一点小误差。

## 下一阶段计划

本阶段基本完成了算法的研究工作，下一阶段开始着手开发软件，主要包括以下工作：

1. 整体架构、整体方案、整体开发流程的拟定。
2. 利用Python完成业务逻辑代码的开发。
3. 利用Node.js完成服务器的搭建以及后台接口的开发。
4. 利用Python完成数据爬虫的开发。
5. 利用MongoDB完成数据库开发。
6. 基于React框架的前端软件开发。