**摘 要**

21世纪是计算机行业与金融行业蓬勃发展的黄金时代，我国的债券交易产业也在进入21世纪之后发展得如火如荼，然而随着金融产业的发展，债券交易市场越来越具有复杂性，许多不可预知的风险威胁着债券交易市场，传统的金融知识以及金融工具已经很难再保障债券交易的安全性，这就需要使用新的工具、新的方法来为债券交易进行保驾护航。众所周知金融知识中运用了许多的数理统计、线性代数与概率论知识，计算的过程中还需要采集大量的历史数据进行分析和推断，使用这些方法进行人工计算往往需要消耗大量的人力物力，但是计算机与大数据概念的出现，使得金融问题的解决有了更高效的方案。该毕业设计主要解决了以下一些问题：

1. 使用Python爬虫获取债券交易数据，实现了用户自定义收益率数据的收集。
2. 通过PCA主成分分析方法研究债券标准期限结构，提取出了3个主成分，计算出了三个主因子对固定收益产品价值波动的贡献度，并分析了这些主因子对债券期限结构的影响。
3. 通过蒙特卡洛模拟法计算了给定债券的在险价值（风险价值），使得用户对于债券的投资风险有了直观的概念。
4. 通过插值法实现了基于Nelson-Siegel模型的收益率曲线拟合功能，使得用户能够拟合自定义的收益率曲线。
5. 通过Web前端技术结合Node Express代理和MongoDB数据库实现软件开发，为用户提供数据展示、数据查询、曲线拟合、PCA分析、VaR计算等功能。

关键词：收益率曲线；Nelson-Siegel；PCA；数据采集；VaR； Node Express；MongoDB

Abstract

The 21st century is the golden age for the booming development of the computer industry and financial industry. China's bond trading industry is also developing rapidly in the 21st century. However, with the development of the financial industry, the bond trading market is increasingly complex and unpredictable. The risk poses a threat to the bond trading market. Traditional financial knowledge and financial instruments can no longer guarantee the security of bond trading. This requires to use new tools and new methods to protect bond transactions. It is well known that many mathematical statistics, linear algebra and probability theory are used in financial knowledge, and a large amount of historical data needs to be collected in the calculation process for analysis and inference. Manual calculation using these methods often requires a lot of manpower and material resources, But with the emergence of computer and the concept of big data, a more efficient solution has been made to the solution of financial problems. This graduation design mainly solves the following problems:

1. Using Python spiders to obtain bond transaction data, realized the collection of user-defined yield data.
2. Using the principal component analysis(PCA) method to study the bond term structure, extracting three principal components, calculating the contribution of the three main factors to the value fluctuation of fixed income products, and analyzing the term structure of bonds with these principal factors’ impact.
3. Calculate the value at risk (VaR) of a given bond through the Monte-Carlo simulation method, so that the user has an intuitive concept of the investment risk of the bond.
4. The Nelson-Siegel model-based yield curve fitting function is implemented by the interpolation method, so that the user can fit a customized yield curve.
5. Software development through Web front-end technology combined with Node Express agent and MongoDB database to provide users with data display, data query, curve fitting, PCA analysis, VaR calculation and other functions.

**Key words**: yield curve; Nelson-Siegel; PCA; data acquisition; VaR; Node Express; MongoDB

目录

第1章 绪论

贵阳高登世德金融科技公司是一家中国资产证券化行业内顶尖的金融科技公司，主要负责为金融机构设计资产证券化产品以及优化证券化产品的评级优化，而在这家公司的实习促进了这个毕业设计的诞生，公司的客户需要一款能够实现债券数据采集、分析、期限结构影响因子挖掘以及在险价值计算的软件，所以这个毕业设计的主要目的就是设计一款符合客户需求的债券分析软件。

1.1 研究背景

在停止发行国债22年之后，1981年我国恢复发行国债，从1981年以来，中国政府债券市场已经经历了近30多年的发展。中国国债在筹集财政资金、支持国家重大项目建设、改善经济结构等多方面都发挥了很大的作用。在这近三十年的发展过程中，政府债券规模不断扩大，政府债券品种不断完善，政府债券市场也从无到有，建立了多层次的政府债券市场体系。但与此同时，中国国债市场还具有许多的不足，比如立法基础薄弱、市场分割、投资者结构不合理、市场流动性不足等等。

我国最早在1984年发行了企业债券，到现在为止，经历了萌芽阶段（1984-1986）、快速发展阶段（1987-1992）、整顿阶段（1993-1995）和规范发展阶段（1996至今）。目前而言，我国上市公司债券规模还不够大，远远落后于股票的发行，但是从1984年发行企业债券以来，我国企业债券发行总量有了大幅度增长，发行主体也越来越呈现出多元化的趋势，企业债券品种越来越丰富，其作为投资工具的地位也越来越重要。随着对企业债券的限制越来越少，企业债券发行节奏明显加快。发行债券的品种和主体都越来越丰富。

伴随着我国国债和企业债市场的飞速发展，诞生了一些优秀的债券数据分析管理企业，比较著名的比如中债登、万得、和讯等企业，这些企业积累了大量的中国债券市场交易数据，比如收益率数据、估值数据、VaR数据、成交指数等等，这些企业还拥有了比较丰富的国内债券数据处理经验和数据，比如中债登采用独特的Hermite插值法拟合中国国债收益率曲线，采用历史模拟法实现VaR的计算。通过这些优秀的企业，可以获取到许多宝贵的数据以及数据处理方法，进而可以对债券进行评估定价和风险预测，达到毕业设计的目的。

1.2 研究目的和意义

在购买债券时，应该如何判断一支债券是否值得购买呢？债券购买者往往通过收益率曲线来判断，研究债券收益率曲线具有重要的意义,对于投资者而言,可以用来作为预测债券的发行投标利率、在二级市场上选择债券投资券种和预测债券价格的分析工具；对于发行人而言,可为其发行债券、进行资产负债管理提供参考，所以收益率曲线以及收益率数据的收集与研究就显得十分有价值。

本设计通过从各大证券交易所收集收益率数据并且对这些数据进行曲线拟合、主因子提炼、分析主要因子的影响以及判断未来曲线的走势预测债券购买价值并给出投资策略。

本设计的意义：通过对已有数据的分析挖掘，提炼出曲线以及方程，根据PCA方法分析主要影响因子，根据这些因子可以给出未来的债券投资策略，达到预测债券走势的目的。

虽然该毕业设计主要虽然研究的是固定收益债券定价和投资策略分析，但是主要研究对象还是收益率曲线。要了解设计的目的，首先要了解收益率曲线。

收益率曲线是根据不同到期期限债券的收益率所绘制的图形，它反映市场当前的收益率水平。收益率曲线不是历史数据图形，不能反映一段时间内的收益率水平，反映一段时间内收益率水平的图形称为历史价格图或者历史收益率图形。

1.3 国内外研究现状

在中国，由于包括国债市场在内的债券市场起步较晚，流动性和市场化程度尚需进一步加强，有关利率期限结构和国债收益率曲线的研究也相对较薄弱，这增加了对国债收益率曲线进行研究的必要性和迫切性。国内最早的关于收益率曲线的公开研究文献是杨大楷和杨勇姚长辉和梁跃军。

