附件5B：

毕业设计（论文）开题报告

R语言corrplot包开发及在多元控制图中的应用

1、课题的目的及意义

1.1 课题的目的

1.1.1 掌握R packages的开发

通过学习掌握，如何设计，开发一个R packages，并参与到R packages 的后续维护，升级当中。提高学生的编程能力，团队意识，项目开发经验。

1.1.2. 结合相关矩阵和多元控制图实现对质量异常原因的推断

了解当前工业工程领域中控制图的应用，了解多元控制图原理。结合高斯混合模型（GMM）实现多元控制图的异常原因的推断

1.2 课题意义

1.2.1 开发R packages的意义

随着人工智能，大数据时代的到来，各个学科和数据科学都有不同程度的交叉，工业工程亦是如此。工业工程师在企业管理中会经常接触数据，有处理数据，分析数据的需求。R语言作为一门专门面向数据分析，挖掘的编程语言对工业工程师来说是一件解决问题的利器。

学习使用，制作和维护R packages，参与到R packages的生命周期当中可以提升编程能力，问题解决能力。更重要的是了解深入到数据科学这个领域当中去。

R作为一个开源的平台，又一大批爱好者组成的社区来支撑，越多的使用者，开发者加入对于R的生态有很好的提升。

1.2.2 结合相关分析研究多元控制图应用的意义

工业工程师经常使用控制图来进行质量控制等，使用控制图本身就是利用数据可视化的方式进行管理。但是当被检查对象的拥有多维变量时，无法使用单一变量进行衡量。为了使用更多维度的数据描述被检查对象的稳定性，同时可以结合相关性分析，利用一些机器学习模型来推导出生产出现问题的原因，这样可以解决很多现实的生产问题。在质量检测中，经常使用多元控制图对拥有多个检测属性的产品进行质量稳定性监控。但是一般多元控制图本身只能起到监控生产稳定性的作用并不能明确出现异常的原因。因此为控制图赋予新的作用：出现异常推断生产原因，是非常有意义的事情。

