8. 综合实践II

一、由实际问题出发

01

背景介绍

手机行业发展现状

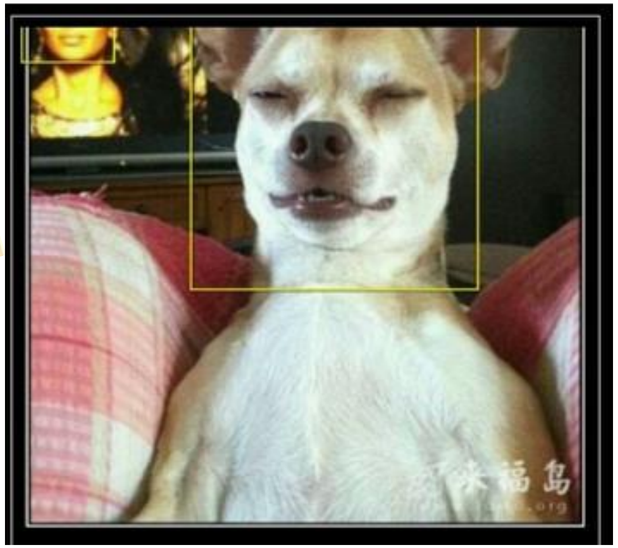
1. 市场容量大
2. 发展空间广阔
3. 市场竞争日益白热化
4. 差异化需求迫切

2017年9月13日，苹果推出了iphone 8

发布时

- ✓ 一体化金属边框
- ✓ 正反两面的玻璃材质
- ✓ 无线充电
- ✓ 人脸识别
- ✓ ……

发布后



如何挖掘**真实**的用户需求，
更好的改进产品呢？？

丰富的线上销售渠道 → 鼓励用户提交使用感受 → 用户反馈展现了对手机的关注点 → 通过关注点抓住客户需求，改善产品和服务

电池还是比较耐用，
充电速度也快。

电池是用户的主要关
注点，应该继续保持

手机整体不错，就
是**包装**太简单了，
不防摔。

包装被吐槽了，应该
使用多使用抗摔材料

开机开不了，一直
这个状态，问**客服**
爱理不理，退货

开机怎么有问题，赶紧
查！**客服**态度不好，还
想不想干了！！

02

数据介绍

数据概况

- 截止2016年11月31日，某知名电商在其自营平台上销售过的手机数据及能爬到的全部用户评论数据。

297部

涉及信息包括三方面（手机的各种参数指标、销售平台的促销情况、评论总数）

01

手机数据

用户评论数据

02

216754条

涉及每条评论的评分以及具体内容

数据变量信息表

手机参数

后置摄像头像素

前置摄像头像素

屏幕尺寸

品牌、价格

GPS定位

指纹识别

数据变量信息表

评论信息

评论评分

手机是原装的未拆封，手感很好，京东包裹不错还要验证码才能收货，手机很流畅，老婆很满意！

玫瑰金色 公开版 128GB 2017-10-08 12:24

评论内容

忘记拍了，拍另外一部嘻嘻也是7p，128用的嗨呀！！！！

玫瑰金色 公开版 128GB 2017-10-07 11:33

给媳妇儿买的手机，不错，她很喜欢。京东快递就是快，上午买的，下午就到货，点赞

玫瑰金色 公开版 128GB 2017-10-02 20:41

购买时间

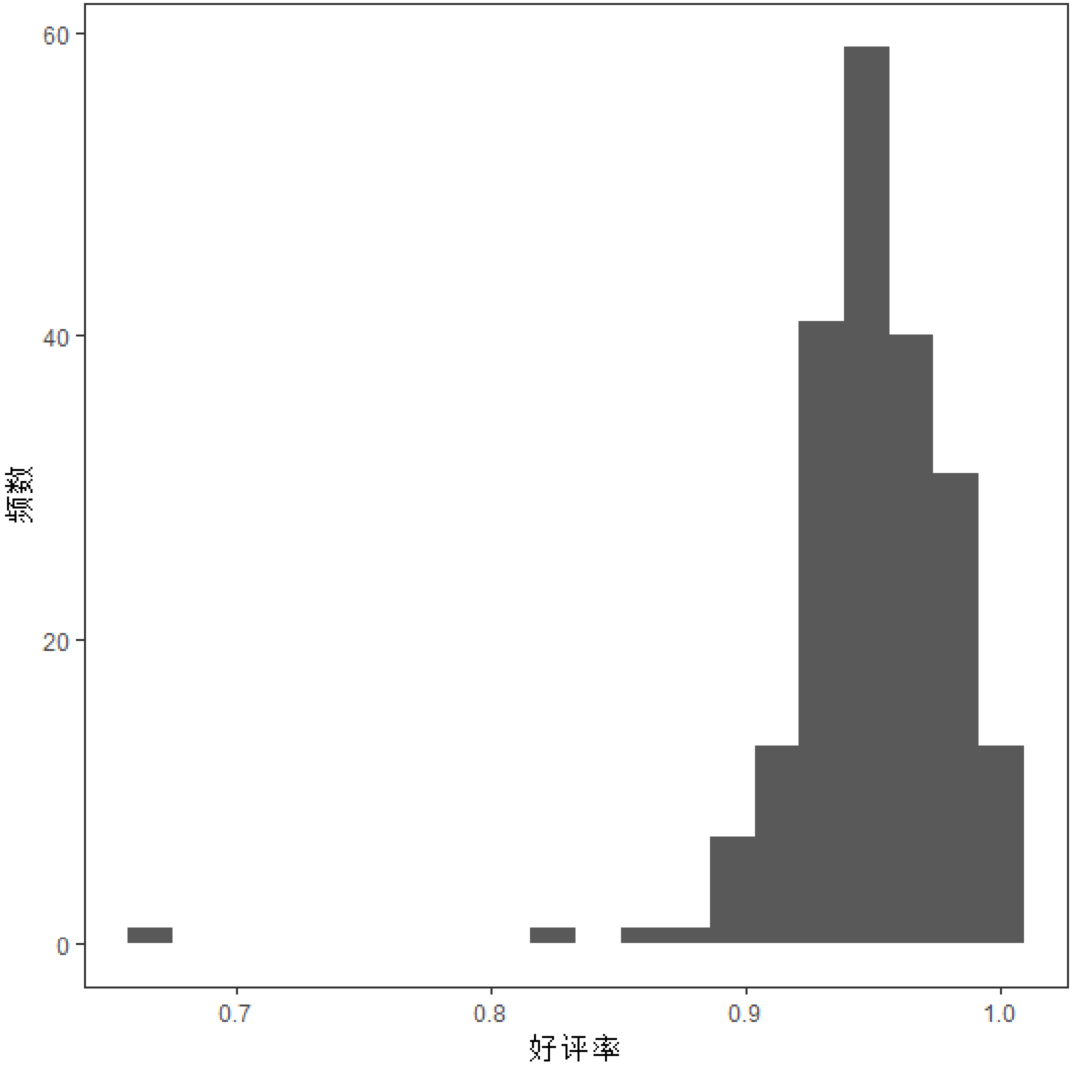
03

评论初探

因变量：好评率

定义手机的好评率=好评数/总评论数

每部手机的好评率为何
如此参差不齐？
评论中都说了什么？



评论的预处理

中文分词

原始评论： 电池还是比较耐用，充电速度也快。
分词后评论： 电池 | 还是 | 比较 | 耐用 | 充电 | 速度 | 也 | 快

去停用词

停用词： 用来维持语义完整性，但是没有特殊含义的词，如“的、了、是”
去停用词之后： 电池 | 耐用 | 充电 | 速度 | 快

提取高频词

统计词频： 计算每个词在所有评论中出现的**总次数**
词频排序： 按词频**由高到低**排序，提取词频最高的前n个词

绘制词云

词云图： 高频词的可视化展示

词云图：大家在评论中谈论了什么

好评（评分>3）词云

差评（评分<3）词云

提取热评词

01

提取有明确业务含义的热评词：

从高频词（前50）中，提取所有描述【服务特征】和【手机特征】的词作为热评词

02

检验热评词是否对评分有显著影响：

“包含该热评词的评论”与“不包含该热评词的评论” 在平均分上是否有显著差异（**t检验**）

03

提取显著的热评词为解释变量

计算任意一部手机的所有评论中出现该热评词的频率，作为新的解释变量

服务特征

速度、物流、送货、快递、包装、客服、售后、发票

手机特征

屏幕、电池、系统、性价比、质量、外观、功能、运行、充电、声音、耳机、信号、开机、软件、拍照

变量汇总表

变量类别		变量名称	说明
因变量		log(好评率)	
自变量（来自手机）	手机特征	价格	以“千元”为单位
		品牌	取总评论数最多的前8大品牌，剩余品牌为“其他”；以“其他”为基准
		屏幕尺寸	
		前置摄像头	以500、1000为界，划分为“高、中、低”三挡；以“低”为基准
		后置摄像头	以1000、1500为界，划分为“高、中、低”三挡；以“低”为基准
		指纹识别	以“不支持”为基准
		GPS	以“不支持”为基准
	促销信息	促销信息	以“无”为基准
自变量（来自评论）	字符统计	平均字符数	每部手机所有评论的平均字符数
	服务特征	7个热评词	速度、物流、送货、快递、客服、售后、发票
	手机特征	14个热评词	屏幕、电池、系统、性价比、质量、外观、功能、运行、充电、声音、耳机、信号、开机、软件