由于中国债券市场的发展起步较晚，对中国债券市场的理论研究落后，收益率曲线研究始于90年代后期。近年来值得一提的工作是应用国外广泛使用的模型对我国债券收益率曲线进行实证研究和比较分析。研究表明，针对中国的现状短期债券发行量小，长期性差，中期债券比例大，不健康的期限结构，Nelson-Siegel模型适用于构建中国国债收益率曲线。应当利用中国债券市场数据，利用国际上公认的几种收益率曲线拟合模型进行实证研究，寻找适合中国债券市场的拟合收益率曲线的方法。

Diebold和Li（2006）、Diebold、Rudebusch、Aruoba（2006）、Diebold Piazzesi和Rudebusch（2005）表明，在美国的封闭经济环境中，广义的Nelson-Siegel模型精确地预估了收益率曲线动态，提供了良好的预测。

Diebold-Li分解单个国家的Nelson-Siegel收益率曲线是：

其中表示月债券的连续复利的零息票名义收益率，、、、、均为待定参数，并且具有标准偏差的干扰。继Diebold和Li之后，通过允许参数随时间变化来动态化模型：

式中解释为为潜在因素，特别是，如Diebold和Li所示，它们分别是水平参数，斜率参数和曲率参数，因为这些因子参数是常数，所以这是一个关于 的递减函数和凹函数。由于收益率因子随时间而变化，这种广义的Nelson-Siegel模型可以生成各种随时间变化的收益率曲线形状。

1.4 研究内容

该论文主要描述一款基于Web前端页面、Node Express后台接口、Python数据采集分析脚本、PCA算法、Monte Carlo算法以及Mongo DB数据库的债券数据采集分析软件的开发过程。这款软件实现了用户自定义债券收益率数据的查询、自定义债券收益率曲线的拟合、自定义债券期限结构数据的PCA分析以及自定义债券的在险价值计算。

这款软件的核心在于后台Python脚本程序，通过Python脚本程序，实现了债券数据采集和过滤、债券数据的PCA分析、债券收益率曲线的拟合以及债券在险价值的计算，这一部分算法以及数据都是整个软件的核心。通过Python程序采集到数据后并进行过滤整理后，将所有数据分类保存在MongoDB数据库中，以便之后Node脚本程序和Python程序对数据进行查询调用。在前端部分，采用了目前最为流行的MVC框架——React进行开发，负责呈现软件的用户交互界面，包括数据展示、收益率曲线展示、用户输入界面等等。后台部分，使用当前比较流行的Node.js以及Node.js构建的Express框架作为后台代理，实现了RESTful风格的后台接口开发以及服务器搭建。另外通过npm模块实现了Node脚本程序与Python脚本程序之间的交互，使得Node服务器可以调用Python程序进行数据处理。

1.5 论文基本结构

这篇论文的正文整体上分为6章，按照总-分-总的架构讲述整个软件设计过程中的思路、原理、算法、心得等方面。

第一章为绪论，主要讲述了这个毕业设计的需求来源、设计研究的背景、目的意义、当前国内外现状、设计的研究内容以及论文的大致结构。

第二章主要讲述整个软件的底层算法，因为这些算法是整个软件的灵魂，所有的逻辑业务都基于这些算法展开，所以把算法的讲述放在第二章，主要包括PCA主成分分析法、Nelson-Siegel模型、基于Monte Carlo模拟法的VaR计算法。

第三章主要讲述整个软件的架构设计，包括整体架构设计、后台架构、前端架构以及前后台交互的设计。

第四章主要讲述服务端（后台）软件设计，包括Python脚本程序设计、Node脚本程序设计、Node Express服务器的搭建、RESTful API接口设计和后台数据库结构设计。

第五章主要讲述前端Web软件设计，包括React框架、Highcharts绘图插件、ant-design界面开发模块、axios前后台之间的AJAX交互模块。

第六章主要讲述整个软件开发过程中得到的结论和见解，以及对于未来的展望。

第2章 软件算法研究

2.1 PCA算法

* + 1. 算法介绍

PCA是一种分析高维数据的方法，能够产生一组新基，当数据在新基上投影时能够有效的识别数据之间的相似性和主要的差异。另外，PCA的一个主要优点是数据的信息主要保存在前几个主成分中，那么，可以把高维数据进行降维处理而不会丢掉很多信息。PCA的思路从数学的角度，主成分分析即构造原变量的一系列线性组合，使各线性组合在彼此不相关的前提下尽可能多地反映原变量的信息，即使其方差最大。可以证明，求某一数据的主成分，等价于求它的协方差矩阵的各特征值及相应的正交单位化特征向量。按特征值由大到小所对应的正交单位化特征向量为组合系数的原变量的线性组合分别为该数据的第一、第二、…主成分，而各主成分的方差等于相应的特征值。

PCA的做法步骤：

1. 第一步：获得数据
2. 第二步：各数据减平均值
3. 第三步：计算数据的协方差矩阵
4. 第四步：计算协方差矩阵的特征值、特征向量
5. 第五步：选择前几个主成分构成新基
6. 第六步：把原数据在新的基下进行投影
7. 第七步：过滤杂音，并回到原基
   * 1. PCA研究中国国债

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：



图2.1 中国国债数据各维度累计贡献度数据

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。

* + 1. PCA综合研究与结论

除了使用PCA算法研究中国国债期限结构之外，还使用PCA算法研究了美国国债（2006~2018年）、SHIBOR（2006~2018年）、LIBOR（2001~2013年）的期限结构，综合以上所有的研究结果进行横向对比，得到了一下结论：

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。

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。

2.2 Nelson-Siegel模型

* + 1. 曲线拟合的概念及目的

通常，债券购买者得到的期限结构数据是二维空间中一系列离散的点，但是为了更加直观的观察债券收益走势，更加合理的利用期限结构获取更多有用的信息，通常会利用这些离散的点构造一条平滑的曲线，即为收益率曲线，利用离散的点构造这条收益率曲线即为收益率曲线拟合。

比如，我们从中债网上可以获取到2018年4月13号当日的债券收益日信息，填写在表格中，即：

|  |  |
| --- | --- |
| Maturity(Year) | Yield(%) |
| 0.0 | 2.2316 |
| 0.08 | 2.5617 |
| 0.17 | 2.5693 |
| 0.25 | 2.6842 |
| 0.5 | 3.0242 |
| 0.75 | 3.1843 |
| 1.0 | 3.2013 |
| 3.0 | 3.3682 |
| 5.0 | 3.5205 |
| 7.0 | 3.6916 |
| 10.0 | 3.7062 |
| 15.0 | 3.9689 |
| 20.0 | 4.0079 |
| 30.0 | 4.1207 |
| 40.0 | 4.162 |
| 50.0 | 4.1808 |

使用Excel描绘散点图：



由此可以得到一个期限结构数据描绘的散点图，但是这张散点图不够直观，如果债券购买者想要获得到期期限为35个月的债券的收益率数据时，将无从下手，所以需要根据这些二维空间中的散点，构造一条光滑曲线，这条曲线并不只是简单的将散点用光滑曲线连接，还需要对未给出的节点做出价格推测。这就是曲线拟合以及曲线拟合的目的和意义。

下图是中债网所给出的收益率曲线：



* + 1. 曲线拟合方法及模型的选择

为了得到一条这样的拟合曲线，首先要选择合适的拟合方法和模型。常见的收益率曲线拟合方法或模型有多项式插值法、Hermite模型插值法、Nelson-Siegel模型、Hull-White等等。接下来会使用同一种数据TestData.csv使用其中几种模型或方法进行曲线拟合，然后对拟合精度进行评估，使用到的评估标准是均方根误差（RMSE）。