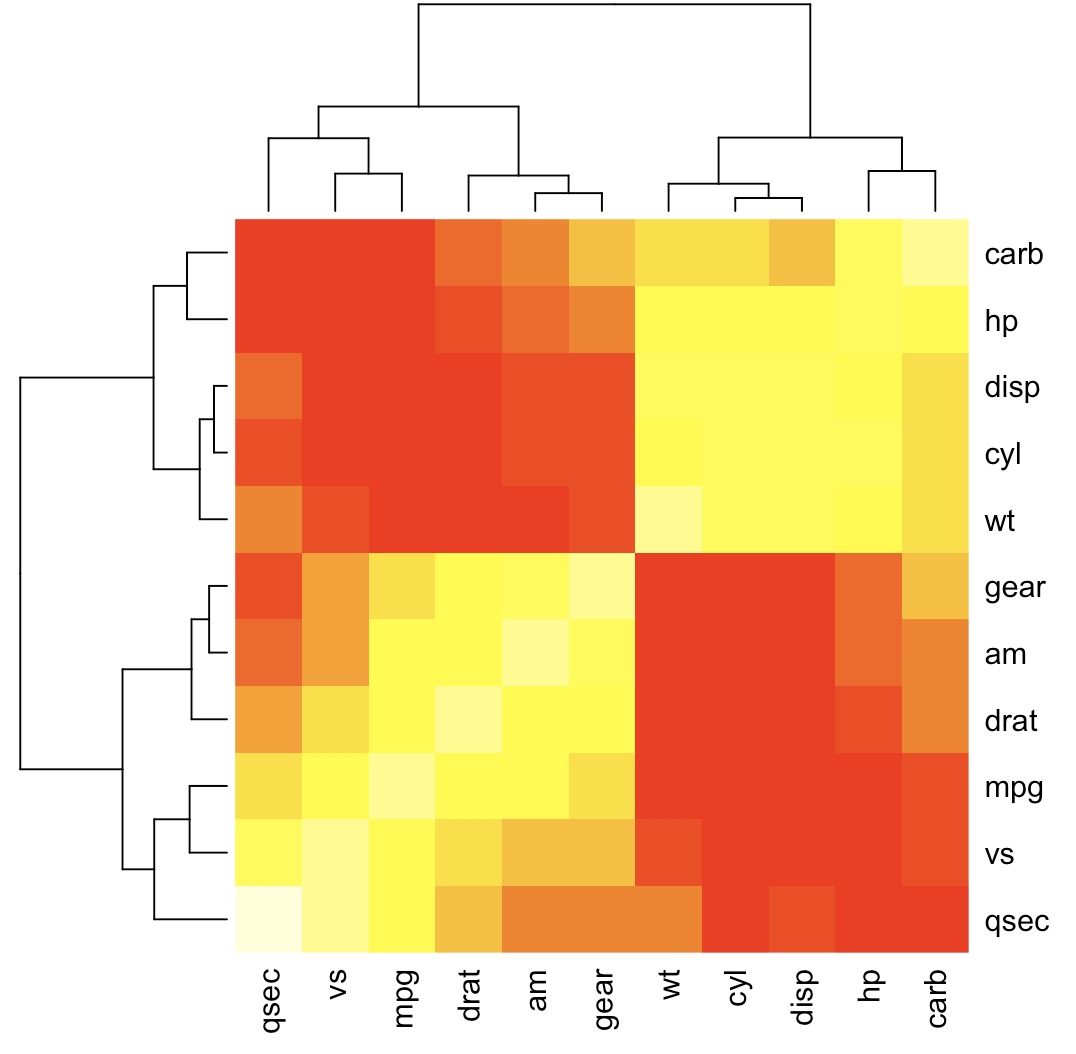