建立回归模型（用BIC选择）

调整后R方

		估计值	P值	显著性
截距		-0.158	0.000	***
价格		0.004	0.000	***
品牌	华为	0.029	0.000	***
	OPPO	0.023	0.022	*
	VIVO	0.024	0.010	*
屏幕尺寸		0.015	0.001	**
平均字符数		0.001	0.000	***
热评词	物流	0.137	0.000	***
	客服	-0.313	0.000	***
	电池	-0.161	0.000	***
	运行	-0.192	0.003	**

模型解读

- ✓ 手机的价格越高，好评率越好，说明价格高的手机功能更趋完善，更能让用户满意；
- ✓ 与“其他”品牌相比，华为、OPPO和VIVO三大品牌的手机好评率更高；
- ✓ 手机屏幕的尺寸越大，手机的好评率越高，说明用户更钟爱大屏手机。

- ✓ 手机评论的字符数越多，手机的好评率越高；
- ✓ 物流在手机评论中出现的频率越高，手机的好评率越高，说明物流是手机的加分项；
- ✓ 客服、电池、运行三个热评词在手机评论中出现的频率越高，手机的好评率反而越低，说明这三点是手机的减分项。

04

深挖热评词

热评词好在哪？差在哪？

提取只包含热评词的句子

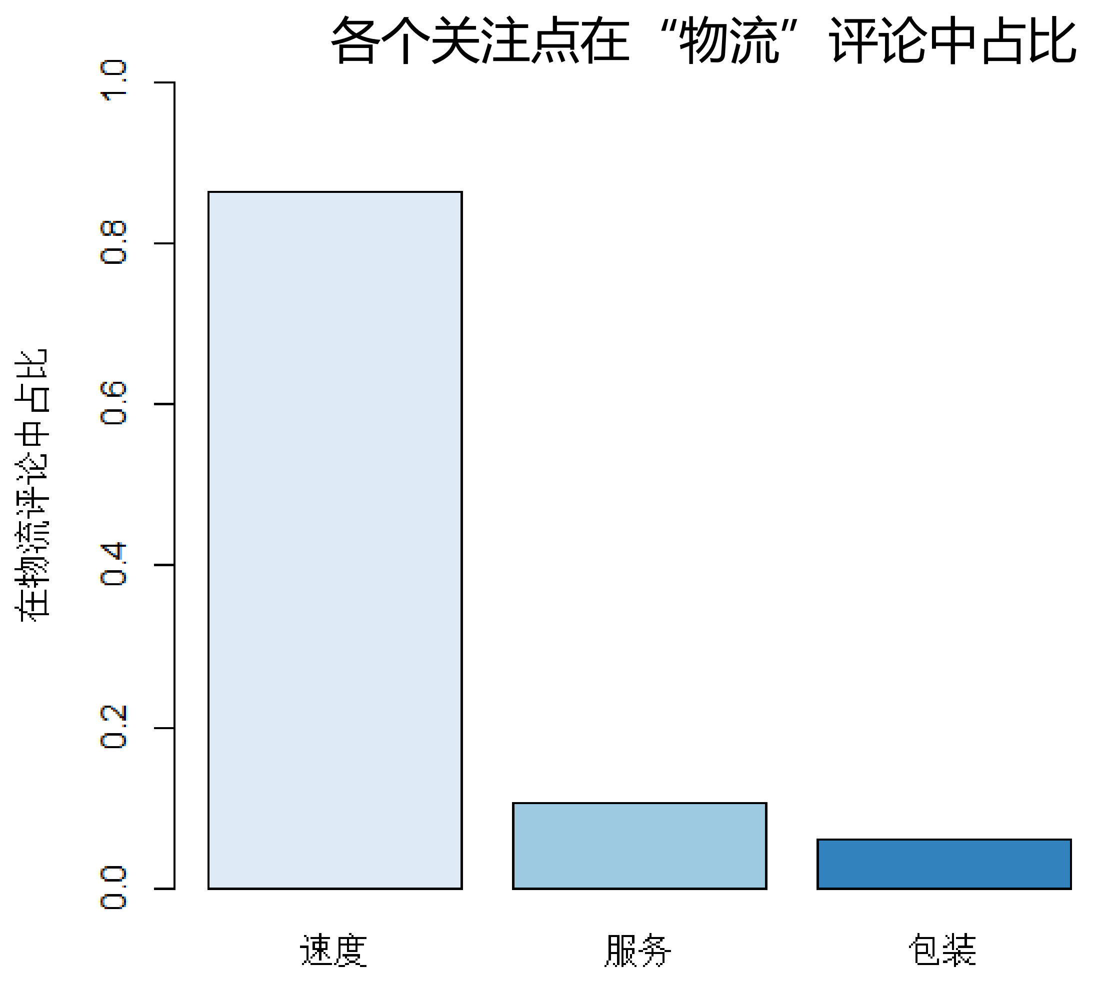
找出这些句子的高频词

归纳总结该热评词的主要关注点

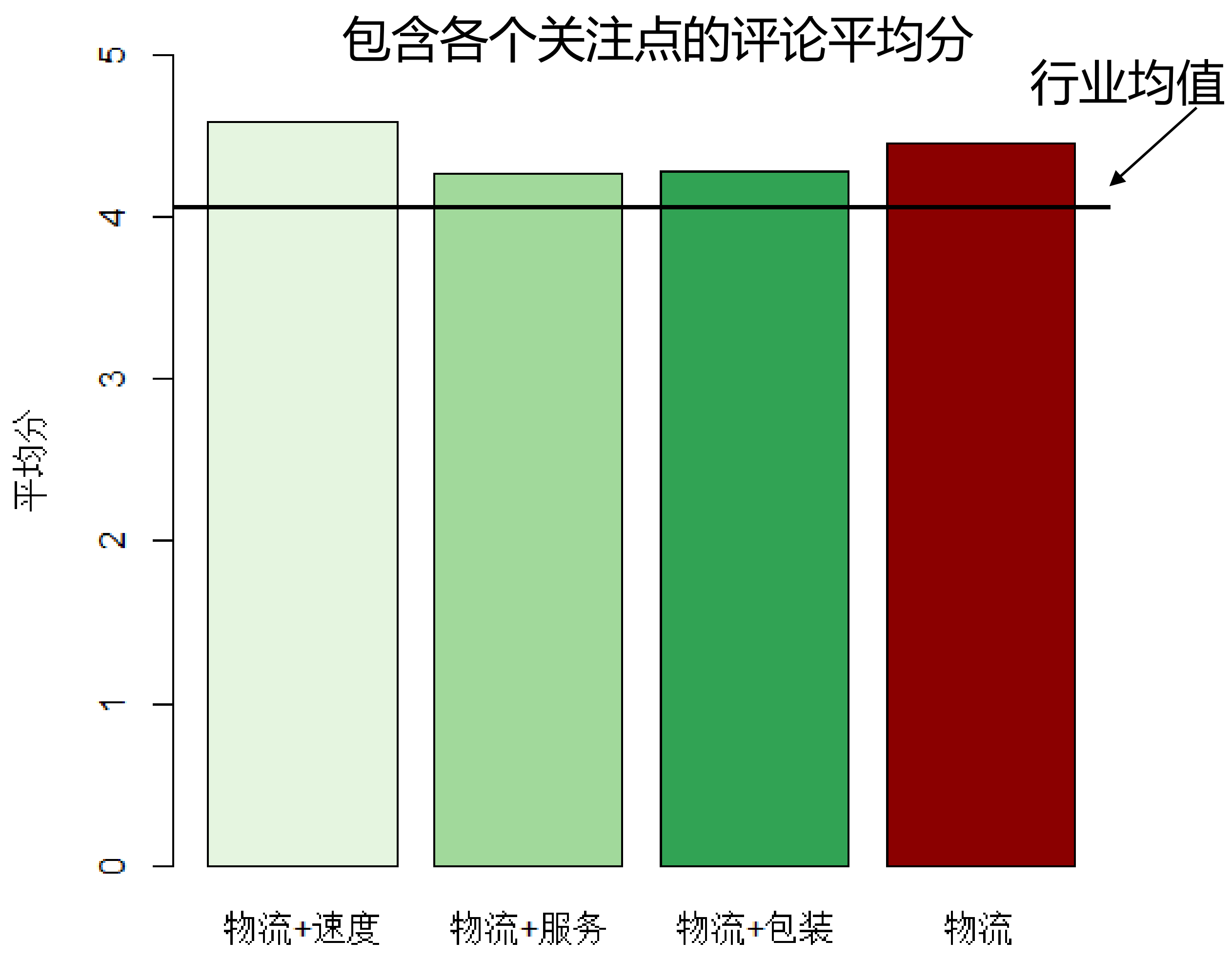
探索每个关注点的正负作用

和行业均值（所有手机评论的平均分）进行对比

物流

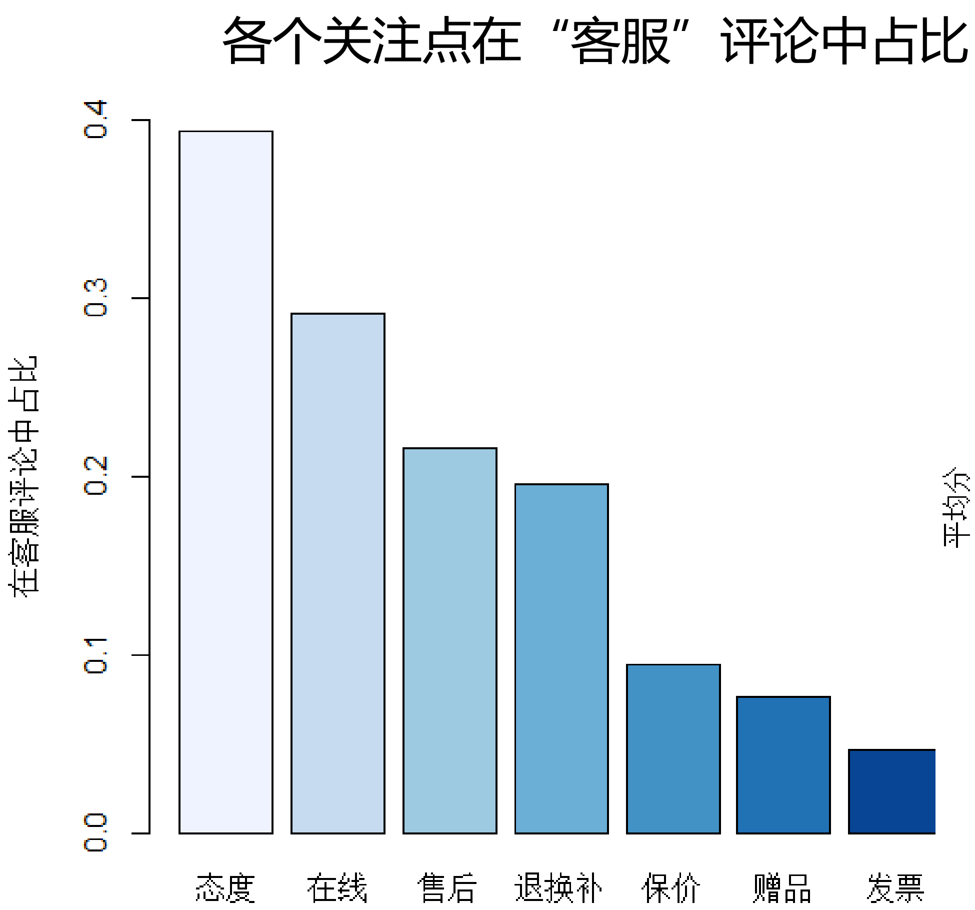


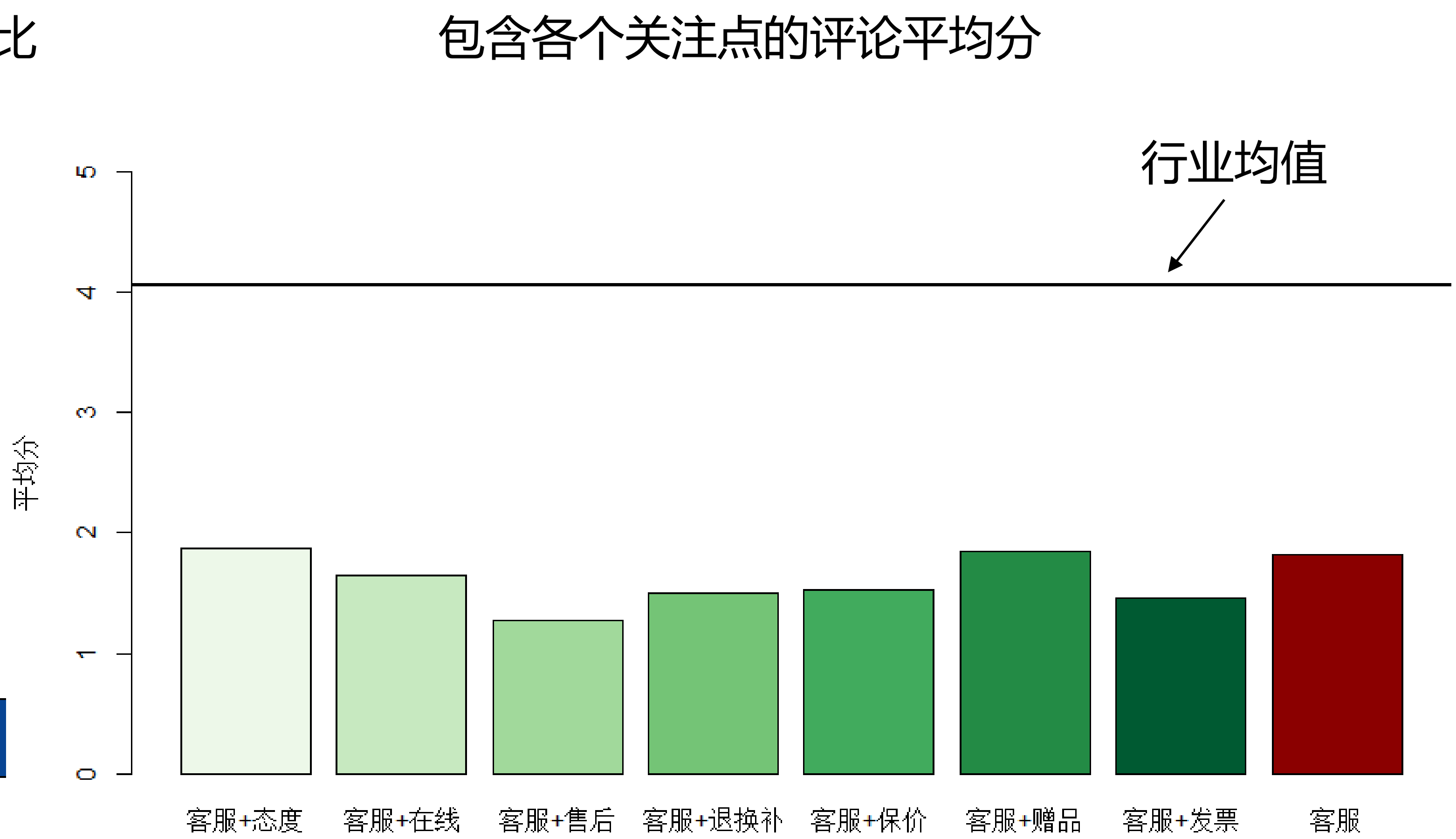
各个关注点在“物流”评论中占比
在物流评论中占比
0.0
0.2
0.4
0.6
0.8
1.0
速度
服务
包装



