

控制与决策

Control and Decision

融合主题模型与文本特征的汽车质量多维动态监测与诊断

宋贲, 王晗, 张久军, 胡雪龙

引用本文:

宋贲, 王晗, 张久军, 等. 融合主题模型与文本特征的汽车质量多维动态监测与诊断[J]. *控制与决策*, 2025, 40(9): 2879–2890.

在线阅读 View online: <https://doi.org/10.13195/j.kzyjc.2024.1469>

您可能感兴趣的其他文章

Articles you may be interested in

[车辆跟随控制策略的状态可达集建模及验证方法](#)

A modeling and verification method of state reachable set for vehicle following control strategy
控制与决策. 2021, 36(7): 1679–1685 <https://doi.org/10.13195/j.kzyjc.2019.1562>

[需求波动下考虑乘运供应能力的网约车平台动态定价](#)

Dynamic pricing for ride-hailing platform with demand fluctuation and supply capacity
控制与决策. 2021, 36(6): 1499–1508 <https://doi.org/10.13195/j.kzyjc.2019.0881>

[基于模糊-两阶段超效率SBM的电网应急能力动态综合评价](#)

Dynamic comprehensive evaluation of power grid emergency capability based on fuzzy-two-stage super efficiency SBM
控制与决策. 2021, 36(6): 1333–1341 <https://doi.org/10.13195/j.kzyjc.2019.1128>

[含混合储能的互联电力系统传感器容错负荷频率控制](#)

Sensor fault-tolerant load frequency control for multi-area interconnected power system with hybrid energy storage system
控制与决策. 2021, 36(5): 1069–1077 <https://doi.org/10.13195/j.kzyjc.2019.1432>

[基于FWADE-ELM的短时交通流预测方法](#)

Short-term traffic flow forecasting based on hybrid FWADE-ELM
控制与决策. 2021, 36(4): 925–932 <https://doi.org/10.13195/j.kzyjc.2019.1103>

融合主题模型与文本特征的汽车质量多维动态监测与诊断

宋 贲¹, 王 晗^{1†}, 张久军², 胡雪龙³

(1. 沈阳农业大学 理学院, 沈阳 110866; 2. 辽宁大学 数学与统计学院, 沈阳 110036;
3. 南京邮电大学 管理学院, 南京 210003)

摘要: 汽车产业是国民经济的重要支柱产业之一, 其质量安全问题备受瞩目. 汽车质量反馈信息延迟性较高, 汽车用户在线投诉成为汽车制造商发现质量缺陷的重要信息途径. 现有研究主要关注汽车质量缺陷的静态提取, 如缺陷识别与类型划分, 较少研究汽车质量问题的动态变化. 鉴于此, 针对网络上汽车质量投诉文本特点, 应用潜在狄利克雷分配模型提取质量缺陷主题, 将投诉文本按照主题分类, 计算各缺陷主题下投诉文本的时间间隔和情感得分, 从频率和强度两个维度同时刻画投诉的变化趋势. 由于时间间隔和情感得分的联合分布难以确定, 基于秩检验和经验 Copula 提出一种二元稳健控制图 (记为 BEWMA-LC), 对于不同投诉主题同时在线监测投诉文本的时间间隔、情感得分以及相关性, 并在报警后能够进一步识别异常原因. 以大众速腾汽车为实证对象, 基于车质网 2010 年 ~ 2021 年在线投诉文本数据, 通过与已有非参数控制图的对比分析, 验证所提出方法在实时监测和诊断汽车质量变化方面的有效性.

关键词: 汽车质量监控; 网络投诉; 潜在狄利克雷分配; 时间间隔; 情感分析; 统计质量控制

中图分类号: O213.1; F272 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.1469

引用格式: 宋贲, 王晗, 张久军, 等. 融合主题模型与文本特征的汽车质量多维动态监测与诊断 [J]. 控制与决策, 2025, 40(9): 2879-2890.

Multi-dimensional dynamic monitoring and diagnosis of automotive quality integrating topic models and textual features

SONG Zhi¹, WANG Han^{1†}, ZHANG Jiu-jun², HU Xue-long³

(1. College of Science, Shenyang Agricultural University, Shenyang 110866, China; 2. School of Mathematics and Statistics, Liaoning University, Shenyang 110036, China; 3. School of Management, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The automotive industry is one of the pivotal sectors of the national economy, and its quality and safety issues have attracted significant attention. Due to the high latency of automotive quality feedback, online complaints from users have become a crucial source of information for manufacturers to identify quality defects. Existing research primarily focuses on the static extraction of automotive quality defects, such as defect identification and classification, with limited attention to the dynamic evolution of quality issues. Therefore, based on the characteristics of online automotive quality complaint texts, this paper applies the latent Dirichlet allocation model to extract quality defect topics, categorizes complaint texts by topics, and calculates the time intervals and sentiment scores of complaint texts under each defect topic, simultaneously depicting the trend of complaints in terms of both frequency and intensity. Since the joint distribution of time intervals and sentiment scores is difficult to determine, we propose a robust bivariate control chart (denoted as BEWMA-LC) based on rank tests and empirical Copula. This method enables simultaneous online monitoring of time intervals, sentiment scores, and their correlations for different complaint topics, while further identifying the root causes of anomalies after triggering an alarm. Using the Volkswagen Sagitar as a case study, based on the online complaint texts from 2010 to 2021 sourced from the "Chezhi" website, we demonstrate the effectiveness of the proposed method in real-time monitoring and diagnosis of dynamic automotive quality changes by comparative analysis with the existing nonparametric control charts.

Keywords: automotive quality monitoring; online complaints; latent Dirichlet allocation; time interval; sentiment analysis; statistical quality control

收稿日期: 2024-12-21; 录用日期: 2025-05-21.

基金项目: 国家自然科学基金青年项目 (12201429); 2022 年度教育部人文社会科学研究青年基金项目 (22YJC910009); 辽宁省社会科学规划基金一般项目 (L24BTJ002).

责任编辑: 阳春华.

†通信作者. E-mail: hanwang@syau.edu.cn.

0 引言

随着人们生活水平的提高,汽车销量逐年上升,随之而来的汽车质量问题也日益凸显.2021年中国消费者协会报告显示,对于汽车以及零部件的投诉量为41624件,同比增长19.28%,在商品大类投诉量中排名第3位^[1].汽车质量是汽车企业发展的根本,是核心竞争力,因此,对于汽车质量的监控是非常重要的.

消费者越来越愿意在网络上发表对于商品的想法和评价,因此,每天在各种网站和软件上均会产生数以万计的在线评论,其中也包括大量关于汽车的在线评论,这些评论为汽车购买者和汽车企业提供了很多有价值的信息.在评论中,“差评”通常蕴含着大量可挖掘的信息^[2],因此,对于汽车企业而言,汽车质量在线投诉文本具有更加重要的研究意义,它能够及时反映某个汽车车型普遍存在的质量问题.与传统获取售后反馈的方式相比,其具有速度快、成本低、数量多等优点.汽车质量在线投诉是非结构化的文本数据,自然语言处理技术的发展使得对于文本数据的挖掘和利用得到了进一步提高^[3],也为汽车质量问题的监控提供了一个崭新的视角.

目前,基于评论文本的汽车质量监控研究尚处于发展阶段,现有大多数研究方法主要包括:1)应用文本分析和机器学习对汽车质量问题的静态提取方法,如汽车缺陷识别与分类,Abrahams等^[4]基于网络媒体,提出了一种使用文本挖掘的新型车辆缺陷发现系统,以实现汽车缺陷的识别;Liu等^[5]构建了一个质量相关评论识别模型,并使用潜在狄利克雷分配(latent Dirichlet allocation, LDA)主题生成了模型、词云图分析和呈现汽车缺陷的主题分布;Abrahams等^[6]基于论坛用户生成内容,总结了一个用于产品缺陷发现的集成文本分析框架,并使用主成分分析和逻辑回归识别了汽车等产品的缺陷.但是,上述研究没有考虑质量缺陷的时序演化特征.2)在此基础上,一些学者进一步研究了提取的质量问题随时间的变化趋势,Kakde等^[7]利用SAS中的文本挖掘功能识别出汽车缺陷主题后,使用泊松统计量刻画了汽车缺陷主题的发生频数随时间的变化趋势;王余行等^[8]基于网络论坛数据提出了一种汽车质量问题的挖掘方法,探究了汽车质量问题的分布状况及其变化趋势.然而,上述研究方法只考虑了投诉频数单一维度指标,未结合投诉强度特征,并只依赖时序曲线的直观分析,虽然能够揭示质量问题的演变趋势,但是缺乏对时序波动中潜在质量风险的异常检测和诊断方法.