TestData.csv数据：

|  |  |
| --- | --- |
| 0 | 2.2316 |
| 0.08 | 2.5617 |
| 0.17 | 2.5693 |
| 0.25 | 2.6842 |
| 0.5 | 3.0242 |
| 0.75 | 3.1843 |
| 1 | 3.2013 |
| 3 | 3.3682 |
| 5 | 3.5205 |
| 7 | 3.6916 |
| 10 | 3.7062 |
| 15 | 3.9689 |
| 20 | 4.0079 |
| 30 | 4.1207 |
| 40 | 4.162 |
| 50 | 4.1808 |

1. 多项式插值法

构造多项式拟合已有收益率数据，然后通过插值构造平滑收益率曲线的方法即为多项式插值法。

多项式中拟合收益率曲线效果最好的是三次多项式，即为立方多项式。可以通过Python程序进行拟合，并查看拟合效果。

首先在Python中通过numpy工具进行三次多项式拟合，拟合的三次多项式系数：

|  |  |  |  |
| --- | --- | --- | --- |
| 三次项系数 | 二次项系数 | 一次项系数 | 常数 |
| 5.77597e-05 | -0.00540 | 0.15539 | 2.74509 |

然后计算出所有的估计值，然后计算出均方根误差进行输出，下面是程序运行的结果：



说明对于TestData.csv这一组数据而言，利用三次多项式进行拟合，计算出的均方根误差是0.045。

下面是拟合数据与原始数据的对比：



从图中可以看出，三次多项式的拟合效果较差，误差较大，甚至出现了折现的情况，不符合我们需要平滑曲线的要求。

1. Hermite插值法

Hermite插值法在国外使用的不是很多，但是这里之所以要提Hermite插值法，是因为它比较适合于拟合中国债券市场的收益率曲线，比较有代表的使用者就是中债网，中债网使用Hermite多段插值法进行收益率曲线拟合，具有较好的拟合效果。

中债网使用的是多段插值法，具体做法是在每两个节点之间，使用三次Hermite进行插值，所以根本上而言，这还是使用的三次多项式模型，但是这种方法不仅要求节点处的值要相同，而且在节点处的倒数也要相同，所以拟合这与一般的三次多项式拟合存在不同。

接下来我们仍然使用Python程序进行拟合，并计算RMSE进行效果比较，这里Python程序没有实现Hermite算法，使用的是已有的工具包——numpy。通过Python程序拟合，我们得到了拟合的三次Hermite多项式。

其各项系数如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 三次项系数 | 二次项系数 | 一次项系数 | 常数 |
| 7.21997e-06 | -1.34920e-03 | 7.77377e-02 | 2.74239 |

拟合后计算出的均方根误差为：



这说明相对于TestData.csv这组数据而言，Hermite插值法拟合曲线得到的数据均方根误差为0.094，这比三次多项式拟合的效果还要差。

其拟合数据与原数据对比如下：



1. Nelson-Siegel插值法

Nelson-Siegel模型是金融届（尤其是债券行业）大名鼎鼎的一个模型，它是1987年Nelson C.和Siegel A.提出来的一个参数拟合模型，它需要确定的参数少（只有四个），但是拟合的效果却很好，而且每一个参数都具有很强的金融含义。

其具体公式如下：

其中，、、是三个具有金融含义的参数；是与时间有关的常数，可称为时间常数；是到期期限，y是到期收益率。

同样的，通过Python程序进行Nelson-Siege模型拟合，由于有四个参数需要确定，所以会存在多组解，我们通过优化方法，使得拟合出的效果最好（以各项数据方差之和为参考标准），最终效果最好的一组参数即为我们拟合的Nelson-Siegel模型的四个参数。

最终拟合的Nelson-Siegel模型的四个参数如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 4.34119 | -1.28633 | 4.455712e-03 | 0.19455 |

最终计算出的RMSE为：



拟合效果：



结合RMSE来看，三次多项式的拟合效果要好于Nelson-Siegel模型拟合的效果，但是通过曲线观察，可以发现Nelson-Siegel模型的拟合效果要好于三次多项式的拟合效果，这是因为在使用Nelson-Siegel模型拟合曲线的时候，由于其他因素限制，只使用了整数到期期限的数据进行拟合，所以导致其RMSE比三次多项式更低，由于这个原因，可以认为Nelson-Siegel模型的拟合效果好于三次多项式，所以该毕设还是采用Nelson-Siegel模型结合插值法进行收益率曲线的拟合。

* + 1. Nelson-Siegel模型的优点

采用Nelson-Siegel模型进行拟合的原因还有以下几点：

1. Nelson-Siegel在业界久负盛名，非常适合与拟合收益率曲线。根据统计[参考中国动态Nelson-Siegel利率期限结构模型研究]，比利时、芬兰、法国、意大利、西班牙、等多个国家的中央银行采用Nelson-Siegel模型作为利率期限结构的估计方法。还有许多其他国家采用的是基于Nelson-Siegel模型的Nelson-Siegel-Svensson模型，所以Nelson-Siegel模型的名气可见一斑。
2. Nelson-Siegel模型构造简单，只需要计算四个参数即可完成构建。
3. Nelson-Siegel模型的四个参数具有强烈的金融含义，所以容易该模型更理解。之前说过参数是只与时间有关的常数，不同日期的收益率数据可能不相同。如果我们确定了，以、、和为参数，我们可以做出当 = 0.0664时， 、、三个参数的系数随发生变化的图形：



上图实际上就是影响债券收益率走势的三个主成分的因子载荷，可以看出对收益率曲线的影响是恒定的（水平位移），对收益率曲线的影响随着期限的增加而逐渐减小（曲线的斜率），对于收益率曲线的影响随着期限的增加先增加然后逐渐减小（曲线的曲度）。根据之前对中国国债、美国国债的研究，代表着第一主成分，即水平因子；代表着第二主成分，即斜率因子；代表着第三主成分，即曲度因子。所以说，Nelson-Siegel模型的参数有着强烈的金融含义。

* + 1. Nelson-Siegel模型的实际拟合效果

在最终的成品软件中，实现了自定义收益率曲线拟合的功能。整个拟合逻辑使用Python程序实现，通过将用户上传债券的收益率数据（至少5个到期期限的到期收益率数据）传递给Python程序，得到经过Python计算之后的各期限数据，即为拟合的收益率曲线数据，通过这些数据，可以在前端页面展示出一条较为平滑且逼近真实的收益率曲线。为了说明Nelson-Siegel模型的实际拟合效果，接下来会使用最终的软件拟合一期实际的中国国债收益率曲线。

真实的收益率曲线数据如下：

|  |  |
| --- | --- |
| **标准期限(年)** | **收益率(%)** |
| 0 | 2.8352 |
| 0.08 | 2.9012 |
| 0.17 | 2.7707 |
| 0.25 | 2.6882 |
| 0.5 | 2.81 |
| 0.75 | 2.9413 |
| 1 | 2.977 |
| 3 | 3.241 |
| 5 | 3.2505 |
| 7 | 3.6601 |
| 10 | 3.6467 |
| 15 | 3.9351 |
| 20 | 3.9741 |
| 30 | 4.1113 |
| 40 | 4.1499 |
| 50 | 4.1661 |