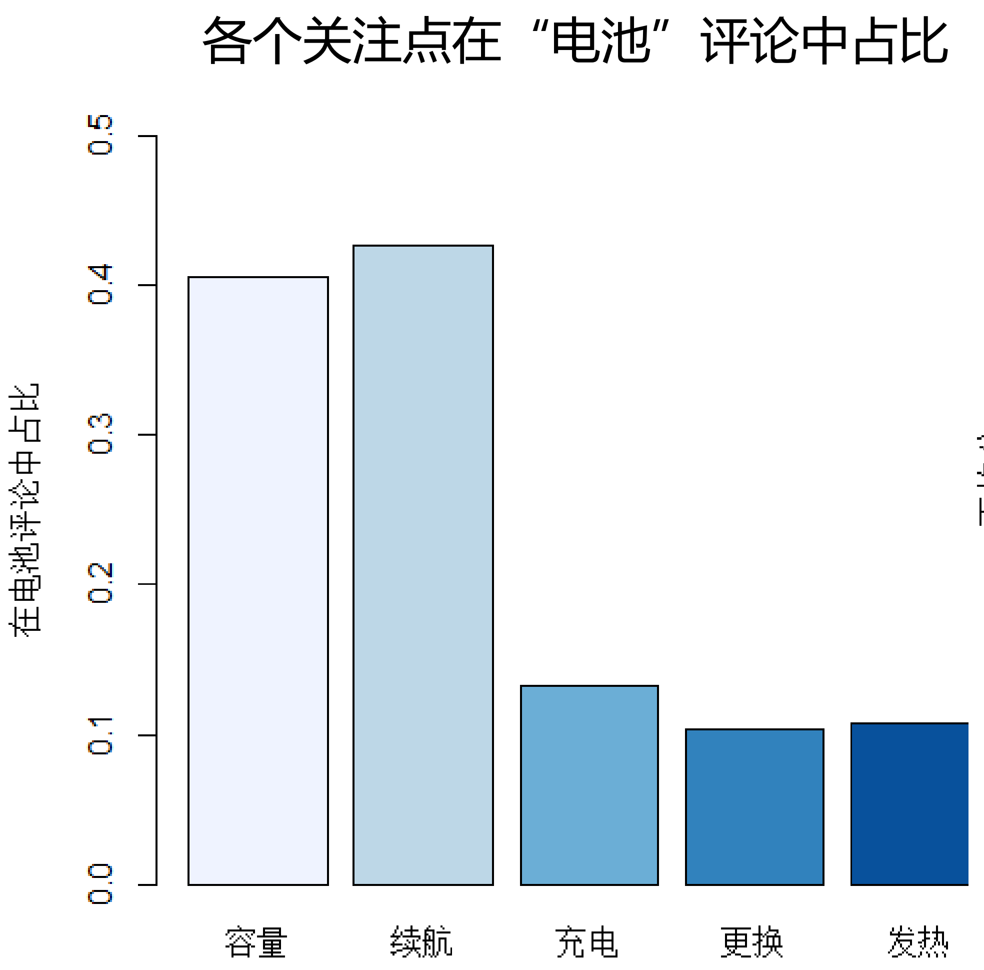
包含各个关注点的评论平均分
行业均值
平均分
0
1
2
3
4
5
物流+速度
物流+服务
物流+包装
物流