对于产品质量的实时在线监控,统计过程控制(statistical process control, SPC)是一种重要的研究方法,其中控制图主要用于监测序贯过程.早期的控制图主要是为了监测工业生产线的生产状态以确保产品质量^[9].近年来,控制图被应用于许多其他领域^[10-11],包括出租车服务质量管理^[12-13]、呼叫中心服务质量管理^[14]、数据质量的度量 and 监控^[15]以及售后评论文本在线监测^[16-18].Zhang等^[17]计算了连续负面评论的时间间隔和情感得分,基于非参数指数加权移动平均(exponential weighted moving average, EWMA)控制图监控连锁酒店在线评论;在此基础上,Zhang等^[18]提出了单侧非参数EWMA控制图检测笔记本电脑在线评论,提高了控制图的监测效率.但是上述研究工作没有提取评论主题,不利于异常报警后的问题诊断.

本文的关注点不仅在于汽车质量网上投诉文本的在线监控,更侧重于报警后的异常源诊断.为此,本文考虑汽车质量投诉的两部分信息:1)投诉的汽车部件,如发动机、变速箱等;2)具体的质量问题,如异响、渗油、有味等.由于在线投诉文本信息的混杂性,应用LDA主题模型、领域知识提取以及关联投诉文本中的汽车部件和质量问题,构成汽车质量缺陷主题.进一步地,在每个缺陷主题下,计算投诉文本的时间间隔和负面情感分数.其中:时间间隔的减少,意味着投诉人数增多,投诉频率增加,汽车出现该种缺陷主题对应的质量问题的可能性增大;而时间间隔的增加意味着该种汽车质量问题很可能得到了改善.负面情感分数越高,用户的使用感受越差,汽车质量问题可能越严重,反之亦然.因此,本文从时间间隔和情感分析两个维度,对每种质量缺陷问题的投诉分别进行动态分析、联合监测和异常诊断.

由于时间间隔和情感得分两个质量特征的联合分布难以确定,本文基于秩检验和经验Copula设计二元稳健控制图联合监控时间间隔和情感得分以及它们的相关结构,控制图的检验统计量由3个分量统计量组合构成:两个分量统计量分别用于监测时间间隔和情感得分的边缘分布,第3个分量统计量应用经验Copula监测时间间隔和情感得分的相关结构.所提出控制图在发出报警的同时能够进一步识别异常来源.当只有时间间隔发生向上偏移时,表明投诉频次显著增加,该缺陷主题可能是共性问题,需要引起企业重视,及时查找原因;当只有情感分数发生向上偏移时,表示部分客户对于该质量问题的反馈强烈,企业可有针对性地开展客户调查并采取相应的改进措施.综上,本文主要研究内容如

下: 1) 针对现有汽车质量在线评论研究方法的不足, 提出一个融合时间间隔和情感分析的多维动态监测与诊断框架. 与以往只关注投诉频数的研究不同, 本文进一步引入投诉强度, 即情感得分, 从两个维度刻画投诉的变化趋势. 由于时间间隔和情感得分的联合分布难以确定, 建立非参数控制图进行实时联合监控, 方法更加稳健, 能够更全面准确地反映汽车质量问题的严重程度和变化趋势. 2) 针对现有控制图方法未考虑变量间相关结构, 提出一种基于经验 Copula 和主成分得分的改进方法, 记为 BEWMA-LC 控制图. 通过计算经验 Copula 伪观测值的主成分得分, 结合 Lepage 非参数统计量和 p 值方法能够有效监测变量间的相关结构变化, 提高控制图的稳健性和监测效率. 应用实例分析比较 BEWMA-LC 与其他非参数控制图的监测性能, 结果表明 BEWMA-LC 图在多数情况下性能优于其他控制图, 并在报警后能够进一步识别异常原因. 尤其是当检验样本容量较小时, 该方法能够更敏感地检测到位置参数与变量间相关结构的小偏移, 能够为汽车质量问题的早期预警和诊断提供更可靠的数据支持.

1 汽车质量投诉文本分析框架

1.1 LDA 模型

由于汽车结构的复杂性, 汽车质量问题的缺陷主题种类多样. 因此, 汽车质量的监控需要对每种质量缺陷主题的投诉文本分别进行分析. 提取汽车质量缺陷主题一般采用监督学习和无监督学习两种方式. 其中: 监督学习往往需要大量的人工标注, 要求较高的人力成本和时间成本; 而无监督学习不需要人工标注, 其可以自动找出潜在的类别规则. 由于汽车质量投诉文本没有标签, 本文采用无监督 LDA 模型, 结合领域知识提取汽车质量缺陷主题.

LDA 模型是由 Blei 等^[19] 提出的, 在文本挖掘中, 该模型常用于文本主题的提取. LDA 模型认为文档是由若干个隐含主题构成, 而每个主题是一个关于单词的概率分布^[20]. 因此, 主题可由词汇分布来表示, 而文档可由主题分布来表示. 具体地, LDA 模型是以主题作为中间层, 若要生成一篇文档, 则其里面的每个词语出现的概率为

$$P(\text{单词}|\text{文档}) = \sum_{\text{主题}} P(\text{单词}|\text{主题}) \times P(\text{主题}|\text{文档}).$$

这里: $P(\text{单词}|\text{文档})$ 是已知的, 可直接通过给定的文档计算出来; $P(\text{单词}|\text{主题})$ 和 $P(\text{主题}|\text{文档})$ 是需要拟合的. 拟合过程即 LDA 的学习过程, 可简要表述如下: 首先, 从狄利克雷先验分布中生成初始值; 然

后, 在迭代过程中利用抽样或变分方法不断调整这两个分布, 直至拟合收敛至符合 $P(\text{单词}|\text{文档})$ 的实际分布; 最后, 利用最终求出的 $P(\text{单词}|\text{主题})$ 分布, 即每个主题下单词出现的概率, 选择每个主题下概率较大的单词, 结合领域知识最后确定出具体的主题含义.

LDA 模型输入的是文档和一个超参数, 即主题个数. 评价 LDA 主题模型的指标主要有两个: 困惑度和一致性^[21]. 一般而言, 困惑度越低或一致性越高, 模型拟合效果越好. 当主题个数越多时, 模型的困惑度通常越低. 但是, 当主题个数过多时, 生成的模型往往会过拟合, 因此, 本文采用一致性来评价 LDA 模型并选择最优主题个数. 一致性是量化文本连贯性的重要指标, 可以衡量 LDA 主题模型的可解释性. 本文利用 Röder 等^[22] 给出的表现最好的一致性 C_v 计算方法, 计算 LDA 输出的每个主题一致性的算数平均数, 作为 LDA 主题模型最终的一致性分数.

在线投诉文本中, 由于语言表达形式的多样性和复杂性, 在最优 (即主题一致性最高) 主题词汇分布的基础上, 还需要根据实际具体投诉文本内容, 结合领域知识进行质量问题分析和提炼, 总结出相应的部件和症状, 最后提取出汽车质量缺陷主题. 综上所述, 本文利用 Python Gensim 包中的 LDA 模型, 结合领域知识提取汽车质量缺陷主题, 具体实现算法如下.

step 1: 对投诉文本进行分词和过滤停用词处理, 并使用词袋模型表示投诉文本;

step 2: 利用 Python Gensim 包构建 LDA 模型, 并计算不同主题个数下的一致性;

step 3: 根据计算出的一致性选择最佳主题数, 并输出该主题个数下每个主题的词汇分布;

step 4: 结合主题词汇分布和投诉文本内容, 利用领域知识提取出汽车质量缺陷主题.