该数据来源于中债网2018年5月3日的日到期收益率数据，官网所拟合的收益率曲线如下：



使用毕业设计所做的软件，将标准期限数据上传之后，得到了拟合的收益率曲线数据，软件所拟合的收益率曲线如下：



可以看到收益率曲线的走势是大致吻合的，下面是两者关于几个标准期限之外的到期期限对应的到期收益率的对比：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Maturity | 8 | 11 | 13 | 18 | 21 | 28 | 33 | 42 | 46 |
| 软件拟合 | 3.61 | 3.72 | 3.78 | 3.89 | 3.95 | 4.04 | 4.09 | 4.15 | 4.17 |
| 中债拟合 | 3.65 | 3.68 | 3.82 | 3.97 | 3.98 | 4.08 | 4.13 | 4.16 | 4.16 |

通过上表中的数据对比，可以发现相比于中债的拟合值而言，误差基本被限制在0.05之内，达到了较好的拟合效果，符合预期的设计要求。

2.3 Monte Carlo模拟法计算VaR

* + 1. VaR的产生

最初提出VaR这个概念，是为了应对金融风险中的市场风险。金融风险曾给投资者带来过极大的损失，1995年2月，英国历史最悠久的巴林银行宣告破产，原因就是交易员Nicholas Leeson在交易中损失了13亿美金。一度被认为保守的巴林银行的破产给许多银行机构敲响了警钟，然而类似的例子还有很多，所以风险衡量就应运而生了。

最初的风险衡量标准是波动性，但是波动具有双向性，可以是向上波动，也可以是向下波动，如果我们使用波动性来衡量风险，那么向上波动也会是风险的一种，当然投资者不会认为收益是一种风险，他们只会关注赔钱的可能性，所以需要有更加合理的风险衡量标准。VaR风险价值就是这个更加合理的风险衡量标准。

* + 1. VaR的概念

VaR，是Value at Risk的缩写，即在险价值。通常来说，VaR由三个部分组成：时间段、置信水平和损失金额。我们可以用VaR的三个部分来描述某个投资组合的风险，比如说我可以以95%的置信水平确定下个月投资者最多会损失100万美金。或者说给定置信度为95%，某投资组合24小时内的VaR为100万美金，那么认为未来24小时内，该投资组合亏损100万美金的概率为5%。

使用数学表达式表述VaR：

其中：

P代表资产价值损失小于最大可能损失的概率，原意为Probability。

ΔP代表某一投资组合在一定持有期Δt的价值损失额。

VaR代表给定置信水平a下的在险价值，即可能的最大损失。

a代表给定的置信水平。

* + 1. VaR的计算方法

VaR的计算方法主要有三种：德尔塔-正态法、历史模拟法和Monte-Carlo模拟法。下面简单介绍一下这三种方法。

1. 德尔塔-正态法

德尔塔-正态法假定投资组合的回报（return）服从正态分布，所以可以利用正态分布置信度与分位数对应的良好特性来简化计算。

由回报服从正态分布我们就可以得到VaR计算公式如下：

其中：

VaR代表风险价值。

代表给定置信度对应的分位数。

代表回报所服从的正态分布的标准差。

代表持有期。

1. 历史模拟法

历史模拟法的原理是根据历史回报数据的变化来推测未来回报数据的走势，其公式如下：

其中：

代表投资组合在时间k的收益率。

代表时间t的投资权重。

代表组合中第i只证券在时间k的收益。

大致计算步骤如下：

1. 建立一个多年所有市场变量的每日变动数据库。
2. 第一次模拟假设每个市场变量的波动率与数据库覆盖时间的第二天相应变量的波动率相同。
3. 第二次模拟假设各市场变量的波动率与数据库覆盖时段的第二天相应变量的波动率相同。

以此类推，每次模拟就可以计算出一个投资组合的样本值。

1. Monte Carlo模拟法

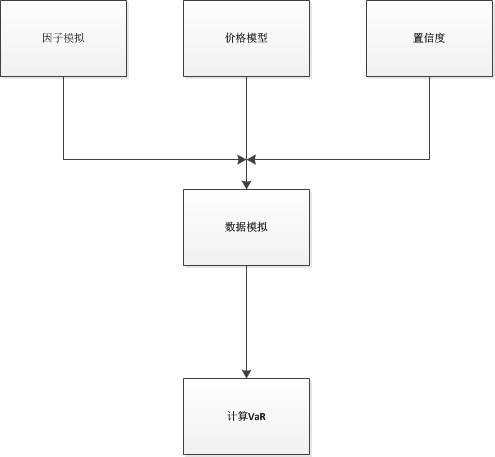
Monte-Carlo模拟法又称为计算机随机模拟方法，是一种基于随机数的计算方法。它起源于第二次世界大战当中的“曼哈顿计划”，为了解决原子弹研制中的中子随机扩散过程，美国科学家提出了这一模拟方法。

蒙特卡洛模拟法利用随机数产生收益率或者资产价格的一个分布。我们可以通过随机数产生一个服从正太分布的数据集来模拟未来数据的变动，从而求得未来可能的损失，得到VaR在险价值。

Monte-Carlo模拟法计算VaR值常常使用以下步骤：

* + 1. 生成随机模拟收益率或者收益。
    2. 重复历史模拟法中的步骤计算VaR。

其原理如下：



在该毕业设计中，使用的就是蒙特卡洛模拟法计算VaR在险价值。

* + 1. 实证分析

在这个毕业设计中，按照要求使用蒙特卡洛模拟法来计算VaR在险价值。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Pα、正态分布标准差σ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σ以及平均值x。
  2. 使用随机数模拟至少10000个数据，服从标准差为σ以及平均值为x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。

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。

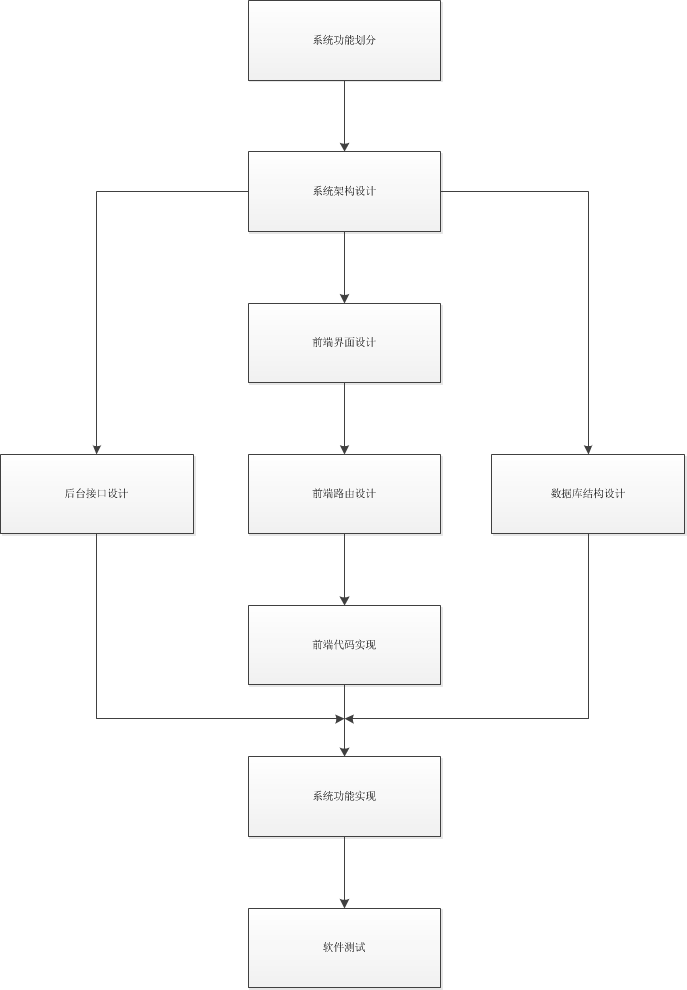
第3章 总体设计方案

3.1 总体开发流程

整个软件的设计开发流程应当遵循以下步骤：

1. 系统功能划分，对应着软件开发中的客户需求分析，在开发之前首先要了解客户需要什么功能，然后才能着手设计开发。
2. 系统架构设计，在了解了客户的真实需求之后，需要对软件设计的整体思路有一个规划，在之后的开发过程中需要遵循这个思路来进行。
3. 软件开发，这一部分按照系统架构的不同，可以分为3个主要部分：前端开发、后台服务器开发和数据库设计。
4. 软件测试，在开发完成之后，需要对代码的可靠性和各种可能出现的异常情况进行测试，做好正向测试与反向测试，保证开发的软件在复杂的环境下具有很高的可靠性。