客服

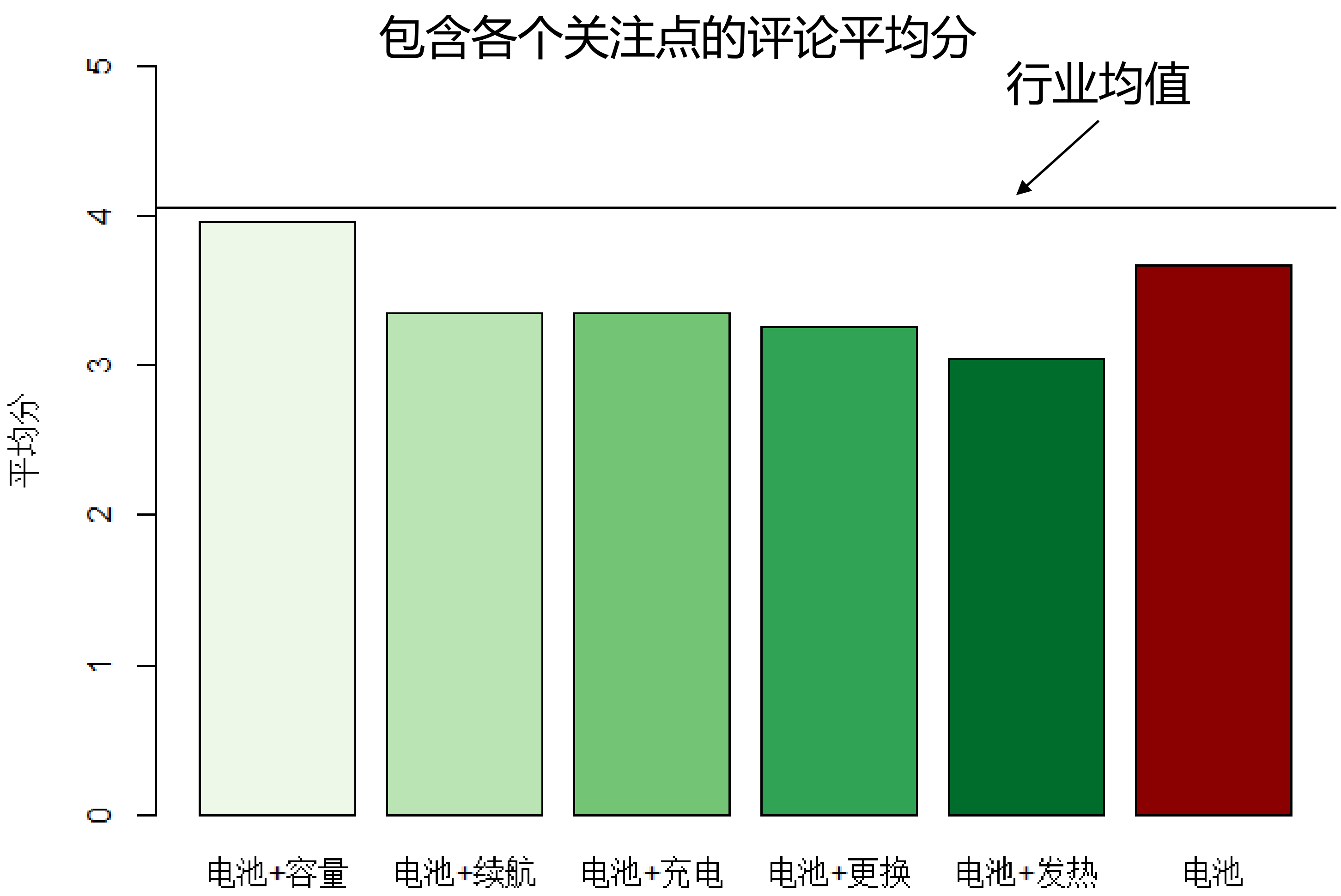

各个关注点在“客服”评论中占比
在客服评论中占比
0.0
0.1
0.2
0.3
0.4
态度
在线
售后
退换补
保价
赠品
发票


包含各个关注点的评论平均分
行业均值
平均分
0
1
2
3
4
5
客服+态度
客服+在线
客服+售后
客服+退换补
客服+保价
客服+赠品
客服+发票
客服

电池

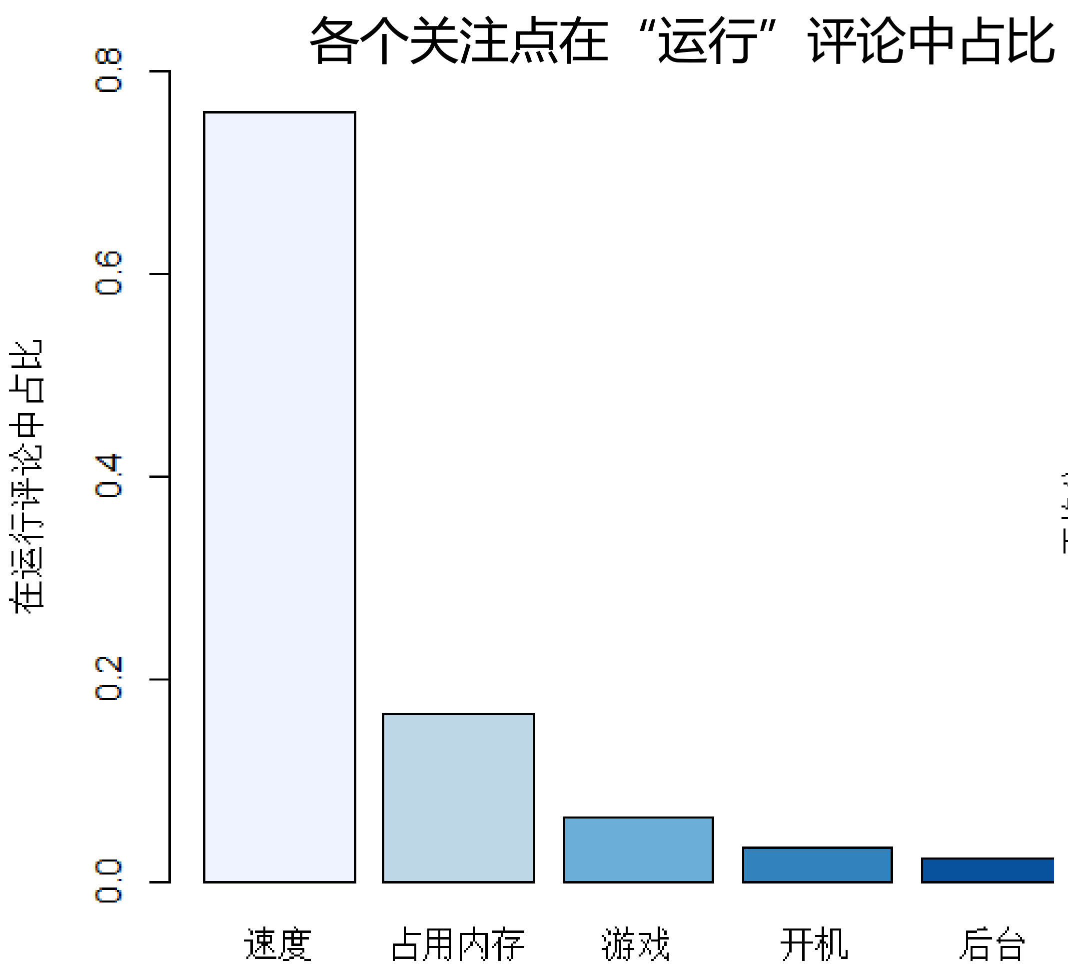


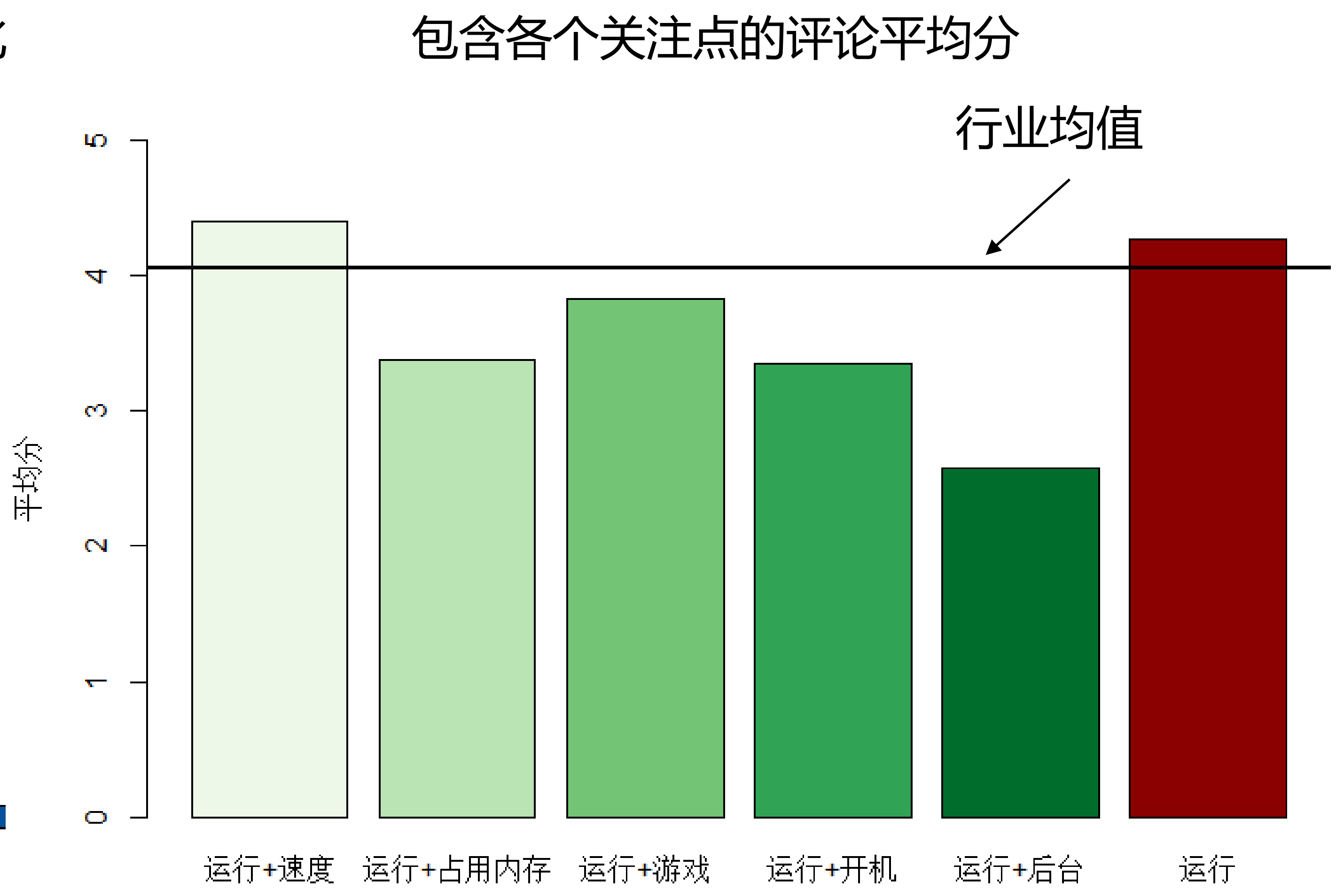
各个关注点在“电池”评论中占比
在电池评论中占比
0.0
0.1
0.2
0.3
0.4
0.5
容量
续航
充电
更换
发热



包含各个关注点的评论平均分
行业均值
平均分
0
1
2
3
4
5
电池+容量
电池+续航
电池+充电
电池+更换