1.2 时间间隔和情感分析

汽车质量投诉的时间间隔和情感分数能够有效地反映汽车质量问题的严重程度和变化趋势. 根据实际投诉情况, 使用连续两次投诉的发布时间差来计算时间间隔. 若一天内的投诉有多条, 则它们的时间间隔为 $1/\text{一天内投诉的总数}$. 如: 一天内共有 4 条投诉, 它们的时间间隔均为 $1/4$, 即 0.25 天. 情感分析, 又称倾向性分析、意见挖掘, 是对文本数据所表现出的态度或情绪倾向性进行判断分析的过程, 具

根据 Sklar 定理^[26], 二元累计分布函数可表示为

$$F(v_1, v_2) = C\{F_1(v_1), F_2(v_2)\}.$$

其中: $C\{\cdot\}$ 为 Copula 函数, 是 $F_1(v_1), F_2(v_2)$ 的联合累计分布函数, 刻画了二元随机向量的相关结构. Sklar 定理阐明了多元联合分布可以表达为单变量的边缘分布函数和 Copula 函数. 因此, 可以方便地通过监测每个边缘分布和 Copula 函数来有效地监测多变量过程. Song 等^[27] 基于 Sklar 定理提出了一种二元稳健控制图, 用于监测位置向量和协方差矩阵并在报警后能够识别异常源. Song 等^[27] 构建的是 Shewhart 控制图, 对于微小而持续的偏移不敏感. 此外, 这种方法还有两个局限性: 1) 为了监控变量间的相关结构, 分别计算参考样本和检验样本的 Copula 伪观测值到原点的欧氏距离, 作为参考样本和检验样本的 Copula 距离样本. 基于 Copula 距离样本计算 Lepage 统计量和相应的 p 值. 但是由于变量间的相关性, Copula 伪观测值是相关的, 导致相应的 p 值不服从区间 $(0, 1)$ 上的均匀分布, 在一定程度上影响了控制图的稳健性. 2) 当检验样本容量很小时 (SPC 应用中的典型情况), 该方法不够灵敏, 对于检测变量间相关结构的小偏移效率不高. 为此, 本文在 Song 等^[27] 的基础上, 应用对于监测小偏移更加敏感的 EWMA 控制图. 此外, 尽管本文监测两个边缘分布的方法与 Song 等^[27] 相同, 但是, 改进了变量间相关结构的监测方法, 解决了上述的两个局限性, 提高了控制图的稳健性和监测效率.

假设 $\Psi_{0m} = \{(X_{01}, Y_{01}), (X_{02}, Y_{02}), \dots, (X_{0m}, Y_{0m})\}$ 为来自受控 (in control, IC) 二元连续过程的参考样本, 样本容量为 m , 具有未知的联合分布函数 $F_0(x, y)$. 值得注意的是, 参考样本需要通过阶段 I 分析进行建立^[28-29]. 假设 $\Psi_{jn} = \{(X_{j1}, Y_{j1}), (X_{j2}, Y_{j2}), \dots, (X_{jn}, Y_{jn})\} (j = 1, 2, \dots)$ 为来自阶段 II 的第 j 个检验样本, 样本容量为 n , 具有未知的联合分布函数 $F_1(x, y)$. 检验样本与参考样本相互独立. 当过程受控时, 应有 $F_0(x, y) = F_1(x, y)$. 因此, 二元过程的监控问题可转换为如下假设检验:

$$\begin{aligned} H_0 &: F_0(x, y) = F_1(x, y), \\ H_1 &: F_0(x, y) \neq F_1(x, y). \end{aligned}$$

等价于同时检验以下 3 个假设问题:

$$\begin{aligned} H_{01} &: F_{01}(x) = F_{11}(x), \quad H_{11} : F_{01}(x) \neq F_{11}(x); \\ H_{02} &: F_{02}(y) = F_{12}(y), \quad H_{12} : F_{02}(y) \neq F_{12}(y); \\ H_{03} &: C_0(x, y) = C_1(x, y), \quad H_{13} : C_0(x, y) \neq C_1(x, y). \end{aligned}$$

其中: $F_{01}(x) = F_0(x, +\infty), F_{02}(y) = F_0(+\infty, y)$ 为 $F_0(x, y)$ 的两个边缘分布函数. 同样地, $F_{11}(x) = F_1(x, +\infty), F_{12}(y) = F_1(+\infty, y)$ 为 $F_1(x, y)$ 的两个

边缘分布函数. 根据 Sklar 定理, 有

$$\begin{aligned} F_0(x, y) &= C_0\{F_{01}(x), F_{02}(y)\}, \\ F_1(x, y) &= C_1\{F_{11}(x), F_{12}(y)\}. \end{aligned}$$

3 个零假设全部成立表示二元过程处于受控状态, 当过程失控 (out of control, OOC) 时, 3 个零假设中至少有一个不成立.

2.1 边缘分布的检验统计量

对于两个边缘分布, 主要监测位置参数和尺度参数的偏移, 考虑位置-尺度模型为

$$\begin{aligned} F_{1i}(x) &= F_{0i}\left(\frac{x - \theta_i}{\delta_i}\right), \quad i = 1, 2, \\ -\infty &< \theta_i < +\infty, \quad \delta_i > 0, \end{aligned}$$

其中 θ_i 和 δ_i 分别为未知的位置参数和尺度参数. 当两个边缘分布受控时, $\theta_i = 0, \delta_i = 1, i = 1, 2$. 当 $(\theta_1, \theta_2) \neq (0, 0), (\delta_1, \delta_2) = (1, 1)$ 时, 表示只存在位置向量偏移; 当 $(\theta_1, \theta_2) = (0, 0), (\delta_1, \delta_2) \neq (1, 1)$ 时, 表示仅存在尺度参数偏移.

对于两样本的位置-尺度检验问题, Marozzi^[30] 给出 Lepage 非参数检验统计量对于不同分布均具有较好的性能, 本文采用了 Lepage 统计量联合监测边缘分布的位置参数和尺度参数. 假设 U_1, U_2, \dots, U_m 为来自一元分布 F_U 的参考样本, W_1, W_2, \dots, W_n 为来自一元分布 F_W 的检验样本, 且相互独立. 将参考样本与检验样本合并, 并按照升序排列全部 $N = m + n$ 个观测值, 检验样本 W_1, W_2, \dots, W_n 的秩记为 $R_k (k = 1, 2, \dots, n)$. 分别用于检验位置参数和尺度参数的 Wilcoxon 秩和 (WRS) 统计量以及 Ansari-Bradley (AB) 统计量定义如下:

$$T_W = \sum_{k=1}^n R_k, \quad T_{AB} = \sum_{k=1}^n \left| \frac{N+1}{2} - R_k \right|.$$

Lepage 统计量是标准化的 WRS 和 AB 统计量的平方和, 定义为

$$L = \begin{cases} \left(\frac{T_W - \frac{n(N+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}} \right)^2 + \left(\frac{T_{AB} - \frac{n(N^2-1)}{4N}}{\sqrt{\frac{mn(N+1)(N^2+3)}{48N^2}}} \right)^2, & N \text{ 为奇数;} \\ \left(\frac{T_W - \frac{n(N+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}} \right)^2 + \left(\frac{T_{AB} - \frac{nN}{4}}{\sqrt{\frac{mn(N^2-4)}{48(N-1)}}} \right)^2, & N \text{ 为偶数.} \end{cases} \quad (1)$$

易见, $E(L|IC) = 2$, 且无论位置参数和尺度参数发生任意类型和方向的偏移, 均有 $E(L|OOC) > 2$. 因此, 本文基于 Lepage 统计量设计单边控制图分别用于监测两个边缘分布 (即情感得分和时间间隔) 的位置参数和尺度参数.

2.2 变量间相关结构的检验统计量

在多元统计建模中, Copula 函数被广泛应用于描述变量间的相关结构. 文献中大多数现有的方法均是假定 Copula 函数的参数化形式, 如 Gaussian Copula、Clayton Copula、Gumbel Copula 等, 然而, 现实过程的复杂性导致难以准确假设参数化的 Copula 函数. 因此, 本文应用一种非参数方法, 通过计算经验 Copula 相对应秩的伪观测值来实现监测变量间的相关结构, 保证了控制图的稳健性.