软件的整体设计流程图如下：



* 1. 总体架构设计

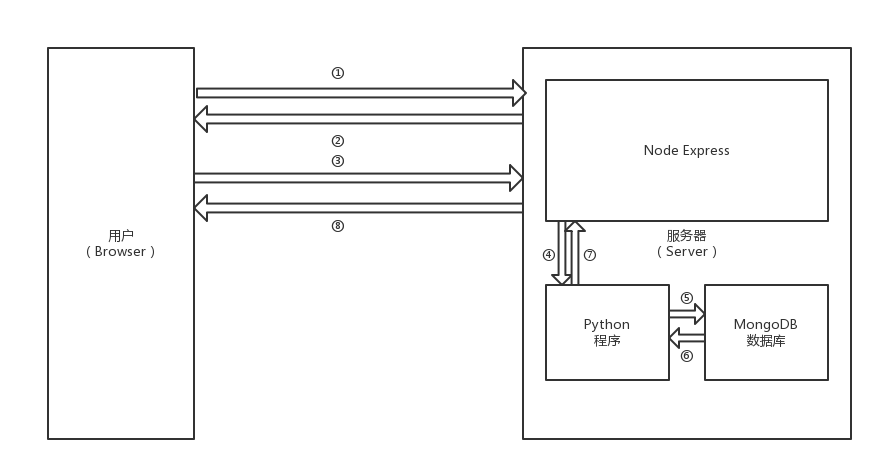
软件架构指的就是软件的基本结构，合适的架构是一款软件成功的重要因素之一。软件架构可以按照结构的不同分为分层架构、事件驱动架构、微核架构、微服务架构以及云架构。其中分层架构是最基本且最常见的一种软件架构，所以该毕业设计就采用分层架构设计软件。

分层架构将软件分成若干个水平层，每一层都有清晰的角色和分工，不需要知道其他层的细节，层与层之间通过接口通信。常见的分层架构将软件设计分为4层：表现层、业务层、持久层和数据库层。由于毕业设计的软件采用NoSQL模式的MongoDB数据库，所以没有太多用于查询数据的SQL脚本，所以用于提示数据的持久层就被取消了，采用3层分层架构设计——表现层、业务层和数据库层，由于设计的实际需求，这三层架构对应成为了用户视图层、Node代理层和应用层。

整体架构设计如下：



其中，用户视图层负责向用户展示视图（数据、图表等），与用户进行交互；Node代理层负责搭建后台服务器，实现后台接口，实现请求派发以及路由控制；应用层负责处理派发的任务，执行指定的代码，返回系统所需的数据。整体工作流程大致如下：



其中，每个步骤所代表的的含义如下：

1. 用户输入网站网址，浏览器向服务器发起HTTP请求。
2. Node Express收到请求，返回HTML页面给用户浏览器。
3. 用户使用软件的某项功能，浏览器向服务器发起HTTP请求。
4. Node Express收到请求以及附带的参数，并把参数传递给任务需要执行 的Python脚本程序。
5. Python程序开始执行，在MongoDB数据库中查询所需要的数据。
6. MongoDB向Python程序返回其所需要的数据。
7. Python执行完所有的任务，并把结果返回给Node程序。
8. Node Express返回数据给用户浏览器，给用户呈现执行结果。

以上基本是一个完整的交互流程，但是有时候不一定需要执行Python脚本程序，Node程序可以直接查询MongoDB数据库中的数据并返回给用户的浏览器。

* 1. 总体解决方案

后台程序主要负责业务逻辑的处理，主要包括数据获取、数据存储、数据分析、数据处理和数据查询等任务。后台开发又分为Python程序开发、Node程序开发和MongoDB数据库开发，其中数据获取、数据分析、数据处理由Python程序实现，服务器搭建、数据查询由Node程序实现，数据存储由MongoDB数据库实现。整个程序设计可以放在同一个文件夹中构成一个完整的解决方案，解决方案的目录结构如下：



其中bonds\_system文件夹代表前端开发的项目文件夹，Data2Mongo代表用于处理数据并存储到MongoDB数据库的Python程序项目文件夹，NodeProject表示用于搭建服务器、查询数据的Node脚本程序的项目文件夹，PythonProject表示用于数据处理、数据分析、数据获取的Python程序项目文件夹。

第4章 后端程序设计

4.1 Python程序设计

Python语法简洁清晰，使用Python编程简明清晰，容易理解。其次Python具有丰富和强大的库。它常被昵称为胶水语言，能够把用其他语言制作的各种模块（尤其是C/C++）很轻松地联结在一起。这就使得在使用Python编写一些难度较大的软件时，能够找到很多已经开发好的模块或者包，简化编程工作，也使得Python这门语言更加强大。使用C/C++作为底层的好处就是Python程序运行速度很快，扩展空间很大，兼容性很好，功能更完善。这也是为什么Python这门语言在当今编程语言百花齐放，开源环境纷繁复杂的情况下，仍然能够占据一席之地，甚至在金融大数据、人工智能、深度学习等方面甚至越来越重要的原因。

第3章3.3节中已经说明数据获取、数据分析、数据处理由Python程序实现，所以本节就按照Python所需执行的任务依次阐述Python程序的实现过程。

* + 1. 爬取数据

1. Python爬虫

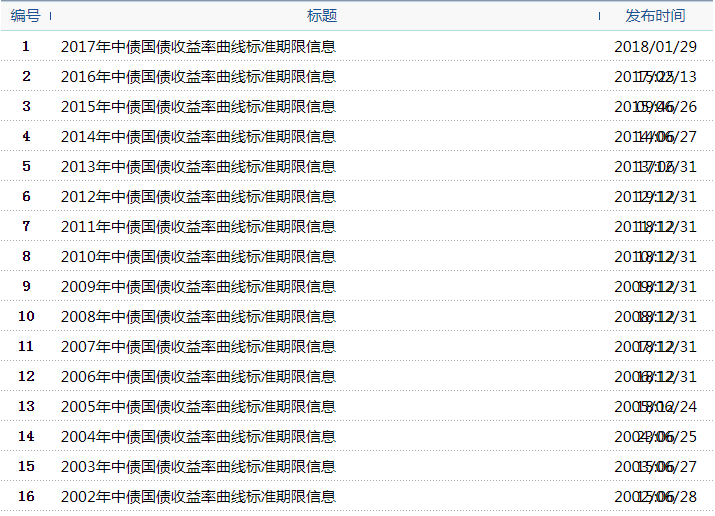
Python爬虫，是一种专门用于抓取网络上指定数据的程序。通过Python爬虫，可以从互联网上抓取大批量的数据，节省大量的人力物力。由于在这次设计中需要从网络上抓取大量的债券数据，所以需要使用到Python爬虫技术。