电池+发热
电池

运行


各个关注点在“运行”评论中占比
在运行评论中占比
0.0
0.2
0.4
0.6
0.8
速度
占用内存
游戏
开机
后台


包含各个关注点的评论平均分
行业均值
平均分
0
1
2
3
4
5
运行+速度
运行+占用内存
运行+游戏
运行+开机
运行+后台
运行

05

商业应用

整体画像

- 计算每部手机在物流、客服、电池、运行四个方面的得分（即该手机包含这些热评词的评论平均得分）
- 与行业标准（所有手机包含这些热评词的评论平均分）进行对比

X手机

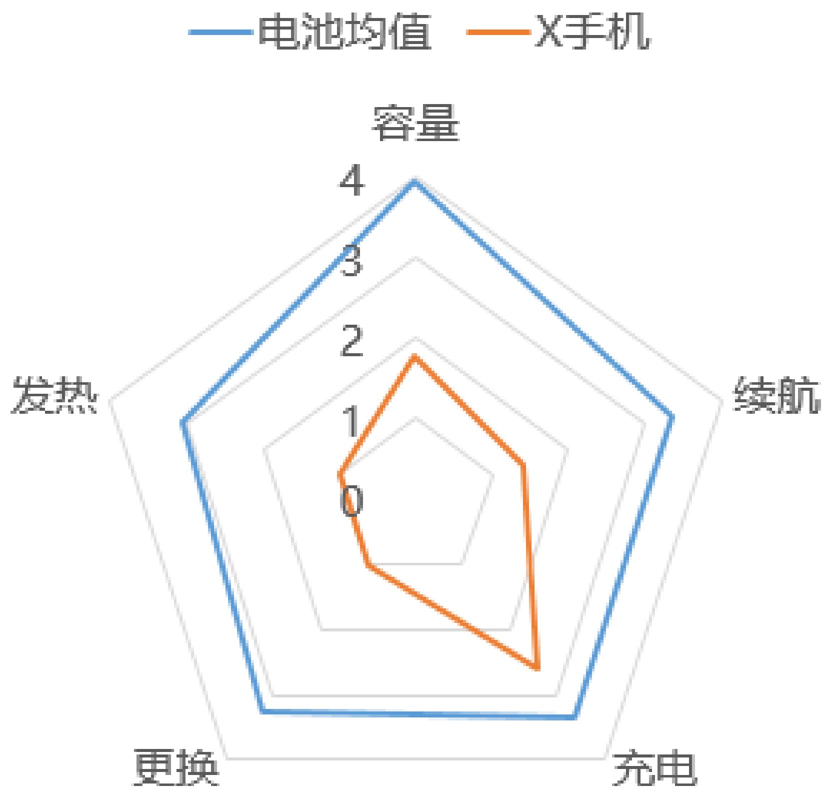
在**电池**和**客服**方面显著低于行业标准

N手机

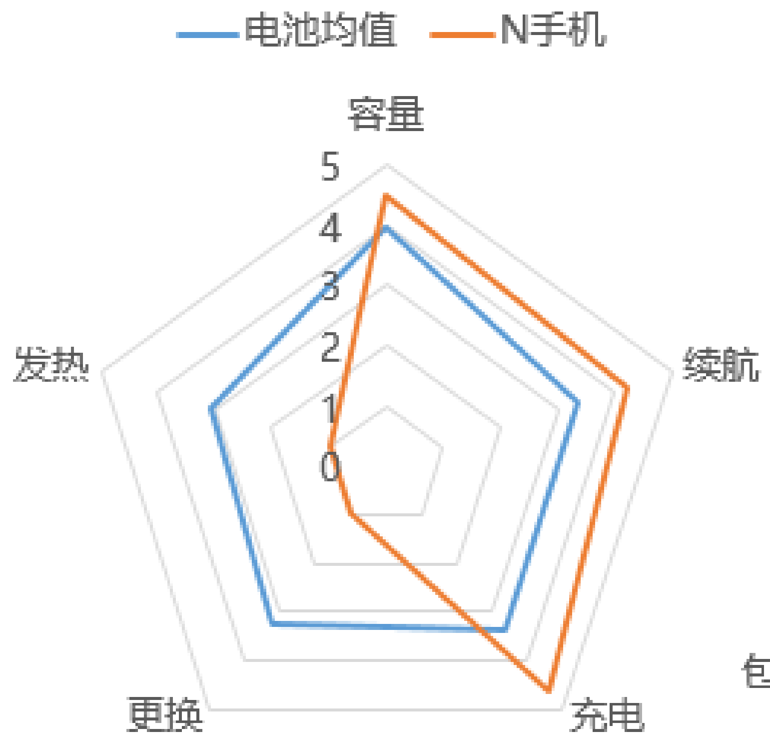
在**电池**方面显著高于行业标准，在**物流**和**运行**方面和行业标准持平