假设 $\mathbf{Y}_i = (Y_{1i}, Y_{2i}) (i = 1, 2, \dots, r)$ 为二元过程的 r 个观测值. 为了估计 \mathbf{Y}_i 的 Copula 函数, 计算基于秩的 Copula 伪观测值为

$$\left(\frac{R_i}{r+1}, \frac{S_i}{r+1} \right), \quad i = 1, 2, \dots, r. \quad (2)$$

其中: R_i 为 Y_{1i} 在 $Y_{11}, Y_{12}, \dots, Y_{1r}$ 中的秩, S_i 为 Y_{2i} 在 $Y_{21}, Y_{22}, \dots, Y_{2r}$ 中的秩. 易看出, 伪观测值可通过计算每个分量的经验分布函数 \times 尺度因子 $\frac{r}{r+1}$ 得到, 尺度因子是渐近可忽略的, 并确保伪观测值在单位超立方体内, 避免了在边界上的密度估计问题. 在数值模拟和实际应用中, 可采用 R 包 “Copula” 中的 “pobs” 函数快速计算伪观测值.

假设 Γ_j 为参考样本和第 j 个检验样本的混合样本 $\{\Psi_{0m}, \Psi_{jn}\}$ 的 Copula 伪观测值矩阵, 有 $(m+n)$ 行 2 列. 假设 Σ_{Γ_j} 为伪观测矩阵 Γ_j 的协方差阵, $\xi_{\Sigma_{\Gamma_j}}$ 为协方差矩阵 Σ_{Γ_j} 的特征向量组成的 2×2 矩阵, 定义如下:

$$\Theta_j = \Gamma_j \xi_{\Sigma_{\Gamma_j}} = (U_{01}, \dots, U_{0m}, V_{j1}, \dots, V_{jn})'. \quad (3)$$

由式 (3) 可见, Θ_j 为 Γ_j 的主成分得分, 是一个 $(m+n) \times 2$ 的矩阵. 易得出 Θ_j 的协方差矩阵为对角矩阵, 表示 Θ_j 的两个列向量是线性无关的. 定义 Θ_j 的第 l 个列向量为

$$\mathbf{Z}_l = (u_{01l}, u_{02l}, \dots, u_{0ml}, v_{j1l}, v_{j2l}, \dots, v_{jnl})', \quad l = 1, 2.$$

即 \mathbf{Z}_1 与 \mathbf{Z}_2 线性无关. 监测二元过程两个分量相关结构的变化可转换为监测参考样本和检验样本的 Copula 函数, 进一步地, 可通过监测 $u_{0il} (i = 1, 2, \dots, m)$ 和 $v_{jkl} (k = 1, 2, \dots, n)$ 的分布来实现. 当 $C_0(x, y) = C_1(x, y)$ 时, u_{0il} 与 $v_{jkl} (l = 1, 2)$ 的分布应

是相同的; 否则, 表示相关结构发生变化. 因此, 对于任一 $l (l = 1, 2)$, 以 $(u_{01l}, u_{02l}, \dots, u_{0ml})$ 作为参考样本, 以 $(v_{j1l}, v_{j2l}, \dots, v_{jnl})$ 作为第 j 个检验样本, 计算 Lepage 非参数统计量 $L_{C,lj}$ 联合监控 u_{0il} 和 v_{jkl} 分布的位置参数和尺度参数, 并定义如下检验统计量:

$$L_{C,j} = \max\{L_{C,1j}, L_{C,2j}\}.$$

2.3 p 值方法

Li 等^[31] 基于 p 值方法提出了动态非参数控制图, 用于监测一元过程的位置参数, 并给出了应用 p 值方法的优势; 在此基础上, Song 等^[27] 基于分量统计量的 p 值设计了两个稳健控制图, 用于监测和诊断二元过程; Zhang 等^[17] 通过计算 WRS 和 Mood 统计量的 p 值提出了单边非参数 EWMA 控制图, 用于监测线上客户投诉的负面情感得分和时间间隔. 本文应用与 Song 等^[27] 相同的方法, 基于随机置换法计算 Lepage 统计量的 p 值, 具体步骤如下.

step 1: 基于参考样本和检验样本计算 Lepage 统计量, 记为 L_0 ;

step 2: 将参考样本与检验样本混合, 采用无放回方法从混合样本中随机抽取 m 个样本作为参考样本, 剩下的 n 个样本作为检验样本, 重新计算 Lepage 统计量;

step 3: 将 step 2 重复 10000 次, 得到 Lepage 统计量 $L_i (i = 1, 2, \dots, 10000)$ 的经验分布;

step 4: 计算 L_0 大于 $L_i (i = 1, 2, \dots, 10000)$ 的个数, 记为 c , 则估计的 p 值为 $c/10000$.

2.4 二元 EWMA 控制图的设计

EWMA 控制图对于过程参数的中小偏移具有较好的监测效率, 因此, 本节对 p 值采取适当的变换后, 设计了一个二元稳健 EWMA 控制图, 记为 BEWMA-LC, 用于监测和诊断汽车质量线上用户评论的情感得分、时间间隔以及它们的相关结构, 具体设计方法和实施步骤如下.

step 1: 应用阶段 I 方法获得受控过程的参考样本 Ψ_{0m} , 样本容量为 m .

step 2: 基于参考样本序贯监测样本容量为 n 的检验样本, 第 j 个检验样本设为 $\Psi_{jn} (j = 1, 2, \dots)$.

step 3: 对于第 $j (j = 1, 2, \dots)$ 个监测阶段, 将二元向量 Ψ_{0m} 的第 1 个分量观测值 $(X_{01}, X_{02}, \dots, X_{0m})$ 作为参考样本 U , 将 Ψ_{jn} 的第 1 个分量观测值 $(X_{j1}, X_{j2}, \dots, X_{jn})$ 作为检验样本 W , 由式 (1) 计算 Lepage 统计量, 进一步基于第 2.3 节给出的步骤计算 Lepage 统计量对应的 p 值, 记为 $p_{X,j}$. 类似地, 将二元向量 Ψ_{0m} 的第 2 个分量观测值 $(Y_{01}, Y_{02}, \dots,$

Y_{0m})作为参考样本 U ,将 Ψ_{jn} 的第2个分量观测值 $(Y_{j1}, Y_{j2}, \dots, Y_{jn})$ 作为检验样本 W ,由式(1)计算Lepage统计量,进一步基于第2.3节给出的步骤计算Lepage统计量对应的 p 值,记为 $p_{Y,j}$.

step 4: 应用式(2)和(3),基于参考样本 Ψ_{0m} 和检验样本 Ψ_{jn} 计算Copula伪随机数的主成分矩阵 Θ_j ,根据第2.2节给出的方法计算Lepage统计量 $L_{C,j}$,进一步基于第2.3节给出的步骤计算相应的 p 值,记为 $p_{C,j}$,用于监测两个变量间的相关结构.

step 5: 构建BEWMA-LC控制图的检验统计量,如下所示:

$$BEL_j = \max\{E_{X,j}, E_{Y,j}, E_{C,j}\}. \quad (4)$$

其中

$$E_{X,j} = \lambda(-\ln p_{X,j}) + (1 - \lambda)E_{X,j-1},$$

$$E_{Y,j} = \lambda(-\ln p_{Y,j}) + (1 - \lambda)E_{Y,j-1},$$

$$E_{C,j} = \lambda(-\ln p_{C,j}) + (1 - \lambda)E_{C,j-1}.$$

这里: λ 为平滑参数, \ln 表示以 e 为底的对数函数, $-\ln p_{X,j}$ 、 $-\ln p_{Y,j}$ 、 $-\ln p_{C,j}$ 为分别用于监测3个假设 H_{01} 、 H_{02} 、 H_{03} 的分量统计量Lepage的 p 值变换.应用基本的概率计算可得到,当过程受控时, $E(-\ln p_{X,j}) = E(-\ln p_{Y,j}) = E(-\ln p_{C,j}) = 1$.因此,EWMA统计量的初值设为 $E_{X,0} = E_{Y,0} = E_{C,0} = 1$.

step 6: 确定BEWMA-LC控制图的上控制限(upper control limit, UCL),记为 H_L .

step 7: 比较检验统计量 BEL_j 与上控制限 H_L .当 $BEL_j > H_L$ 时,表示第 j 个检验样本失控,控制图发出异常报警,需要进一步诊断失控原因;否则,表示过程受控,继续监测下一个检验样本.

step 8: 二元控制图异常报警后的诊断至关重要,能够为最终的决策提供信息和依据.当 $BEL_j > H_L$ 时,只需要进一步比较3个分量EWMA统计量与上控制限 H_L ,若分量EWMA统计量超出上控制限,则对应的变量或相关结构可能发生异常,如:当 $E_{X,j} > H_L$, $E_{Y,j} < H_L$, $E_{C,j} < H_L$ 时,表示二元过程第1个分量可能发生异常,而第2个分量和相关结构是受控状态.

