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

一、课题的目的及意义

（一）课题的背景

数据可视化是一个非常受欢迎的话题。它将能够让人类接收大量数据的信息。R语言corrplot包就是这样一个为相关矩阵可视化而产生的软件包。它在R镜像下每月下载量达3.3万左右，被顶级期刊论文引用超过30余。corrplot从2008年由魏太云创建起，从0.1版发展到今天的0.84版。经过很多R爱好者一起开发完善。目前0.84版的corrplot相较其他类似的软件包在可视化和统计模型层面有很大的优势。

corrplot在很多领域如医学，生物学，数学还有工程领域得到广泛使用。对于相关矩阵的可视化，工业工程领域也经常使用这种方法。比如，质量屋的搭建、分析，企业管理中的数据分析，人岗匹配，质量管理等等相关领域。

随着大数据和人工智能时代的到来，各个学科的相互交叉愈发频繁。很多领域都有对数据分析和挖掘的需求。corrplot包本身建于R的环境之上，要使用好corrplot，制作美观合理的图需要一定的R语言编程。对于其他研究领域的人员来说这就增加了其使用corrplot包的成本。如果可以开发一个可交互性的新版本方便大家使用，节约使用时间，这是非常有意义的一件事情。同时如果使用html，js等语言进行开发可以让corrplot的新版本在可视化方面更进一步，这样可以很好地在可重复性报告，成果展示等方面得到很好发挥效果。

在质量的统计控制过程领域（SPC）控制图经常被作为对生产稳定性的监控。随着对生产的数据监控需求越来越多，人们开始收集不止一个维度的数据，于是多元控制图就出现了。多元控制图监控主要的基本研究是两个方面：均值是否服从正态分布，协方差是否服从正态分布。接下来的多数研究基本在这两个方面讨论。可是传统多元控制图只能进行稳定性检测，并不知道出现异常是否真的出现质量问题，也不会推导出哪个工艺出现问题。所以目前的大多数研究都在这个方面，即将稳定性和合格率相结合来研究控制图。多元控制图的研究和应用上会使用相关性或者协方差矩阵，如果可以结合corrplot的可视化效果为控制图的使用带来便利那么会是一件非常有意义的事情。

（二） 课题的目的

本课题的目的在于开发可交互版本的corrplot包并将其运用到多元控制图当中去。

通过学习掌握如何设计、开发一个R packages，并参与到corrplot的后续维护、升级当中。了解corrplot的应用场景，了解相关矩阵可视化的现状，比较其他相关矩阵可视化R语言包的优劣。了解控制图的应用现状。结合一个生产管理的场景使用控制图来解决现实问题。分析相关分析和多元控制图的结合带来的便利。最终制作一个将相关分析和多元控制图结合分析的可视化应用，使之流程化。

（三） 课题意义

1. 开发corrplot新版本的意义

corrplot是一个基于R的相关矩阵可视化的软件包，使用corrplot需要一定的编程基础。同时corrplot经过发展，功能很完善，合理的使用corrplot需要了解很多函数和参数。开发可交互版本的corrplot包可以方便非计算机领域的使用者使用，降低使用门槛，减少阅读使用文档的时间。