细节画像：X手机

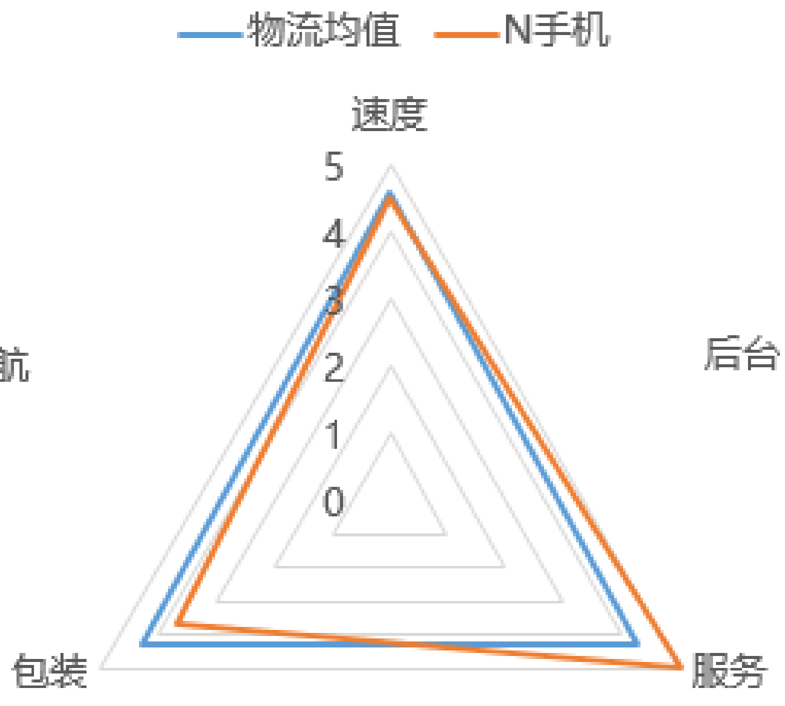
- 计算每部手机在各个关注点的得分（即该手机包含【热评词】+【关注点】的评论的平均得分
- 与该关注点的行业标准（所有手机包含【热评词】+【关注点】的评论的平均分）进行对比



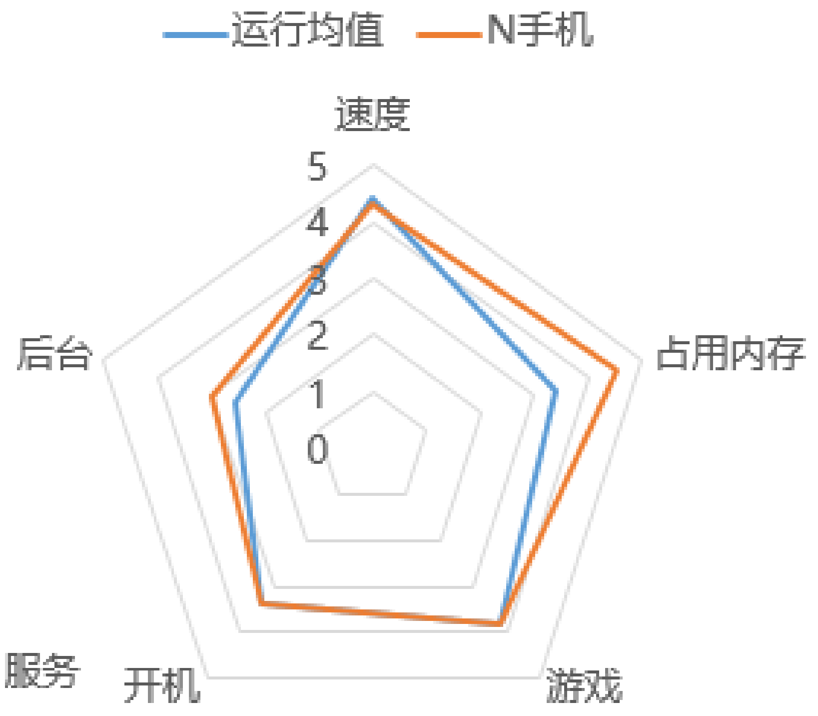
细节画像：N手机



电池均值
N手机
容量
续航
充电
更换
发热
5
4
3
2
1
0



物流均值
N手机
速度
服务
包装
5
4
3
2
1
0



运行均值
N手机
速度
占用内存
游戏
开机
后台
5
4
3
2
1
0

结论

1 用户评论是反映用户需求的一种直接表现形式。通过分析手机产品的用户评论，我们找到了四个影响手机好评率的关键点：物流、客服、电池和运行。

2 通过深挖评论内容，我们为每个关键点设计了一套具体的评价体系。该体系可以用来进行手机画像。

3 通过描绘每部手机的整体画像和细节画像，可以帮助产品快速查找不足，确定改进方向。

二、论文扩展

Sequential Text-Term Selection in Vector Space Models

Outline

1. Introduction
2. The Model
3. Simulation
4. Real Data Analysis
5. Conclusion

中國人民大學
RENMIN UNIVERSITY OF CHINA

1. Introduction

Intro to Text Mining

Text mining has wide applications and becomes increasingly important with the increasing accumulation of text documents in all fields.

Text V.S. Contiguous Responses

Our focus here is using text documents to explain ***continuous responses***.

好评率

99%

手机收到了，正品无假，朋友介绍在这家店买的，用了一段时间了，没有任何问题，性价比高，东哈哈

好评率

95%

物流是真心慢啊，等的花都谢了……手机已经激活了，现在还没发现什么问题，但数据线很硬是X的线都这样，懒得追究了，后续有什么问题再反应吧

好评率

87%

用了一个月，系统各种bug，微信视频通话声音完全不正常的小，而且找不到原因，上，充电中自动拨出四五个电话而且是在半夜里....！！！！

More Examples

欢迎#172girls官宣# @青春有你2-金子涵Aria @青春有你2-刘令姿 @青春有你2-曾可妮 @青春有你2-戴燕妮 加盟江苏卫视天猫#618超级晚#！ 超级晚，超…

江苏卫视 6月3日 17:07

21490 | 19628 | 632544

百度热榜 换一换

1	官方发布弗洛伊德最终尸检报告	481万
2	英国首相喊话特朗普反对种族主义	464万
3	民航局调整国际客运航班	448万
4	27地设摊贩规范点发展地摊经济	432万
5	美国将暂停所有中国客运航班	417万
6	黑人之死涉事4名警察被拘留	403万
7	钟美美模仿志愿者 新	388万
8	劳动成高中必修课	375万
9	男子4个月没回家床头长竹子	362万
10	董子健 恭喜我们都成功守住零点 新	349万

统计学理论精要及R代码实现

602 人开通

已购

数理统计 Mathematical...

630 人开通

已购

Vector Space Models

- Given that textual data are highly ***unstructured***, the first step in analyzing text documents is to make them ***structured***.
- A popular paradigm of structuralizing text documents is **vector space models** (Salton et al., 1975; Salton, 1989; Belew and Rijsbergen, 2000).

Unstructured Text

1. 手机收到了，是正品(*^▽^*)
2. 物流是真心很慢啊
3. 通常质量好差，各种bug

……

Structured Vectors

dict	手机	收到	正品	……	物流	通话	……
1	**1**	**1**	**1**	**……**	**0**	**0**	**……**
2	**1**	**0**	**0**	**……**	**1**	**0**	**……**
3	**0**	**0**	**0**	**……**	**0**	**1**	**……**
…				**……**			

Word or Phrase ?

The term used in vector space models can be both “***word***” and “***phrase***”.

- Word
  - the smallest element expressing semantic meaning
  - i.e. 电池, 发烫, 物流, 很快
- Phrase
  - a sequence of words, conveying an idiomatic meaning
  - i.e.电池+发烫, 物流+很快

Word or Phrase ?

- Using word as terms in VSM
  - Neglect order of words
  - Size of dictionary is relatively small

- Using phrase as terms in VSM
  - Take word order into consideration
  - Size of dictionary is much large

Phrase Better

Word order is critical since it influences the meaning of documents dramatically

我爱你 ≠ 你爱我

我不爱你 ≠ 你不爱我

Term Selection

- Due to the high dimensionality and sparse assumption of dictionary, ***term selection*** is necessary.