3 实例应用

3.1 数据获取和预处理

本文以近年来销量很高且稳定的车型大众速腾为例,选择其在全国汽车消费者投诉受理处置服务平台“车质网”(http://www.12365auto.com)中的投诉文本进行分析和监控.首先,利用Python中的Requests库编写爬虫程序,获取“大众速腾”车型从2010年

4月26日~2021年9月17日的全部结构化和非结构化用户投诉数据,包括:投诉编号、投诉具体车型、投诉时间、问题简述和具体投诉内容,共7837条,对其进行去重处理,去重后为5015条;然后,对5015条投诉文本中的问题简述和具体投诉内容进行Jieba中文分词和过滤停用词,其中在分词过程中,加入汽车领域专业词典,以提高分词精度.

3.2 汽车质量缺陷主题的提取和分类

对“大众速腾”汽车的具体投诉数据预处理后,构建词典,将语料向量化表示.应用第1.1节给出的算法步骤创建LDA模型,将主题个数设置为1~15进行循环,求出不同主题个数下的一致性,同时得到各主题的词汇分布.计算结果显示,当主题个数为6时,一致性最大.因此,选择最佳主题个数为6,并输出主题个数为6时每个主题概率最高的前3个词汇分布,具体分别为{“异响”“刹车”“变速箱”}、{“变速箱”“DSG”“离合器”}、{“灯”“亮”“传感器”}、{“异响”“发动机”“油泵”}、{“悬挂”“断裂”“悬架”}、{“轮胎”“纵臂”“变形”}.在主题词汇分布的基础上,结合实际具体投诉文本内容和领域知识,提取出6个汽车质量缺陷主题,分别为变速箱异响、刹车异响、变速箱顿挫、发动机异响、悬挂断裂、故障灯亮.

确定汽车质量缺陷主题后,对每条投诉文本进行分类.为保证分类精度,根据问题简述中包含的汽车部件和症状词汇,找到每种质量缺陷主题下的全部投诉文本.在一条投诉文本中,往往可能存在多种质量问题,因此,不同质量缺陷主题下的投诉文本可能会重复.对各汽车质量缺陷主题下的投诉文本数进行统计,结果为变速箱异响579条、刹车异响316条、变速箱顿挫260条、发动机异响247条、悬挂断裂161条、故障灯亮145条.

3.3 时间间隔和情感分数的计算

按照第1.2节中所述方法,计算各缺陷主题下投诉文本的时间间隔和情感得分.在计算情感分数时,采用Hownet情感词典,包括积极情感词典、消极情感词典、否定词表和程度词表.程度词表中按照程度的大小区分为6个等级的子词表,分别为“最(most)”“非常(very)”“过于(over)”“比较(more)”“稍微(ish)”“不足(insufficiently)”.加权计算时,参见Ku等^[25]、刘思远^[32],并结合实际投诉文本,设置这6个等级对应的权重分别为2、1.8、1.6、1.4、1.2、0.5,同时设置否定词的权重为-1.在中文情感表达中,句尾表示结束的标点符号具有一定的情感增强作用,

当句中含有计入得分的情感词且句尾存在感叹号或问号时,情感得分增加值为2.

3.4 汽车质量缺陷主题的监测与诊断

对于每个汽车质量缺陷主题的时间间隔和情感得分,选取不同常用分布,如正态分布、Gamma分布等,进行拟合优度检验,得到的 p 值均小于0.05,表明时间间隔和情感得分的参数分布难以确定.因此,本节应用无需事先指定过程分布的稳健的 BEWMA-LC 控制图对汽车质量缺陷主题的时间间隔和情感得分进行监控和诊断.为评估不同控制图在实际应用中的监测效果,将所提出 BEWMA-LC 控制图与现有控制图进行性能对比,包括: Mukherjee 等^[33]基于 Mathur 统计量设计的二元非参数控制图,记为 BEWMA-Mathur; Mukherjee 等^[34]提出了一类基于距离度量的非参数 Shewhart 控制图,用于监控多变量和高维过程,文章结论指出,从总体上看 HDSOR 控制图的性能更优,为方便比较,基于 HDSOR 统计量构造了相应的二元 EWMA 控制图,记为 BEWMA-OR;经典的基于数据深度的多变量控制图,由 Liu^[35]和 Hamurkaroglu 等^[36]可知,数据深度有许多不同的定义,其中基于 Mahalanobis 距离的深度概念是最简单和最受研究人员欢迎的概念之一,选取其对应的二元 EWMA 控制图进行比较,记为 BEWMA-QMD 控制图.

首先,应确定阶段 I 样本(参考样本),采用 Capizzi 等^[29]提出的非参数阶段 I 分析方法,应用 R 包“dfphase1”的“mshewhar”函数,计算并绘制阶段 I 控制图,用于监测多变量过程的位置和尺度参数.6个汽车质量缺陷主题的阶段 I 分析结果如下:1)对于主题变速箱异响,根据阶段 I 控制图结果,选取从2015年2月6日~2017年12月25日共102对数据作为参考样本;2)对于主题刹车异响,基于阶段 I 控制图结果,选取从2017年2月24日~2020年6月14日共104对数据作为参考样本;3)对于主题变速箱顿挫,阶段 I 控制图结果显示从2018年5月2日~2020年5月25日的的数据是受控的,采用这80对数据作为参考样本;4)对于主题发动机异响,由阶段 I 分析结果可见,从2018年11月24日~2020年12月8日的94对数据没有报警点,可作为参考样本;5)对于主题悬挂断裂,根据阶段 I 控制图结果,采用2015年1月4日~2015年3月31日的53对数据作为参考样本;6)对于主题故障灯亮,基于阶段 I 分析结果,选取2015年2月13日~2019年8月21日共41对数据作为参考样本.

确定参考样本后,应用 BEWMA-LC、BEWMA-Mathur、BEWMA-OR 和 BEWMA-QMD 控制图对后续数据(检验样本)进行在线监控.链长(run length, RL)定义为从监测启动到控制图首次触发异常警报所需的样本组数,其数学期望称为平均链长(average run length, ARL),是统计过程控制中衡量控制图性能的核心指标.在实际应用中,通常采用标准化对比方法,通过预先设定过程受控时的 ARL(记为 ARL_0 ,如工业中广泛采用 $ARL_0 = 370.4$,对应0.27%的误报率),评估不同控制图在异常偏移(过程失控)下的 ARL(记为 ARL_1). ARL_1 越小,控制图对特定参数偏移的检测效率越高.本文选取检验样本容量 $n = 10$,设定控制图参数 $\lambda = 0.1$, $ARL_0 = 370.4$.通过蒙特卡洛模拟分别找到6个汽车质量缺陷主题对应的 BEWMA-LC、BEWMA-Mathur 和 BEWMA-OR 控制图的上控制限,以及 BEWMA-QMD 控制图的下控制限(lower control limit, LCL).对于 BEWMA-LC 控制图,应用第2.4节给出的步骤计算每个主题的检验样本对应的检验统计量值 $BEL_j = \max\{E_{X,j}, E_{Y,j}, E_{C,j}\}$ (式(4)),当检验统计量超出上控制限时,在该样本点处控制图发生报警,结果如图2所示.注意到,计算获得的检验统计量值存在无穷大(“ ∞ ”)的情况,表示3个分量统计量对应的 p 值中至少有一个为0,即观测到的 Lepage 统计量值很大,是一种强烈的失控报警信号.为了绘图清晰,图2中超出2倍上控制限($2 \times H_L$)的检验统计量值均用 $2 \times H_L$ 的值表示,并用星号“*”标记.控制图发生失控报警后及时准确地识别报警原因有助于企业制定合理有效的措施.在 BEWMA-LC 控制图发出报警信号后,只需要分别比较3个分量统计量与上控制限的值,超出上控制限的分量统计量对应的变量或相关结构可能发生异常.图3~图5分别为 BEWMA-Mathur、BEWMA-OR 和 BEWMA-QMD 控制图的监测结果,即每个主题的检验样本对应的检验统计量值以及相应的控制限.对于 BEWMA-Mathur 和 BEWMA-OR 控制图,当检验统计量超过上控制限 H_M 和 H_O 时,表示在该样本点处控制图发生报警.对于 BEWMA-QMD 控制图,异常信号表现为检验统计量低于下控制限 H_Q ,此时触发报警.