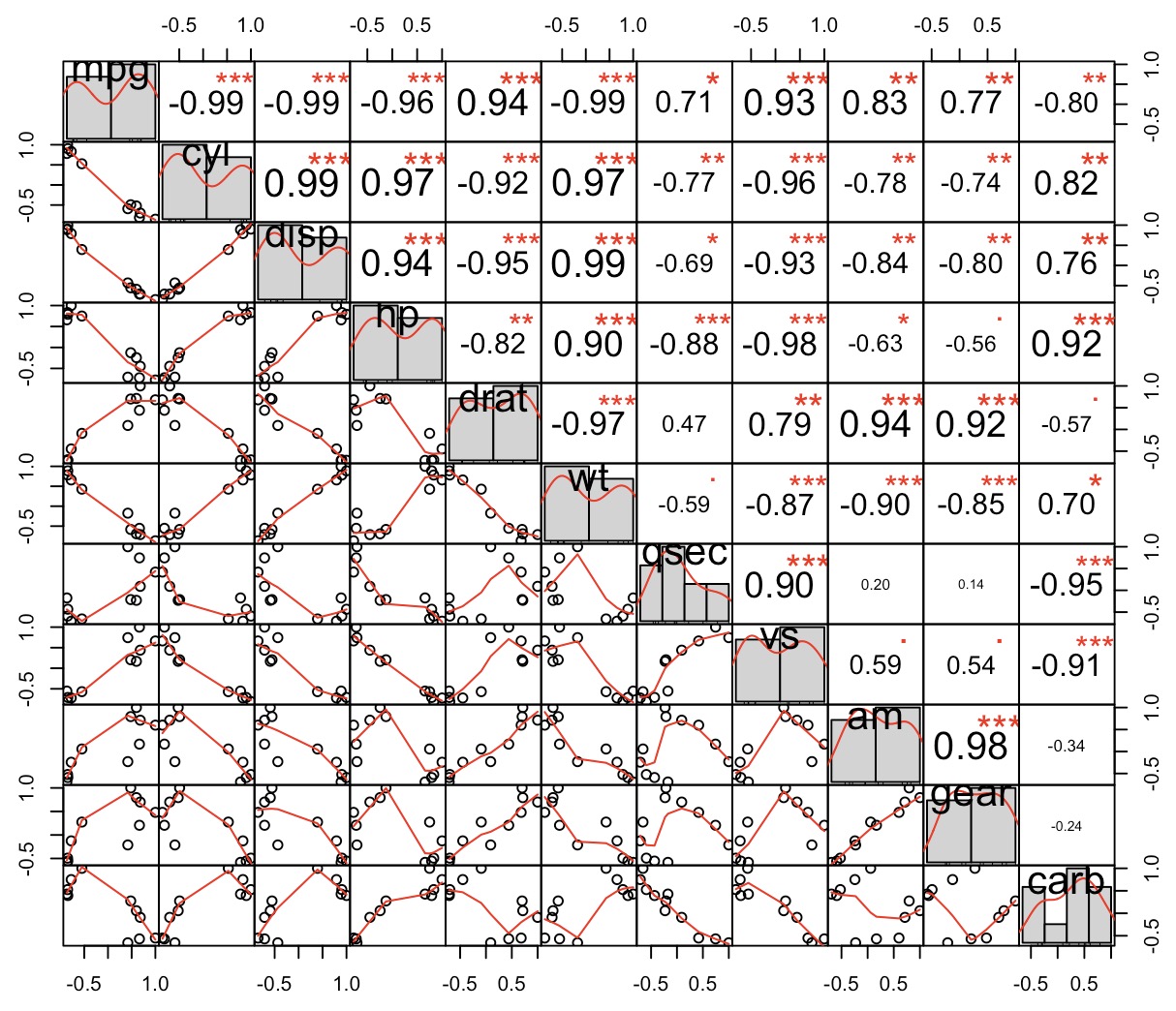
1.3 国内外研究现状