爬取数据可以分为网页数据爬取和接口数据爬取，根据爬取对象的不同，Python爬虫也大致分为两种，一种直接请求指定接口的数据，另一种通过请求HTML网页文件，然后解析HTML获取数据。本次毕业设计两种方式都有使用到。

接下来以爬取中国国债日收益率历史数据为例，阐述Python程序的开发， 由于数据爬取过程相对而言比较简单，所以没有使用已经封装好的Python 爬虫框架，直接使用Python自带的模块实现。

1. 爬取中国国债数据

本次数据爬取的对象确定为中债网，中债网的中国国债日收益率以Excel文件的形式提供给用户下载：



通过查看其浏览器的控制台可以看到其HTML代码如下：



可以发现是一个ul标签，内嵌了一系列的li标签，形成了一个链接列表，其中的a标签指明了下载Excel数据文件的url和标题，所以只需要解析到这个a标签即可，然后根据对应的url下载服务器资源。

使用到的Python包为urllib.request和pyquery，其中urllib.request用于发起HTTP请求，而pyquery是一个类似jQuery的HTML DOM操作库，但是它是在服务器端对DOM进行渲染和操作，而非浏览器，可以用于对返回的HTML文件进行解析。由于pyquery是第三方模块，所以需要使用pip（Python包管理器）安装。整个程序只有两个函数，一个用于获取Excel下载的url，一个用于根据提供的url下载对应资源，并保存到本机指定位置。

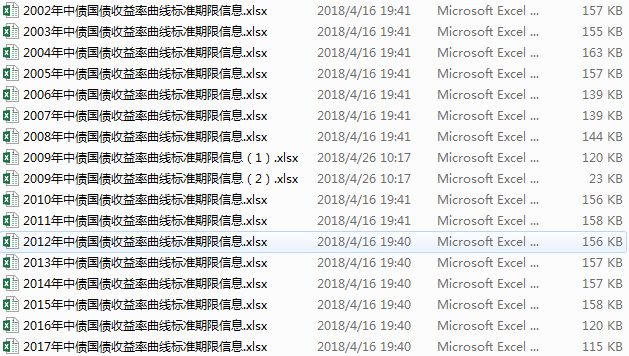
获取url的函数代码如下：



根据url下载对应Excel数据文件的函数代码如下：



依次执行以上两个函数，所有Excel文件会被下载到指定的路径下：



以上就是爬取中国国债日收益率历史数据的方法，其他数据的爬取方式与此大同小异，不再赘述。

* + 1. 处理数据

4.1.1节中，已经将所有的Excel数据文件爬取到了指定目录下，但是程序运行过程中不直接使用Excel文件，最好是将数据存储到数据库中，这样数据结构统一，程序运行时查询数据的速度也远远快于直接访问Excel文件。所以对于下载好的Excel数据文件，还需要对其中的数据进行整理，如果Python进行数据分析，还有可能需要使用CSV格式的文件。接下来仍然以中国国债日收益率历史数据为例进行阐述。

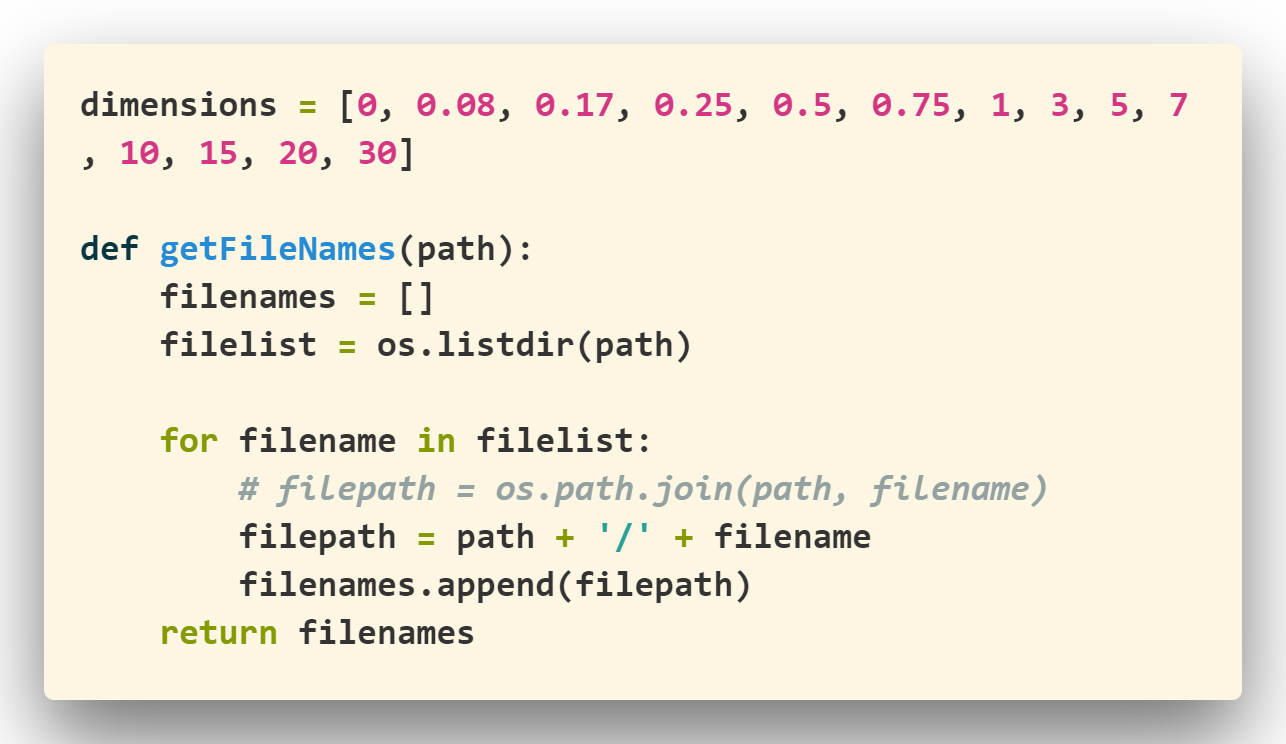
1. 整合成CSV文件

经过对比发现，中国国债由于还在飞速发展的过程中，所以期限结构仍然还不是特别稳定，从2002年1月4号开始到2009年11月27号共有19期标准期限。从2009年11月30号到2015年12月31号共有21期标准期限。从2016年1月4号至今共有16期标准期限。为了便于研究，取所有数据的共有标准期限，然后整合成一个CSV文件，这14期标准期限为：0, 0.08, 0.17, 0.25, 0.5, 0.75, 1, 3, 5, 7, 10, 15, 20, 30，单位是年。

由于需要处理所有的Excel文件，所以需要遍历指定的文件夹，使用Python读取Excel文件需要使用到xlrd模块，xlrd模块是第三方模块，需要使用pip进行安装。