由图2~图5可见:对于主题变速箱异响, BEWMA-Mathur 和 BEWMA-OR 控制图在第10个检验样本首次发出报警, BEWMA-LC 控制图从第12个检验样本开始报警直至第43个样本,而 BEWMA-QMD 控制图在第28个样本点才发出异常

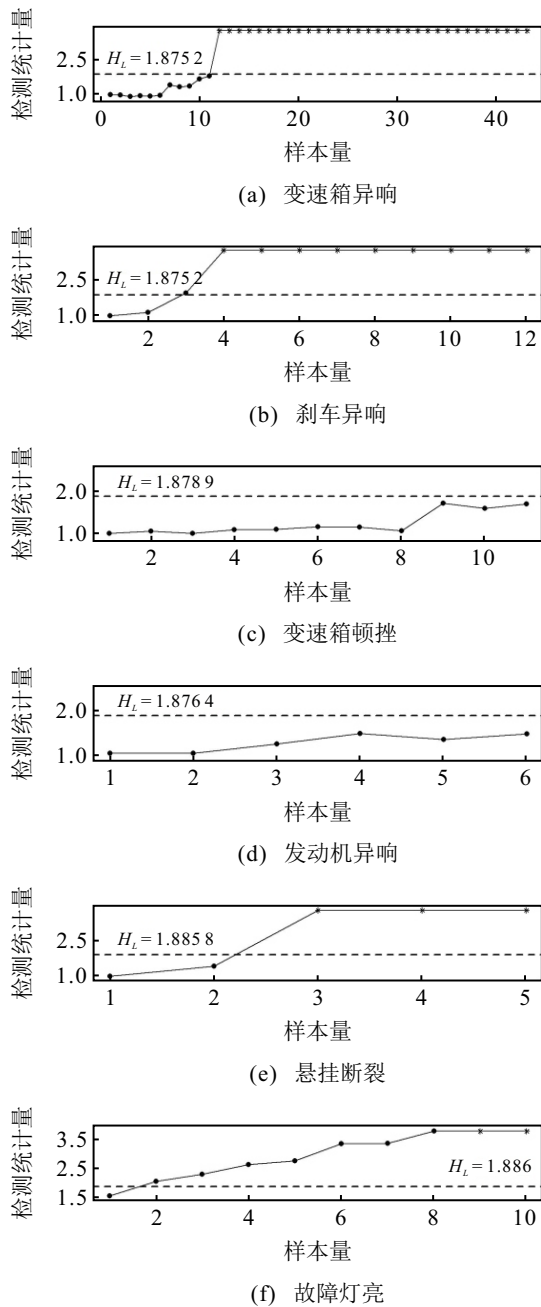


图2 阶段 II 监测汽车质量缺陷主题的时间间隔和情感得分的 BEWMA-LC 控制图

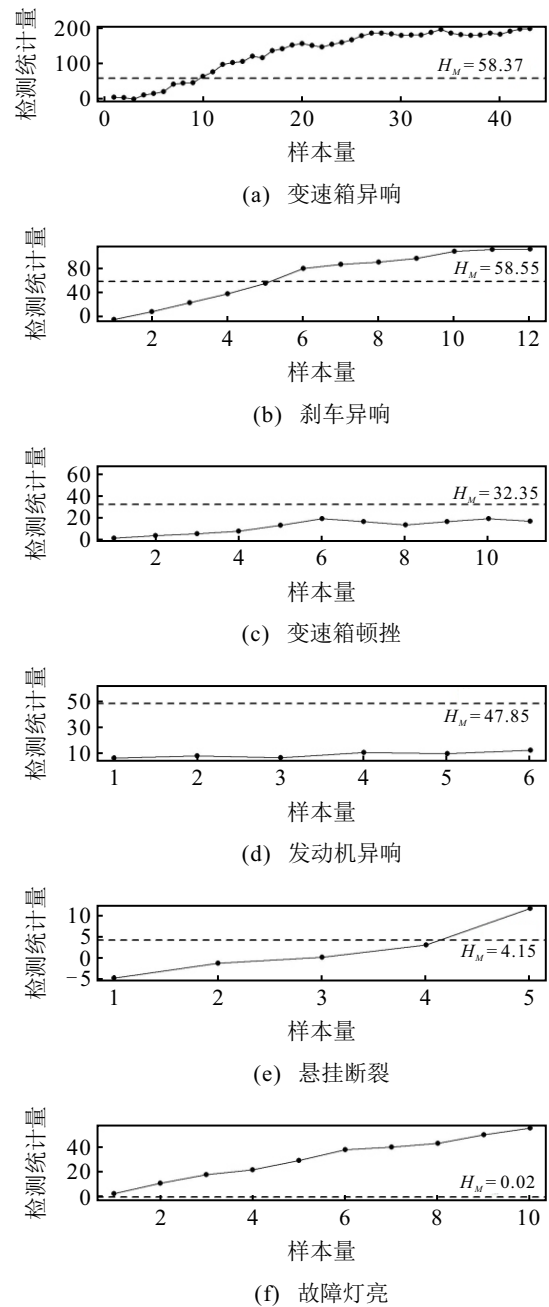


图3 阶段 II 监测汽车质量缺陷主题的时间间隔和情感得分的 BEWMA-Mathur 控制图

信号. 应用 BEWMA-LC 控制图进一步诊断报警原因, 样本点 12 ~ 样本点 43 均是由于时间间隔和相关性可能发生变化引起的过程失控, 观察原始数据可以发现时间间隔明显存在减少的趋势, 表明从 2019 年 9 月 20 日开始投诉的人数增多, 频率增加, 可能会存在变速箱异响的共性问题, 需要引起企业的重视. 对于主题刹车异响, BEWMA-LC 控制图在第 3 个样本点 ~ 第 12 个样本点处发生报警, 而 BEWMA-OR 控制图从第 5 个样本点开始报警, BEWMA-Mathur 和 BEWMA-QMD 控制图从第 6 个样本点开始报警. 根据 BEWMA-LC 控制图的诊断, 对于样本点 3 和样本点 4 报警原因是时间间隔可能

发生了变化. 对于样本点 5 ~ 样本点 12 是由于时间间隔和相关性同时发生变化. 对于主题变速箱顿挫和发动机异响, 4 个控制图均未发出报警, 在监测中处于受控状态. 对于主题悬挂断裂, BEWMA-LC、BEWMA-OR 和 BEWMA-QMD 控制图在样本点 3 ~ 样本点 5 发出报警, 而 BEWMA-Mathur 仅在第 5 个样本点给出报警信号. 经过诊断分析, 在第 3 个样本点处, 报警信号是由于时间间隔对应的分量统计量超出上控制限引起的. 对于样本点 4 和样本点 5 报警原因是时间间隔和相关性可能同时发生了变化. 结合原始数据可以发现, 从 2015 年 6 月 11 日开始关于悬挂断裂的投诉时间间隔增加, 表示该汽车