corrplot的新版本具有很好的展示效果，因此可以在数据可视化方面有很大的应用，可以支持shiny，rmarkdown等工具，支持开发可重复的报告和app。

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。

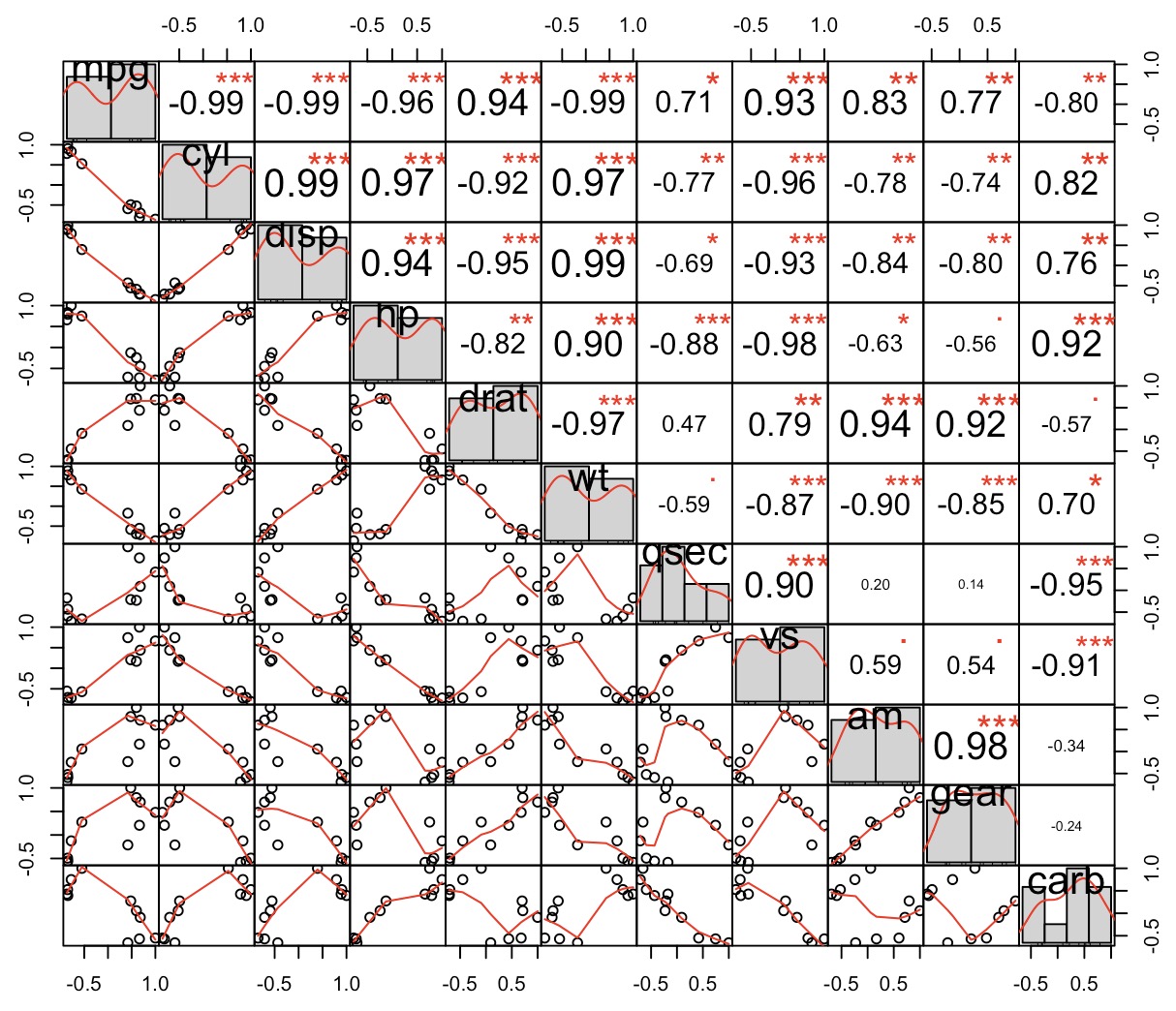
（四） 国内外研究现状

1.corrplot目前的应用发展现状

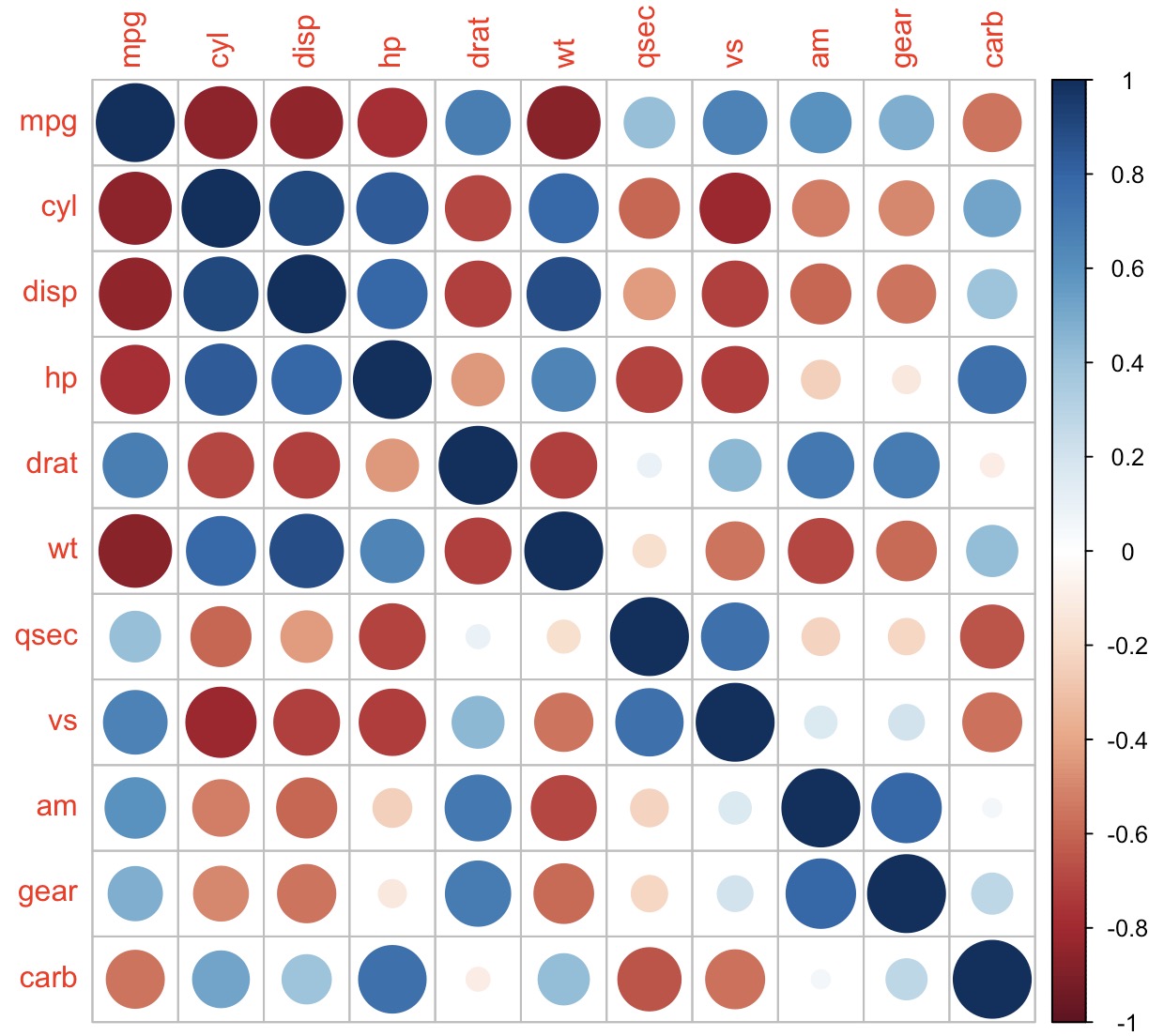
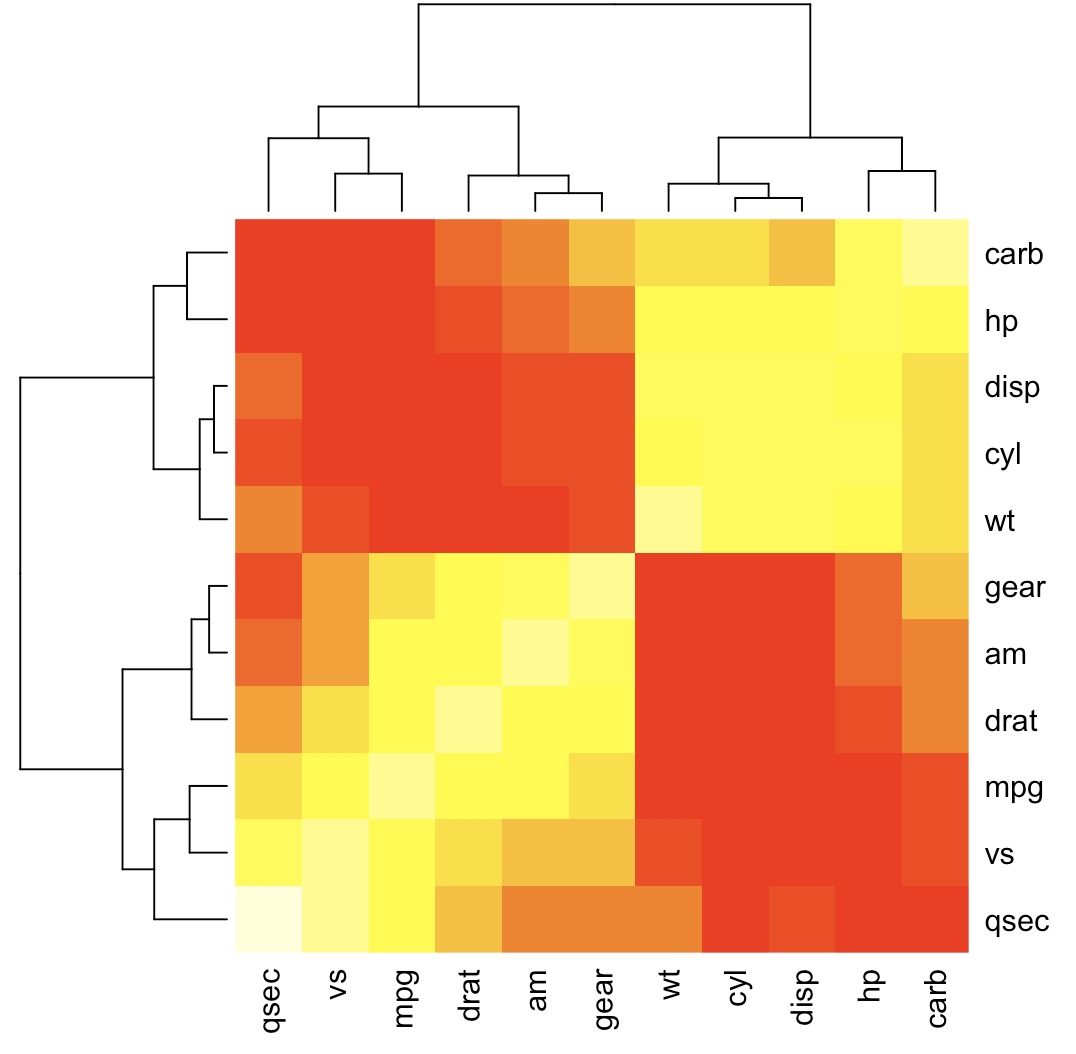
corrplot解决的问题就是相关矩阵的可视化，相关矩阵的绘制是矩阵绘制问题的一种。目前矩阵的绘制主要有两种方式：1.网络法（network） 2.网格法（grid）。网络法主要适用于矩阵稀疏时，用节点表示变量，边表示变量之间的关系，该方法适用于社交网络分析。网格法适合稠密的矩阵，适合相关矩阵的可视化。