- Traditional selection methods in text mining (Ng et al., 1997; Yang and Pedersen, 1997; Sebastiani, 2002)
  - Information gain
  - Mutual information
  - Chi-square
  - ……

- Model-based selection / screening methods
  - Regularization methods (e.g. LASSO, Tibshirani, 1996)
  - Regularized inverse regression (Taddy, 2013)
  - SIS-based method (Fan and Lv, 2008; Fan and Song, 2010; Li et al., 2012; Liu et al., 2014; Liu et al., 2015)
  - ……

2. The Model: Sequential Term Selection

中國人民大學 RENMIN UNIVERSITY OF CHINA

Notations

- Let $W = \{w_1^*, w_2^*, \dots, w_d^*\}$ be a set (or dictionary) of d distinct words, such as W = {手机，电池，耐用，物流，很快，…}

- Let S be a ***term***, which is represented by a sequence of m words $\boldsymbol{S = < w_1 w_2 \dots w_m >}$. For example, S =< 电池 耐用 >. If all words in S appear in W, we say S is generated by W.

- If S_1 =< 电池 耐用 > and S_2 =< 物流 很快 >, then:
  (1) $S_3 = S_1 \cup S_2$ =< 电池 耐用 物流 很快 > be a new term
  (2) $S_1 \subset S_3$ and $S_2 \subset S_3$

The Model

- Let $C = \{S_1, S_2, \ldots, S_n\}$ denote the collection of documents. All terms (length <= q) construct a term dictionary T_q. Each document S_i is paired with a univariate and continuous response variable Y_i.

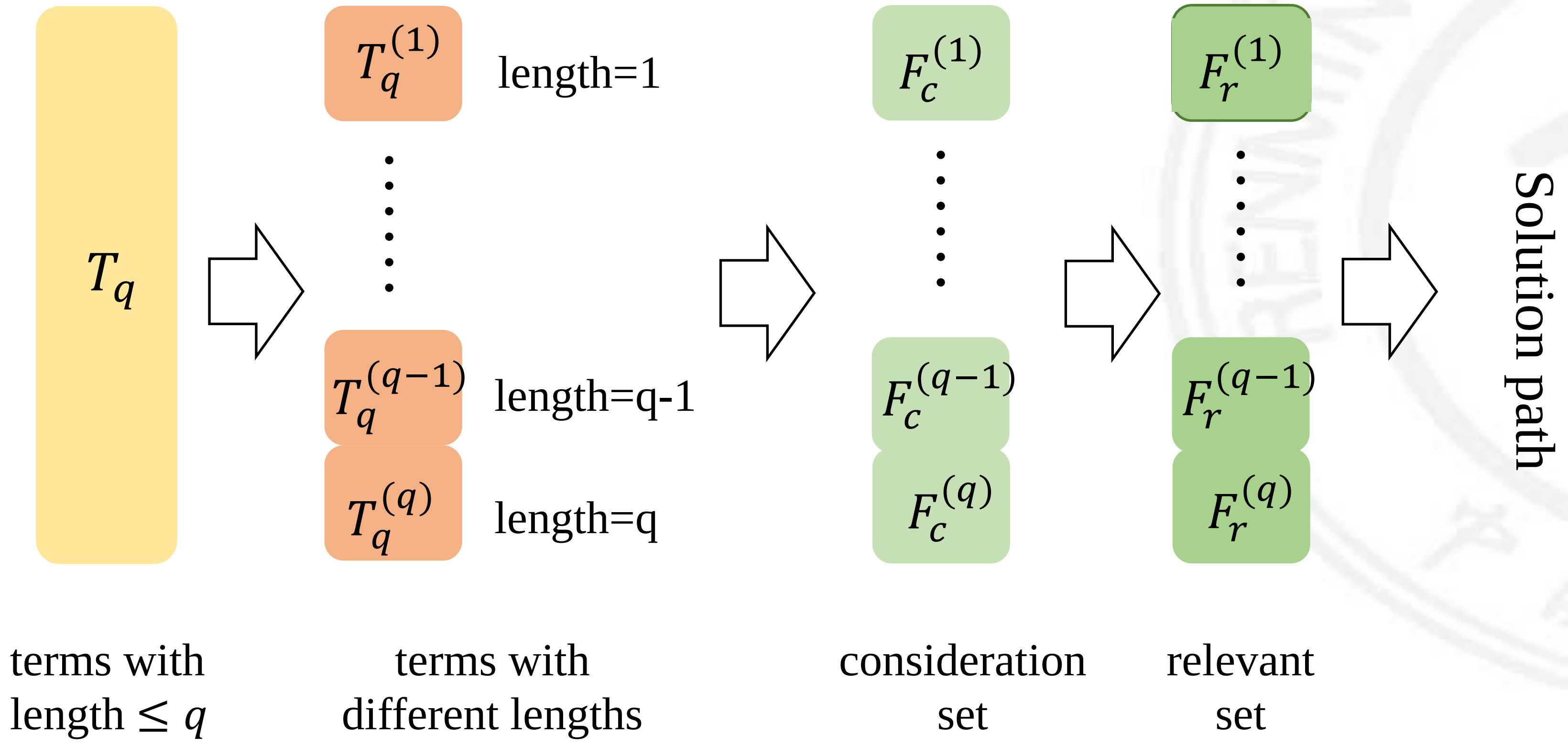
- To study the association between S_i and Y_i, we assume the following model

$$Y_i = \beta_0 + \sum_{1 \le j \le p} \beta_j \, I(S_j^* \subset S_i) + \varepsilon_i$$

where p is the size of T_q, and S_j^* is a term in T_q.

Sequential Term-Selection Method

- Split the whole term space T_q into sub-spaces according to term length



Sequential Term-Selection Method

length=1 $F_c^{(1)} = T_q^{(1)}$ $\xrightarrow{\text{SIS}}$ $F_r^{(1)}$

length=2 $F_c^{(2)} = \boldsymbol{F_r^{(1)}} \otimes F_r^{(1)}$ $\xrightarrow{\text{SIS}}$ $F_r^{(2)}$

⋮

length=m $F_c^{(\mathrm{m})} = \boldsymbol{F_r^{(1)}} \otimes F_r^{(m-1)}$ $\xrightarrow{\text{SIS}}$ $F_r^{(m)}$

Solution path

$F_r^{(i)}$ = {电池, 物流}
$F_r^{(j)}$ = {耐用, 很快 }
$F_r^{(i)} \otimes F_r^{(j)} =$
{
<电池 耐用> <电池 很快>
<物流 耐用> <物流 很快>
<耐用 电池> <耐用 物流>
<很快 电池> <很快 物流>
}

Sequential Term-Selection Method

- Summary of sequential term-selection method :
  - Step 1 (*Initialization*). Set $F_c^{(0)}$=$F_r^{(0)}$=$L^{(1)}$ and $F^{(0)}$=0
  - Step 2 (*Sequential Selection*). In the *m* sub-space,
    - 2.1 (*Consideration set*). Define ⊗ is the operator of right join

$$\mathcal{F}_c^{(m)} = \begin{cases} \mathcal{F}_r^{(0)} & \text{if } m = 1 \\ \mathcal{F}_r^{(1)} \otimes \mathcal{F}_r^{(1)} & \text{if } m = 2 \\ \{\mathcal{F}_r^{(m-1)} \otimes \mathcal{F}_r^{(1)}\} \cup \{\mathcal{F}_r^{(1)} \otimes \mathcal{F}_r^{(m-1)}\} & \text{if } 3 \le m \le q \end{cases}$$

    - 2.2 (*Term Selection*). Top *d* ones with the highest $\widehat{w}_j$ construct $F_r^{(m)}$

$$\hat{\omega}_j = \frac{(\mathbb{X}_{(j)} - \overline{\mathbb{X}}_{(j)})^\top (\mathbb{Y} - \overline{\mathbb{Y}})}{\sqrt{(\mathbb{X}_{(j)} - \overline{\mathbb{X}}_{(j)})^\top (\mathbb{X}_{(j)} - \overline{\mathbb{X}}_{(j)}) (\mathbb{Y} - \overline{\mathbb{Y}})^\top (\mathbb{Y} - \overline{\mathbb{Y}})}}$$

  - Step 3 (*Solution Path*). Iterate Step (2) for *q* times, which results in a total of *q* candidate models, $F^{(1)}, \dots, F^{(m)}$, with $F^{(m)} = \cup_{1 \le i \le m} F_r^{(i)}$

Together with Backward Method

- Remark 1: the choice of d (number of selected terms)
  - hard threshold rules: d=[n/log(n)] or (n-1)/q
  - data-based rules
- Remark 2: $F^{(m)}$ is not the final model.
  - It serves as a “quick-and-dirty” way to rule out unimportant ones
  - A backward elimination method and the extended BIC (Chen and Chen, 2008) is then applied to recover the final sparse model.

$$BIC(M) = \log\{\hat{\sigma}^2_{(M)}\} + n^{-1}|M|(\log n + 2\log|M|),$$

Theoretical property

- Under some usual conditions, the solution path is verified to be ***screening consistent***

(A1) Let ω_j be the correlation between the jth term indicator $\mathcal{I}(\mathcal{S}_j^* \subset \mathcal{S})$ and the response; then, for some $c_1 > 0$, $\kappa > 0$, $\min_{j \in \mathcal{F}_1} |\omega_j| \geq 2c_1 n^{-\kappa}$.