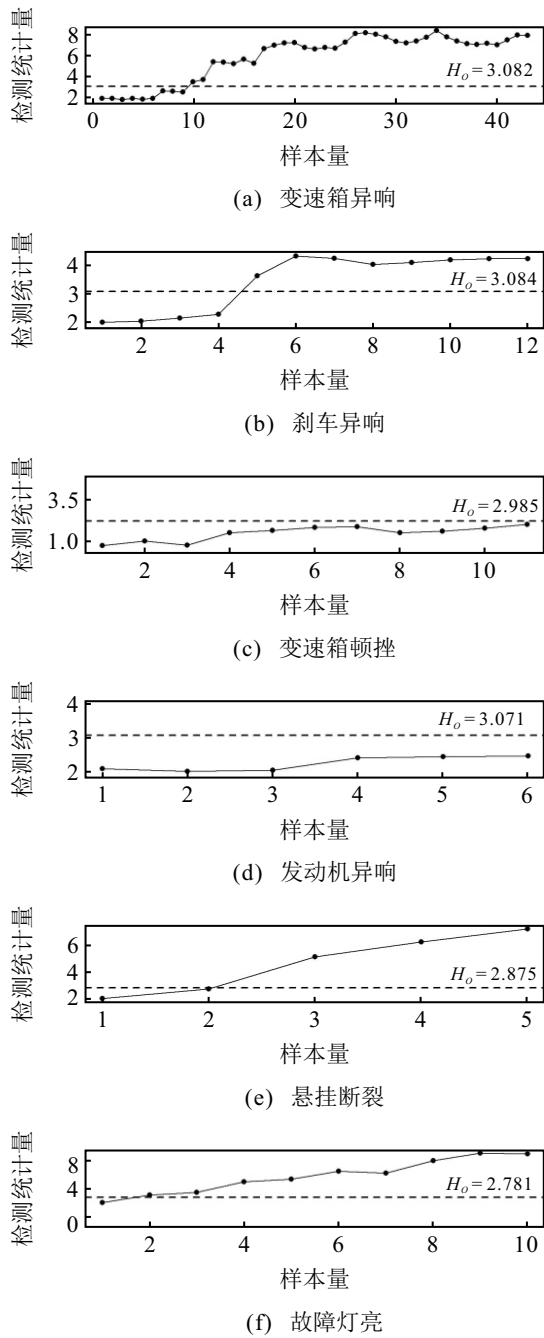


图4 阶段 II 监测汽车质量缺陷主题的时间间隔和情感得分的 BEWMA-OR 控制图

质量问题的投诉人数减少. 经核查, 在 2012 年 ~ 2014 年速腾频繁发生断轴问题, 一汽大众采取相应措施较好地解决了该质量问题, 减少了投诉. 对于主题故障灯亮, BEWMA-LC 和 BEWMA-OR 控制图在第 2 个样本点 ~ 第 10 个样本点处发出异常报警, 而 BEWMA-Mathur 和 BEWMA-QMD 控制图从第 1 个样本点给出异常. 应用 BEWMA-LC 进行诊断, 异常报警均是由时间间隔和相关性可能发生了变化引起的. 综上, BEWMA-LC 控制图在监测各质量缺陷主题时能够较早地发出报警, 且只有 BEWMA-LC 控制图在报警后可以异常诊断, 因此, 在实际应用

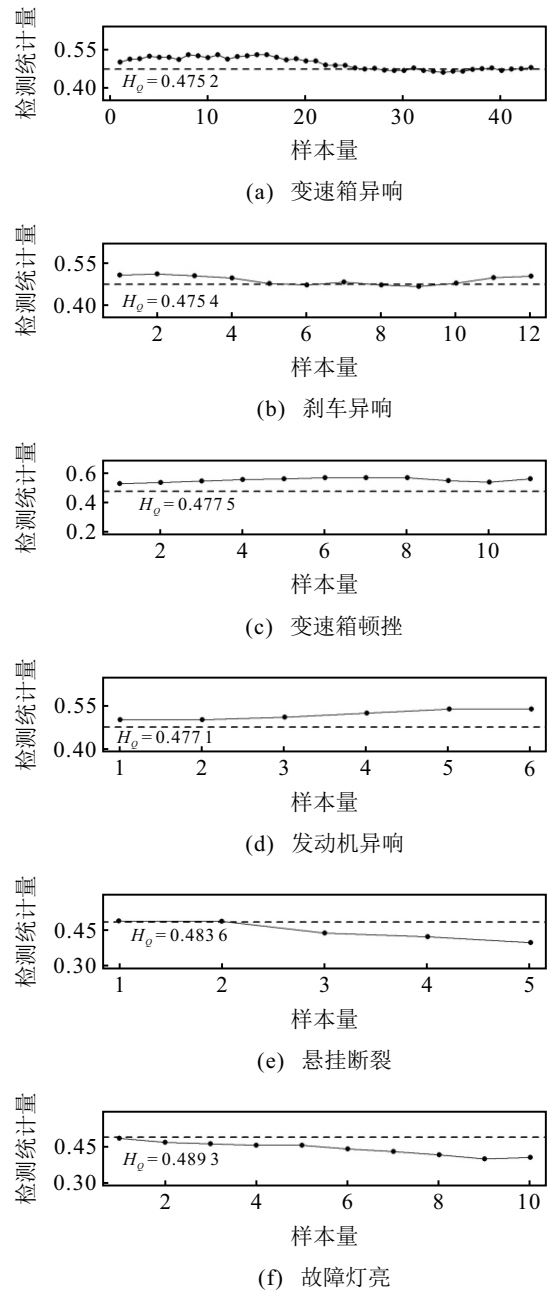


图5 阶段 II 监测汽车质量缺陷主题的时间间隔和情感得分的 BEWMA-QMD 控制图

中推荐使用 BEWMA-LC 控制图.

4 结论

合理利用汽车质量在线投诉文本是实现汽车质量监控的有效途径, 符合利用数据驱动产品质量管理工程的发展趋势. 本文选取了“车质网”中“大众速腾”车型在 2010 年 ~ 2021 年的投诉文本, 利用 LDA 模型提取了汽车质量缺陷主题, 并将投诉文本按照主题分类, 便于诊断异常来源. 对于每个缺陷主题下的投诉文本, 计算时间间隔和情感得分, 从投诉频率和投诉强度两个维度进行了分析. 由于时间间隔和情感得分的分布难以确定, 基于 Sklar 定理结合 Lepage 非参数统计量和经验 Copula 方法提出了

BEWMA-LC 控制图联合监测投诉文本的时间间隔、情感得分以及它们的相关结构. BEWMA-LC 控制图具有稳健和易于识别报警原因的优势. 在实例分析中, 将 BEWMA-LC 控制图与 BEWMA-Mathur、BEWMA-OR 和 BEWMA-QMD 控制图进行比较, 对比结果显示, 在多数情况下, BEWMA-LC 控制图能够更及时地发出警报, 并在报警后可以进行异常诊断, 从而帮助车企做出更有效的决策.

本文在挖掘汽车质量问题的基础上, 对汽车质量进行了进一步的动态分析和联合监控. 在未来的研究中, 可以尝试使用其他提取主题的方法获得汽车质量缺陷主题, 同时提高主题分类准确性^[37]. 将投诉文本的其他特征, 如文本长度、相似度等作为监控指标, 提高监测效果. 另外, 还可以通过数据预处理、多源数据融合、模型优化等手段进一步改进反馈机制, 构建更全面、准确的汽车质量预警体系.

参考文献 (References)