|  |  |  |
| --- | --- | --- |
|  | 网格式方法 | 网络式方法 |
| 优点 | 1.适合于所有数字矩阵  2.不会出现节点重合、边交叉现象  3.可以展示节点间较为复杂的关系 | 1.仅限于关系矩阵  2.直观性强、更紧密，更易读  3.矩阵稀疏时更有效 |
| 缺点 | 1.矩阵维数较高时，占用空间较多  2.不如网络图直观易懂  3.路径追踪难度大 | 1.矩阵密集时容易出现节点、边大量重合  2.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语言软件有corrplot包，PerformanceAnalytics包，ggcorrplot，heatmap函数等。这些包各有特点，其中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和论文引用量。PerformanceAnalytics中有一个函数来实现相关性可视化。ggcorrplot是建造在ggplot2上面的包。它的独立性受到限制，绘制风格是ggplot2的风格。heatmap是一个R自带的函数，可以绘制简单的绘制相关矩阵。



图一 PerformanceAnalytics绘制的相关矩阵图



图二 corrplot绘制的相关矩阵图

图三 heatmap绘制的相关矩阵图

corrplot的功能分为两个层面（魏太云，2010），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 经常在生物学，医学方面研究经常得到使用。比如寻找miRNAs之间关系(Bot AM2013)，BotAM在研究mRNA时，绘制了miRNAs表达水平之间的相关性，挖掘了假阳性动物与受试动物表达差异。对67个选定细胞特征进行相关性可视化，使用皮尔森系数，最后进行聚类发现规律。Jeffery Li（Jeffery Li，2015）对67个选定细胞特征进行相关性可视化，使用皮尔森系数，最后进行聚类。corrplot在google学术上确定的引用量在214次，其中主要为医学生物方面的数据研究。而工业工程是企业的诊断在数据研究方面同样有很高的需求，corrplot会有很广的用途。

2．多元控制图研究应用现状

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为μ，协方差为Σ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。(梁文娟.2016)

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。

根据有关研究（蒋家东，2007），传统控制图主要有这几种：

（1）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Σ 己知时的多元控制图和总体协方差矩阵未知时的多元控制图。（2）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 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 。（3）多元离差控制图：使用广义方差协方差矩阵的行列式值来代替协方差，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。 而广义方差作为数值标量可以表征多元分布的离散情况，同时能够进行数量比较。（4）其他控制图包括，三向控制图，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等。

有时变量并不会一定以数值的形式呈现，而是以等级的方式呈现，兼顾到变量间的相关性，可以使用模糊方法多元多项式控制图(Fernández M N P,2015)。

在多元统计过程控制中，当过程发生变化时，并非所有变量而是少数变量都可能从控制状态转移。KazuyaNishimura提出了一种基于AIC的变量选择多元EWMA控制图。

近年来控制图的研究主要集中在模式识别相关的领域。由于控制图本身只能监控数据的稳定性并不能够识别出异常产品是否合格，更不能推断出现异常情况的原因。而为了能够跟上智能制造的要求，许多研究将控制图和机器学习模式识别相互结合起来。比如Yu, J在2012年和2011年的两篇文章（Yu, J，2011）（Yu, J，2012）中提出将高斯混合模型（GMM）与控制图相互结合，利用数据训练出多个高斯混合模型来达到数据压缩和对异常模式的识别。这两篇基本上理论是一致的，2011年的重点在理论模型，2012年重点在应用的改进。

二、课题任务、重点研究内容、实现途径

（一）课题任务

本课题有：

1）开发corrplot新版本，达到新的可交互版本可以达到满足非编程用户方便使用的效果。学习css，html，js等前端开发语言。

2）查阅文献资料，了解corrplot的应用场景。

3）学习研究控制图的研究应用现状，了解控制图原理，能够将控制图用于实际生产场景当中的去，研究可能的使用方式如结合模式识别或者数据压缩。

4）结合corrplot新版本的使用和当前的研究现状对控制图的应用进行改善。

（二）重点研究内容

1）研究如何开发corrplot可交互版本，达到美观、便捷的使用效果。为corrplot新版本增加一些新的函数。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。

2）研究当前相关矩阵可视化和工业工程相关的应用领域，如在SPC等数据分析领域。

3）利用相关性、控制图对某企业的质量检测进行优化。

（三）实现途径

本课题的实现途径使用图的方式表达如下图

企业质量控制改进

企业

场景

数据压缩

模式识别

筛选当前具有价值和意义的控制图研究

开发corrplot

多元控制图文献综述

学习css html，js等前端工具

基于R shiny的质量控制的可视化app

图四 本课题的实现途径

参考文献：

[1]黄云云, 杨晓翔, & 黄剑平. (2010). MCUSUM、MEWMA及Hotelling T2控制图的比较分析. 现代制造技术与装备, (03), 64–65+68.

[2]李钢, & 代海飞. (2008). 小批量多元控制图建模方法研究. 计算机应用, (10), 2718–2720.

[3]梁文娟. (2016). 关于多元控制图的若干问题研究 (博士). 华东师范大学.