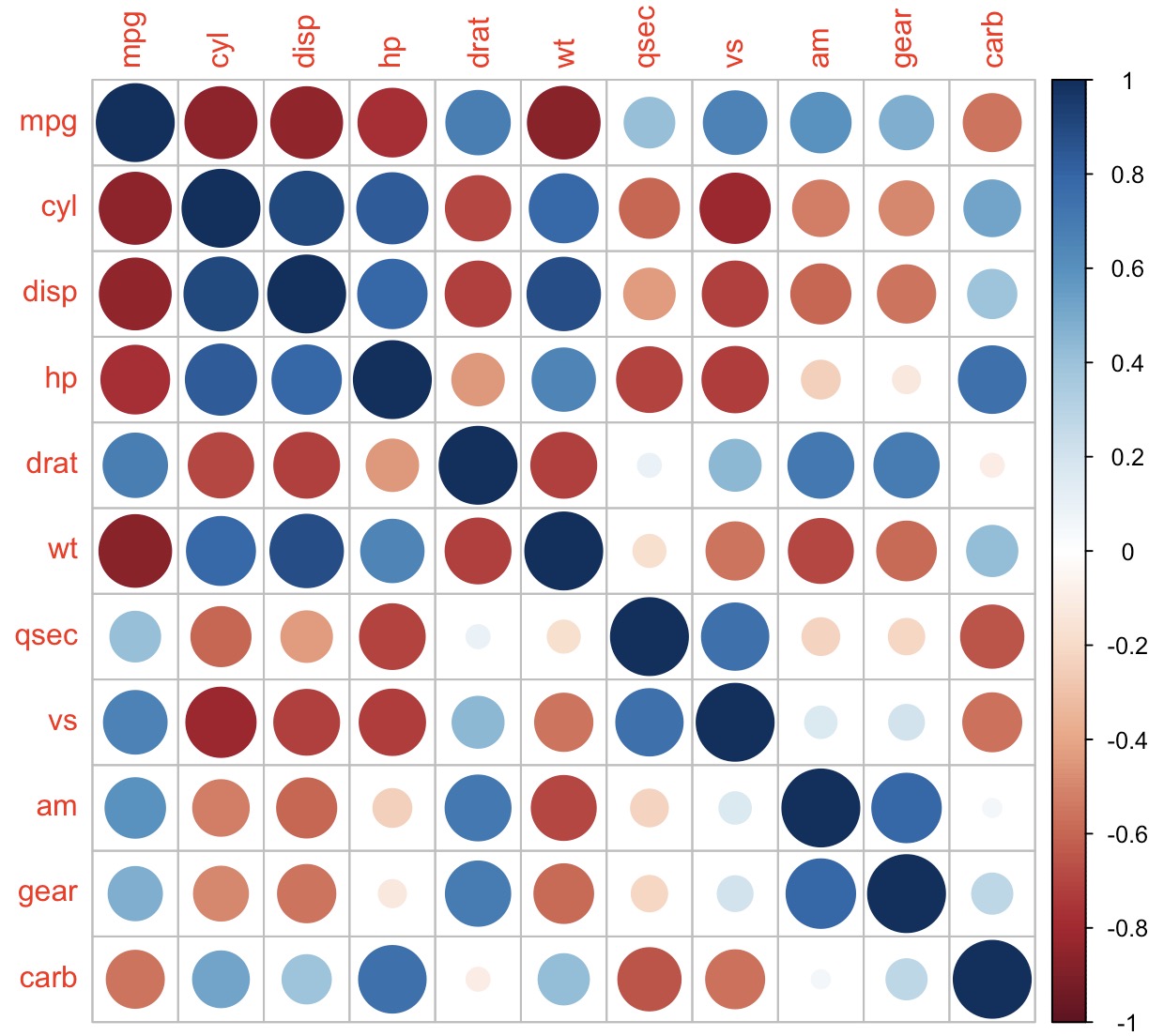
1.3.1相关矩阵绘制现状

相关矩阵的绘制是矩阵绘制问题的一种。目前矩阵的绘制主要有两种方式：1.网络法（network） 2.网格法（grid）（参考文献）。网络法主要适用于矩阵稀疏时，用节点表示变量，边表示变量之间的关系，该方法适用于社交网络分析。网格法适合稠密的矩阵，适合相关矩阵的可视化。

|  |  |  |
| --- | --- | --- |
|  | 网格式方法 | 网络式方法 |
| 优点 | * 适合于所有数字矩阵 * 不会出现节点重合、边交叉现象 * 可以展示节点间较为复杂的关系 | * 仅限于关系矩阵 * 直观性强、更紧密，更易读 * 矩阵稀疏时更有效 |
| 缺点 | * 矩阵维数较高时，占用空间较多 * 不如网络图直观易懂 * 路径追踪难度大 | * 矩阵密集时容易出现节点、边大量重合 * 需要较好布局算法防止节点、边重合 * 节点间关系复杂时效果较差 |
| 共同点 | * 都依赖具体统计、数学模型为基础来探求变量、个体之间的关系，且这些关系发掘算法有很多是可以互通的 | |

目前相关矩阵的可视化R语言包有corrplot包，PerformanceAnalytics包，heatmap函数等。这些包各有特点，但是corrplot是专注相关矩阵可视化的R语言包，同时在相关矩阵可视化中有着较高认可度。





corrplot的功能分为两个层面，统计模型和展示层面。统计模型主要是指是对矩阵变量的顺序进行重排，以便发现规律。展示层面是指利用图形，颜色等计算机图形来清晰地表达矩阵的复杂信息。

corrplot诞生于数据可视化渐渐升起热潮的时代，在数据可视化快速发展的今天，corrpot绘制的图形需要具有可交互性便于使用者更好，更方便地使用。所以开发corrplot包的新版本是一件非常有趣，有意义的事情。

1.3.2 多元控制图研究现状

控制图在质量控制中用来衡量产品生产的稳定性。显然，如果对一件产品检查越多的属性，那么对这件产品度量准确性就越高，将多个变量数据用来衡量产品生产的稳定性就需要多元控制图。

多元控制图一般有两个用途：1. 在假设检测变量服从正态分布的情况下，收集大量数据来估计其分布，确定其均值和方差；2. 根据估计正态分布来检测后续生产是否受控。

因此多元统计控制图的基本假设就是，检测的质量特性共同组成的变量服从均值，协方差为sigma的多维正态分布。当均值，和协方差同时保持稳定时，该生产过程被认为处于统计受控状态。