此处进行数据处理的函数也分为两个：getFileNames函数用于获取指定目录下所有文件的文件路径，readXLSX函数用于根据传入的Excel文件路径集合读取所有指定的Excel，如果是dimensions中的共有标准期限，那么就将数据存储到集合中，最后将所有数据写入一个指定的CSV文件中，达到数据过滤整理的目的。

getFileNames函数代码如下：



readXLSX函数代码如下：



依次运行两个函数，可以得到一个汇总数据的CSV文件，总共有4077条数据，14个维度。

1. 存储数据到MongoDB

由于需要遍历目录、读取Excel以及插入数据到MongoDB，所以需要使用到的模块有os、xlrd和pymongo，其中pymongo是一个第三方的用于MongoDB操作的Python驱动，需要使用pip进行安装。该部分代码分为两个函数，getFileNames函数与前一部分相同，用于获取指定目录下所有的文件的路径；readXLSX函数区别于前一部分，因为需要向MongoDB插入数据，所以需要将每一条数据构建为一个字典，然后将所有数据汇总到一个list结构中，最后使用pymongo的MongoClient构造方法构造的对象中的insert\_many方法批量插入到MongoDB中，以便提高插入速度。

readXLSX函数代码如下：



插入MongoDB数据库的代码如下：



* + 1. 分析数据

Python程序要实现的数据分析任务包括PCA分析、收益率曲线拟合以及VaR在险价值计算。由于Python所要实现的数据分析程序要被Node程序调用，所以在本节之后的所有Python程序都需要引入sys模块，以接受外部程序传入的参数。接下来按照不同的任务依次阐述设计过程。

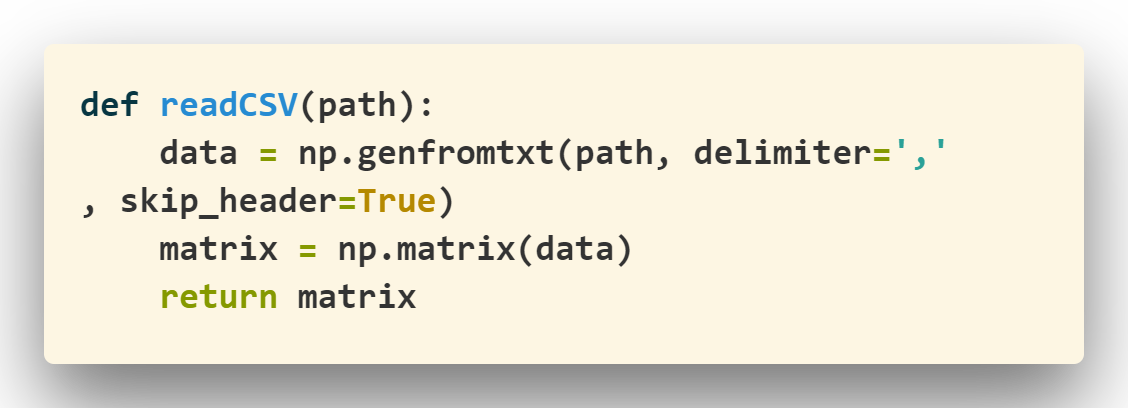
1. PCA分析

在第二章第一节中已经研究了PCA算法的理论，并使用SPSS统计学软件进行了计算分析，接下来使用Python程序进行PCA分析。Node程序调用Python进行PCA分析时，需要将CSV数据的地址以及所要提取的主成分个数传入，Python使用传入的数据进行PCA分析，提取出指定个数的主成分之后，将各个维度的贡献度、累计贡献度以及因子载荷矩阵数据存储到指定路径下的CSV文件中，Node程序通过读取CSV文件中的数据向用户返回分析结果数据。

该程序主要使用第三方模块numpy进行主成分分析，需要先使用pip进行安装，安装之后在文件开头通过import引入。为了防止生成的CSV文件重名，此处使用uuid模块生成全球独一无二的标志，最终作为文件名保存。

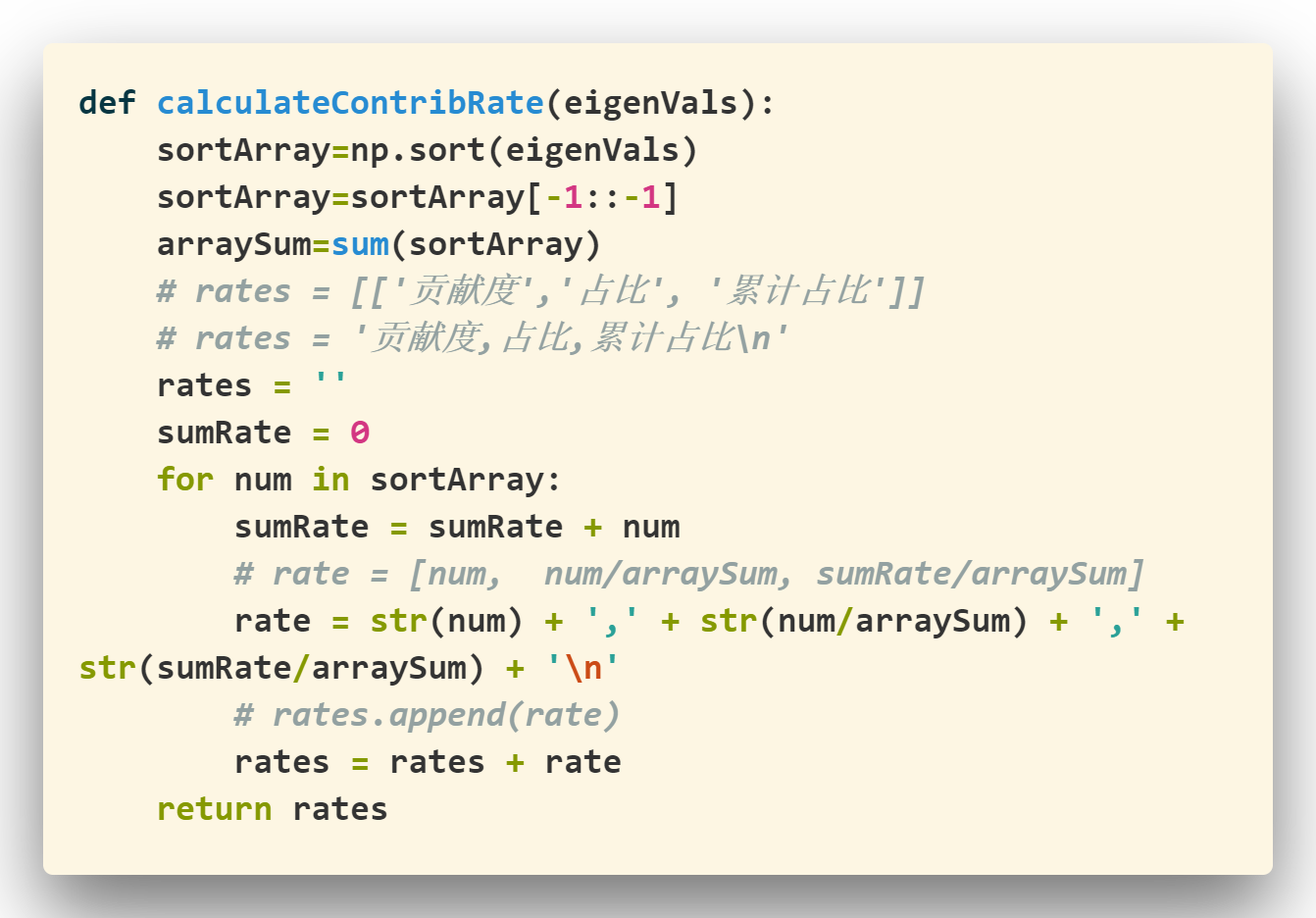
这个程序分为四个主要函数：readCSV函数负责读取指定路径的CSV文件中的数据，返回一个numpy matrix矩阵结构变量；calculateContribRate函数负责计算个维度累计贡献度；D2Matrix2CSV函数负责将numpy matrix变量写入CSV文件；最重要的PCA方法实现PCA分析，传入多维数据，返回特征向量和因子载荷矩阵。