[4]魏太云 .(2010).矩阵可视化的比较研究及其应用.中南大学

[5]张灿华. (2012). 基于主成分分析法的结构损伤识别研究 (硕士). 重庆交通大学.

[6] 周茂袁, & 耿薇. (2013).一个非参数多元控制图(英文). 南开大学学报(自然科学版), (03), 69–74.

[7]赵永满, 何桢, 何曙光,等. (2013)基于PSO的支持向量机多元控制图均值偏移诊断模型[J]. 天津大学学报(自然科学与工程技术版)(5):469-475.

[8]谭超. (2016)基于参数估计的多元控制图性能研究[D]. 湖南大学

[9]张兵.(2016) 多元控制图在发动机监控中的应用研究[D]. 中国民航大学

[10]陆永婷.(2015) 二维平面多元统计控制图[D]. 上海交通大学

[11]黄虎, 柯华, 王晶.(2016) 基于逆加权参数估计方法的改进型Q控制图研究[J]. 系统工程学报, 31(4):568-574.

[12]谭超, 刘坚, 张星.(,2016) 基于参数估计的贝叶斯均值控制图研究[J]. 统计与决策 (21):22-25.

[13]陈瑞兴, 程琳.(2012) 基于PCA和多元控制图的大坝安全多测点数据监控方法[J]. 水利与建筑工程学报10(5):68-72.

[14]蒋家东,蒋小琴,冯允成,商广娟.多元统计控制图的比较研究[J].航空标准化与质量,2007(01):17-21.

[15]Yu, J. (2011). Pattern recognition of manufacturing process signals using Gaussian mixture models-based recognition systems. Computers & Industrial Engineering, 61(3), 881–890.

[16]Yu, J. (2012). Gaussian mixture models-based control chart pattern recognition. International Journal of Production Research, 50(23), 6746–6762.

[17]Rea A, Rea W.(2014) Visualization of a stock market correlation matrix[J]. Physica A Statistical Mechanics & Its Applications, 400(2):109-123.

[18]Wei T, Simko V. (2013)corrplot: Visualization of a Correlation Matrix[J]. Mmwr Morbidity & Mortality Weekly Report, 52(12):145-51.

[19]Mcguire M K, Meehan C L, Mcguire M A, et al. (2017)What’s normal? Oligosaccharide concentrations and profiles in milk produced by healthy women vary geographically[J]. American Journal of Clinical Nutrition, 105(5):1086-1100.

[20]Tuerhong G, Kim S B.(2014) Gower distance-based multivariate control charts for a mixture of continuous and categorical variables[M]. Pergamon Press, Inc.

[21]Mingwei L U, Rudy R J. MULTIVARIATE CONTROL CHART[M]// Recent Advances In Reliability And Quality Engineering. 1993:61-74.

[22] Fernández M N P, García A C, Barzola O R(,2015). Multivariate multinomial control chart using fuzzy approach[J]. International Journal of Production Research, 53(7):2225-2238.

[23] Nishimura K, Matsuura S, Suzuki H(2015). Multivariate EWMA control chart based on a variable selection using AIC for multivariate statistical process monitoring[J]. Statistics & Probability Letters, 104:7-13.

三、进度计划

|  |  |  |
| --- | --- | --- |
| 序号 | 起止周次 | 工 作 内 容 |
| 1 | 1周至5周 | 开发R语言corrplot包，学习前端开发的知识，如：js，css ，html等 |
| 2 | 5周至7周 | 学习研究控制图的使用，研究现状 |
| 3 | 7周至9周 | 选择当前控制图的方向结合企业实际应用 |
| 4 | 9周至11周 | 利用企业数据进行验证，修改模型 |
| 5 | 11周至14周 | 撰写修改论文 |
| 6 | 14周至15周 | 完成毕业设计 |

学生签名：

年 月

四、指导教师意见

指导教师签名：

校外指导教师签名：

年 月 日

说明：

1．开题报告应根据教师下发的毕业设计（论文）任务书，在教师的指导下由学生独立撰写。

2．本页不够，请加页。