可以看到，多元控制图的基本要求就是均值和协方差要同时处于稳定状态。

有很多控制图基于被提出来：

多元均值控制图：多元均值控制图是多元统计过程控制中使用最为广 泛的统计控制图 包括总体协方差矩阵Z 己知时的多元才 控制图和总体协方差矩阵艺未知时的多元 控制图。

多元离差控制图：在讨论协方差未知的均值向量控制图时 总是假定 被控制的p 个质量特性的协方差矩阵保持统计稳定状态对于多特性过程的情况 只有协方差阵艺处于受控状 态 均值向量控制图才有意义 因为统计量护的计算式 中含有协方差阵 。

多元离差控制是一个比较复杂的问题 需要借助广 义方差 协方差矩阵的行列式 的概念 因为包含各分 量的方差及各分量间的协方差的协方差矩阵是一个矩阵 表达式 不便进行数量的比较。 而广义方差作为标量 不仅可以表征多元分布的离散情况 而且能够进行数量比较。

三向控制图 三向控制图是一种特定形式的多元统计控制图 由

单值控制图、 移动极差控制图和极差控制图张时序控制 图联合组成 用于监控同一 部件不同位置土以相同方式 测量的同一质量特性的统计控制状况 。 三向控制图可以 同时控制多个类似的质量特性 识别任一质量特性的失 控或其他非随机条件 以便采取措施纠正任何的非随机 现象。 使用三向控制图 既可以监控部件与部件之间的波动。

多元累积和控制图

多元指数加权移动平均控制图

基于主成分模型的多元统计控制图

主成分分析方法是将多个相关的统计变量转化为 少数几个相互独立的变量的一种有效的统计降维分析方 法 简称PC 方法。 基于主成分分析方法建立的、 能够 反映生产过程正常运行状态的统计降维模型称为主成分模型。

引用参考文献（质量管理，检测方面；多元控制图方面）

参考文献：

2、课题任务、重点研究内容、实现途径

2.1课题任务

本课题有两个任务 1）开发新版本的corrplot包，2）结合相关分析使用多元控制图，提高对生产过程的问题进行控制。

1）开发新版本的corrplot包

学习如何制作R packages，学习d3.js,html,css等前端知识，在github上和其他人一起合作完成。

2）结合控制图，优化质量检测成本

利用corrplot的新版本成果，结合相关分析，在保证控制图中数据一定稳定性时，对生产过程中的质量检测进行一定优化。

2.2重点研究内容

由于该课题的内容主要分为两大块，R packages的开发和控制图方面中的应用。由于时间和能力的限制，本科毕业设计不可能在这两个领域中同时作出非常成果，因此该课题将重点放在corrplot新版本的开发上。

传统的多元控制图只能对质量进行控制并不能对指出出现异常的原因。本文希望在实际质量控制案例背景中，结合控制图中使用，观察corrplot新函数带来的便利。使用GMM高斯混合模型来反相推导出产生异常的原因。

2.3实现途径

为实现需要的交互效果，corrplot新版本需要增加一些新的函数。新的函数需要建立在一个R packages htmlwidgets的基础之上，在htmlwidgets框架基础之上，通过d3.js对html的svg元素实现操作，实现矩阵可视化的绘图，布局，动态等。

（细节）

由于控制图不能推断异常原因，可以结合一些机器学习的方式解决该问题。GMM高斯混合模型是建立数据服从正太分布的模型，可以非常好地用于SPC。

1. 进度计划

|  |  |  |
| --- | --- | --- |
| 序号 | 起止周次 | 工 作 内 容 |
| 1 | 1周至6周 | 开发R语言corrplot包，学习前端开发的知识，如：js，css ，html等 |
| 2 | 6周至7周 | 学习研究控制图的使用，研究现状 |
| 3 | 7周至9周 | 学习研究GMM模型应用到控制图中 |
| 4 | 9周至14周 | 撰写修改论文 |
| 5 | 14周至15周 | 完成毕业设计 |

学生签名：

年 月 日

4、指导教师意见

指导教师签名：

年 月 日