- [1] 中国消费者协会. 2021 年全国消协组织受理投诉情况分析 [EB/OL]. (2022-01-28)[2023-09-15]. <https://www.cca.org.cn/Detail?catalogId=475800865030213&contentType=article&contentId=521578370699333>. (China Consumers' Association. Analysis of complaints handled by consumer associations across the country in 2021 [EB/OL]. (2022-01-28)[2023-09-15]. <https://www.cca.org.cn/Detail?catalogId=475800865030213&contentType=article&contentId=521578370699333>.)
- [2] Xu X. Examining the relevance of online customer textual reviews on hotels' product and service attributes [J]. *Journal of Hospitality & Tourism Research*, 2019, 43(1): 141-163.
- [3] 刘孝炎, 唐焕玲, 王育林, 等. 一种去注意力机制的动态多层语义感知机 [J]. *控制与决策*, 2024, 39(2): 588-594. (Liu X Y, Tang H L, Wang Y L, et al. A dynamic multi-layer semantics perceptron without attention mechanism [J]. *Control and Decision*, 2024, 39(2): 588-594.)
- [4] Abrahams A S, Jiao J, Wang G A, et al. Vehicle defect discovery from social media [J]. *Decision Support Systems*, 2012, 54(1): 87-97.
- [5] Liu Y, Jiang C, Ding Y, et al. Identifying helpful quality-related reviews from social media based on attractive quality theory [J]. *Total Quality Management & Business Excellence*, 2017, 1: 1-20.
- [6] Abrahams A S, Fan W, Wang G A, et al. An integrated text analytic framework for product defect discovery [J]. *Production and Operations Management*, 2015, 24(6): 975-990.
- [7] Kakde D, Chaudhuri A. Leveraging unstructured data to detect emerging reliability issues [C]. *Annual Reliability and Maintainability Symposium*. Tucson, 2016: 1-6.
- [8] 王余行, 党延忠, 徐照光. 针对论坛数据特点的汽车质量问题挖掘 [J]. *中国管理科学*, 2021, 29(9): 201-212. (Wang Y H, Dang Y Z, Xu Z G. Mining automobile quality problems based on the characteristics of forum data [J]. *Chinese Journal of Management Science*, 2021, 29(9): 201-212.)
- [9] Song Z, Mukherjee A, Qiu P H, et al. Two robust multivariate exponentially weighted moving average charts to facilitate distinctive product quality features assessment [J]. *Computers & Industrial Engineering*, 2023, 183: 109469.
- [10] 胡雪君, 王建江, 崔南方. 基于统计过程控制的两阶段缓冲监控方法 [J]. *控制与决策*, 2020, 35(6): 1453-1462. (Hu X J, Wang J J, Cui N F. Two-stage buffer monitoring method based on statistical process control [J]. *Control and Decision*, 2020, 35(6): 1453-1462.)
- [11] 宋贇, 刘艳春, 陶桂洪. 阶段 II 联合检测过程位置和尺度的非参数 EWMA 控制图 [J]. *应用概率统计*, 2019, 35(6): 639-653. (Song Z, Liu Y C, Tao G H. A distribution-free phase II EWMA control chart for joint monitoring of location and scale [J]. *Chinese Journal of Applied Probability and Statistics*, 2019, 35(6): 639-653.)
- [12] Song Z, Mukherjee A, Tao G H. A class of distribution-free one-sided Cucconi schemes for joint surveillance of location and scale parameters and their application in monitoring cab services [J]. *Computers & Industrial Engineering*, 2020, 148: 106625.
- [13] Song Z, Mukherjee A, Ma N, et al. A class of new nonparametric circular-grid charts for signal classification [J]. *Quality and Reliability Engineering International*, 2021, 37(6): 2738-2759.
- [14] Mukherjee A, Marozzi M. A distribution-free phase-II CUSUM procedure for monitoring service quality [J]. *Total Quality Management & Business Excellence*, 2017, 28(9/10): 1227-1263.
- [15] 吴纯杰, 李泓玮. 基于非参数控制图的数据质量监控方法改进研究 [J]. *数理统计与管理*, 2024, 43(1): 1-19. (Wu C J, Li H W. Improvement on data quality monitoring methods based on nonparametric control charts [J]. *Journal of Applied Statistics and Management*, 2024, 43(1): 1-19.)
- [16] 王晗, 宋贇. 在线投诉文本驱动的汽车质量监控方法研究 [J]. *数理统计与管理*, 2024, 43(5): 789-799. (Wang H, Song Z. Research on automobile quality monitoring method driven by online complaint text [J]. *Journal of Applied Statistics and Management*, 2024, 43(5): 789-799.)
- [17] Zhang T X, He Z, Zhao X J, et al. Joint monitoring of post-sales online review processes based on a distribution-free EWMA scheme [J]. *Computers & Industrial Engineering*, 2021, 158: 107372.
- [18] Zhang T X, He Z, Mukherjee A. Monitoring negative sentiment scores and time between customer complaints via one-sided distribution-free EWMA schemes [J].

- Computers & Industrial Engineering, 2023, 180: 109247.
- [19] Blei D M, Ng A, Jordan M I. Latent dirichlet allocation[J]. *The Journal of Machine Learning Research*, 2003, 3: 993-1022.
- [20] 蒋翠清, 王齐林, 刘士喜, 等. 中文社交媒体环境下半监督学习的汽车缺陷识别方法[J]. *中国管理科学*, 2014, 22(S1): 677-685.
(Jiang C Q, Wang Q L, Liu S X, et al. Semi-supervised learning for automobile defect identification in the context of Chinese social media[J]. *Chinese Journal of Management Science*, 2014, 22(S1): 677-685.)
- [21] 赵越, 郝琨, 时彩云, 等. 跨模 LDA 融合的多模态数据主题分析方法[J]. *控制与决策*, 2024, 39(4): 1325-1332.
(Zhao Y, Hao K, Shi C Y, et al. Multimodal data topic analysis method based on cross-modal LDA fusion[J]. *Control and Decision*, 2024, 39(4): 1325-1332.)
- [22] Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures[C]. *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. Shanghai, 2015: 399-408.
- [23] 李杨, 徐泽水, 王新鑫. 基于在线评论的情感分析方法及应用[J]. *控制与决策*, 2023, 38(2): 304-317.
(Li Y, Xu Z S, Wang X X. Methods and applications of sentiment analysis with online reviews[J]. *Control and Decision*, 2023, 38(2): 304-317.)
- [24] Gomes H, Neto M, Henriques R. Text mining: Sentiment analysis on news classification[C]. *Proceedings of the 8th Iberian Conference on Information Systems and Technologies*. Lisboa, 2013: 1-6.
- [25] Ku L W, Chen H H. Mining opinions from the web: Beyond relevance retrieval[J]. *Journal of the American Society for Information Science and Technology*, 2007, 58(12): 1838-1850.
- [26] Sklar A. Fonctions de répartition à n dimensions et leurs marges[J]. *Publications de l'Institut Statistique de l'Université de Paris*, 1959, 8: 229-231.
- [27] Song Z, Mukherjee A, Zhang J J. Some robust approaches based on copula for monitoring bivariate processes and component-wise assessment[J]. *European Journal of Operational Research*, 2021, 289(1): 177-196.
- [28] Jones-Farmer L A, Woodall W H, Steiner S H, et al. An overview of phase I analysis for process improvement and monitoring[J]. *Journal of Quality Technology*, 2014, 46(3): 265-280.
- [29] Capizzi G, Masarotto G. Phase I distribution-free analysis with the R package dfphase1[C]. *Frontiers in Statistical Quality Control 12*. Magdeburg, 2018: 3-19.
- [30] Marozzi M. Nonparametric simultaneous tests for location and scale testing: A comparison of several methods[J]. *Communications in Statistics-Simulation and Computation*, 2013, 42(6): 1298-1317.
- [31] Li Z H, Qiu P H. Statistical process control using a dynamic sampling scheme[J]. *Technometrics*, 2014, 56(3): 325-335.
- [32] 刘思远. 基于情感词典的美妆领域在线评论文本情感分类研究[D]. 武汉: 华中师范大学, 2022: 1-63.
(Liu S Y. Sentiment classification of online comment texts in beauty makeup based on sentiment dictionary[D]. Wuhan: Central China Normal University, 2022, 1-63.)
- [33] Mukherjee A, Cheng Y, Gong M. A new nonparametric scheme for simultaneous monitoring of bivariate processes and its application in monitoring service quality[J]. *Quality Technology & Quantitative Management*, 2017, 14(1): 1-14.
- [34] Mukherjee A, Marozzi M. Nonparametric phase-II control charts for monitoring high-dimensional processes with unknown parameters[J]. *Journal of Quality Technology*, 2022, 54(1): 44-64.
- [35] Liu R Y. Control charts for multivariate processes[J]. *Journal of the American Statistical Association*, 1995, 90(432): 1380-1387.
- [36] Hamurkaroğlu C, Mert M, Saykan Y. Nonparametric control charts based on Mahalanobis depth[J]. *Hacettepe Journal of Mathematics and Statistics*, 2004, 33: 57-67.
- [37] 张书涛, 杨志强, 苏建宁, 等. 用户评论异质情感的主题聚类仿真[J]. *控制与决策*, 2024, 39(11): 3645-3654.
(Zhang S T, Yang Z Q, Su J N, et al. Topic clustering simulation of heterogeneous emotions in user comments[J]. *Control and Decision*, 2024, 39(11): 3645-3654.)

作者简介

宋赞 (1982-), 女, 副教授, 博士, 主要研究方向为统计质量控制、异常诊断, E-mail: zhisong@syau.edu.cn;

王晗 (1991-), 女, 讲师, 主要研究方向为统计质量控制、机器学习, E-mail: hanwang@syau.edu.cn;

张久军 (1978-), 男, 教授, 博士, 博士生导师, 主要研究方向为统计质量控制、异常诊断, E-mail: zjjly790816@163.com;

胡雪龙 (1988-), 男, 副教授, 博士, 主要研究方向为管理科学与工程、机器学习, E-mail: hx10419@njupt.edu.cn.