readCSV函数代码如下：

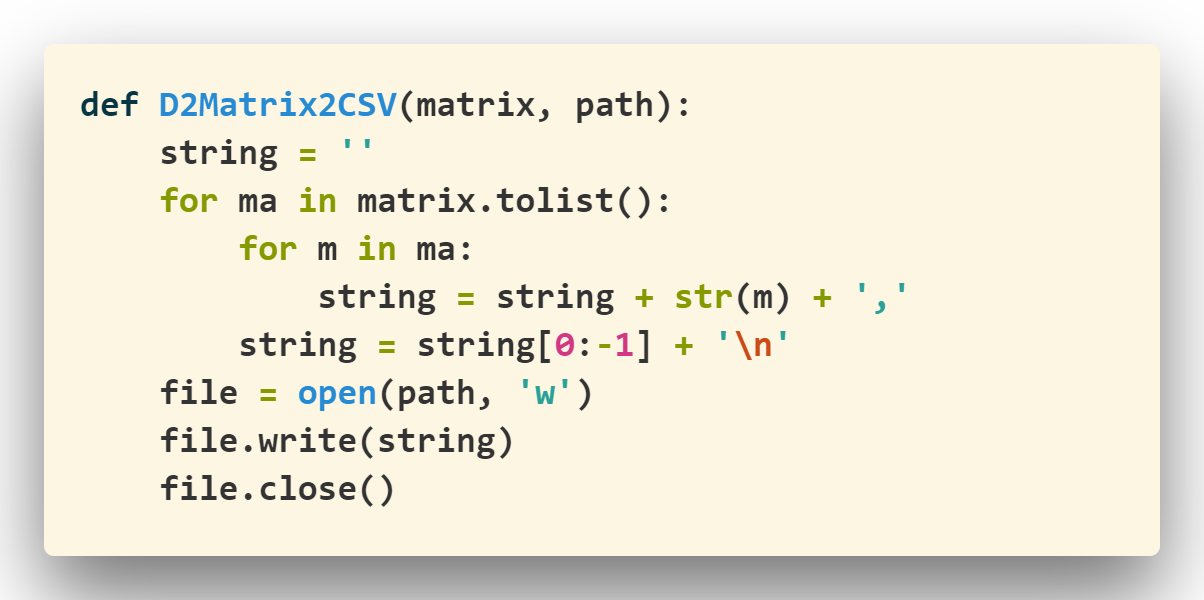


此处使用到的是numpy模块的genfromtxt方法读取CSV文件中的数据。

calculateContribRate函数代码如下：



D2Matrix2CSV函数代码如下：



这里主要是使用matrix矩阵转化为list变量，然后通过for循环遍历其中的元素，构造一个符合CSV格式的字符串，最终保存到指定路径的文件中即可。

PCA函数代码如下：



所有注释基本都写在了代码中，基本步骤完全遵循2.1.1节。

1. 收益率曲线拟合

基于2.2节的研究，选择Nelson-Siegel模型作为收益率曲线拟合模型。使用第三方数学分析模块scipy的optimize.fmin函数优化Nelson-Siegel参数，选择最合适的四个参数确定Nelson-Siegel模型，最后返回拟合的数据。

此程序除了依赖scipy模块之外，还依赖pandas、numpy、pymongo等第三方模块，这些模块都需要使用pip进行安装。另外，还需要引入内置模块json，以便于Node与Python进行数据交互。

程序代码如下：



1. 基于Monte Carlo模拟法计算VaR

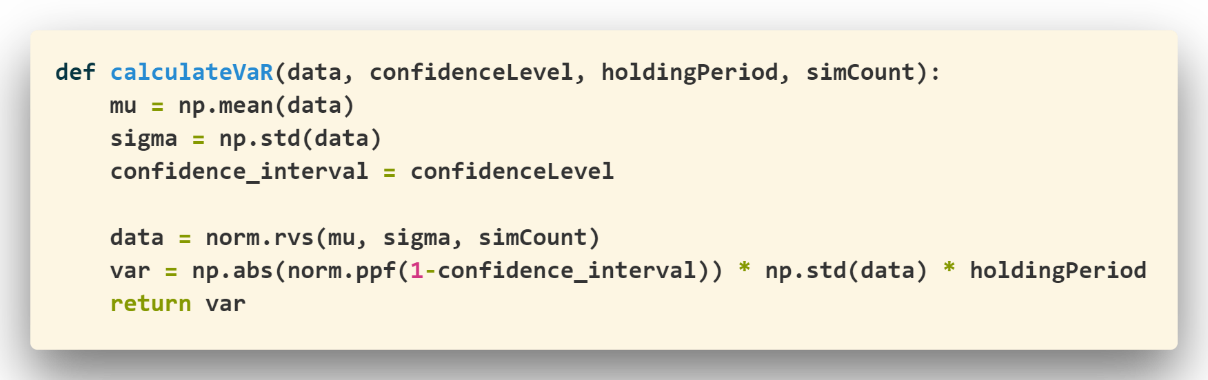
基于2.3节的结论，很容易得到VaR的计算方法，其大致步骤如下：

1. 接受一组一维参数，代表某支债券的历史到期收益率数据。
2. 计算这组一维历史数据的平均值和标准差。
3. 根据求出的平均值和标准差，使用随机函数生成一组符合要求的且服从正态分布的随机数。
4. 根据这组随机数，计算VaR。

当然在实际开发中还要考虑到其他参数，比如置信区间、债券的持有期 以及Monte Carlo模拟的次数。

这个程序的开发依赖一下模块：numpy、scipy.stats、xlrd、os、re，其中 numpy、scipy.stats和xlrd需要使用pip命令安装。

整个程序的核心是一个函数：calculateVaR，它负责计算VaR。其具体代 码如下：



1. 实证分析

在实现所有数据分析功能之后，仍然使用中国国债数据作为样本，验证程序分析的正确性。基于软件的数据分析如下：

1. PCA分析

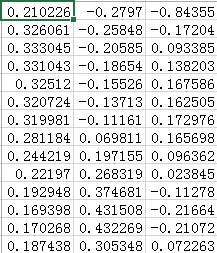
通过软件将中国国债数据上传之后，点击“PCA分析”，得到贡献度与因子载荷矩阵如下：



与图2.1所代表的SPSS软件的PCA分析结果图对比，相差不超过0.05%。



上图是软件分析出的中国国债数据的因子载荷矩阵，下图是SPSS软件分析出的因子载荷矩阵：



对比也可以发现结果非常的接近，可以证明程序的分析效果很好。

1. 曲线拟合

关于软件进行收益率曲线拟合的实证分析，请阅读本文2.2.4节——Nelson Siegel模型的实际拟合效果，本节不再赘述。

1. VaR计算

关于软件进行VaR计算的实证分析，请阅读本文2.3.4——实证分析，本节不在赘述。

4.2 Node程序设计

Node程序负责搭建服务器、查询数据以及运行Python脚本，充当的是一个代理的角色。