(A2) The random error ε follows subexponential tail probability condition: for some $s_0 > 0$ and all $s \in [0, s_0)$, we have $E\{\exp(s\varepsilon^2)\} < \infty$.

Theorem 1. Under conditions (A1) and (A2), the proposed algorithm is screening consistent, i.e.,

$$\Pr\left(\mathcal{F}_1 \subset \mathcal{F}^{(m)} \in \mathbb{F}, \text{ for some } 1 \leq m \leq q\right) \to 1.$$

3. Simulation

$2,\ldots,W*d\}$

Simulation Setting

- Data: consumer reviews for Cellphones in Jingdong.
- Let S_1 =< 电池 >, S_2 =< 物流 >, S_3 =< 耐用 >, S_4 =< 很快 >, S_5 =< 电池 耐用 >, S_6 =< 物流 很快 >. We consider the following three simulation settings,

Setting 1. $$Y_i = -1.5 + \mathcal{I}(s_1 \subset \mathcal{S}_i) + \mathcal{I}(s_2 \subset \mathcal{S}_i) + \varepsilon_i$$

Setting 2. $$Y_i = -2 + 1.5\mathcal{I}(s_5 \subset \mathcal{S}_i) + 1.5\mathcal{I}(s_6 \subset \mathcal{S}_i) + \varepsilon_i$$

Setting 3. $$Y_i = -1 + 0.5\mathcal{I}(s_1 \subset \mathcal{S}_i) + 0.5\mathcal{I}(s_3 \subset \mathcal{S}_i) + \mathcal{I}(s_5 \subset \mathcal{S}_i) + \varepsilon_i$$

Simulation Setting

- Competing methods:
  - information gain
  - mutual information
  - Chi-square
  - LASSO (Tibshirani, 1996)
  - Forward regression (Wang, 2007)
  - SIS (Fan and Lv, 2008)
  - DC-SIS (Li et al., 2012)

Evaluation Criteria

- Let F_1 be the true model, $\hat{F}_t$ be the selected model in the t-th simulation run. We consider the following four evaluation criteria.

$$\text{Coverage Probability} = T^{-1}\sum_{t=1}^{T}\mathcal{I}(\mathcal{F}_1 \subset \hat{\mathcal{F}}_{(t)}).$$

$$\text{Percentage of Correctly Fit} = T^{-1}\sum_{t=1}^{T}\mathcal{I}(\mathcal{F}_1 = \hat{\mathcal{F}}_{(t)}).$$

$$\text{Percentage of Correct Zeros} = \frac{1}{T(p-p_1)}\sum_{t=1}^{T}\sum_{j=1}^{p}\left\{I(\hat{\beta}_{j(t)}=0)\times I(\beta_j=0)\right\}$$

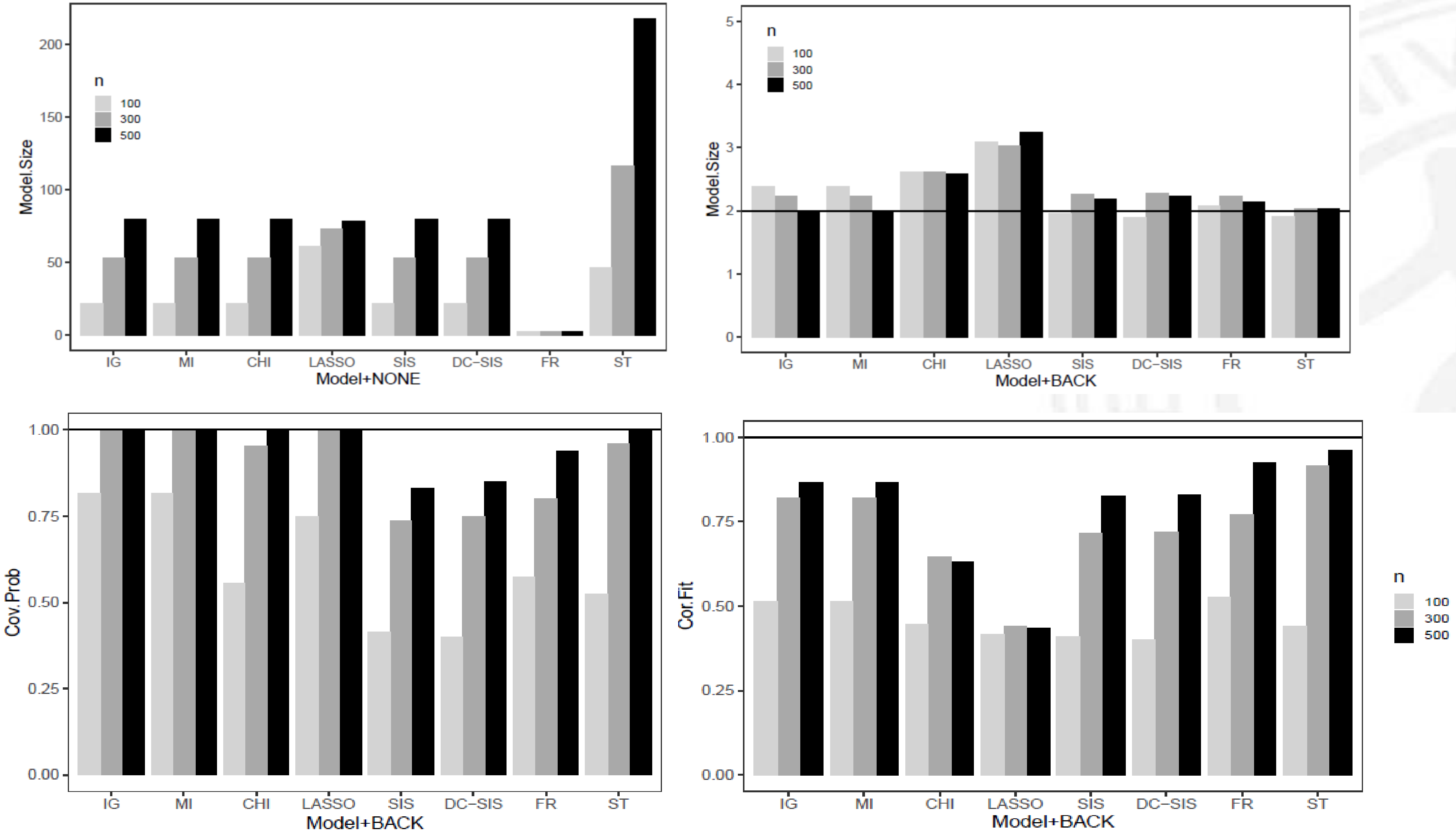
$$\text{Percentage of Incorrect Zeros} = \frac{1}{Tp_1}\sum_{t=1}^{T}\sum_{j=1}^{p}\left\{I(\hat{\beta}_{j(t)}=0)\times I(\beta_j\neq 0)\right\}$$

Simulation Results

Setting 1 (Model+Back)

n	Model (+BACK)	Model Size	Cov-Prob	Cor-Zeros	Incor-Zeros	Cor-Fit	Model Size	Cov-Prob	Cor-Zeros	Incor-Zeros	Cor-Fit	Model Size	Cov-Prob	Cor-Zeros	Incor-Zeros	Cor-Fit
		Theoretical $R^2 = 30\%$					Theoretical $R^2 = 50\%$					Theoretical $R^2 = 70\%$				
100	IG	1.7	60.5	100.0	28.3	59.0	2.4	91.5	100.0	5.3	71.5	2.2	96.5	100.0	1.8	80.0
	MI	1.7	60.5	100.0	28.3	59.0	2.4	91.5	100.0	5.3	71.5	2.2	96.5	100.0	1.8	80.0
	CHI	1.7	24.5	100.0	68.5	21.0	2.6	55.5	100.0	41.0	44.5	2.4	53.0	100.0	43.5	44.5
	LASSO	3.0	43.5	100.0	49.5	23.0	3.1	77.0	100.0	21.5	41.5	3.1	92.5	100.0	7.0	41.0
	SIS	1.5	17.0	100.0	66.5	16.5	2.0	41.5	100.0	42.5	41.0	2.1	56.5	100.0	29.0	56.0
	DC-SIS	1.5	19.0	100.0	65.0	19.0	1.9	40.0	100.0	45.0	40.0	2.1	58.0	100.0	25.0	56.0
	FR	1.7	42.0	100.0	57.0	39.5	2.1	57.5	100.0	42.5	52.5	2.6	58.0	100.0	42.0	55.0
	ST	1.6	30.0	100.0	38.8	22.0	1.9	52.5	100.0	26.3	44.0	1.9	60.5	100.0	22.0	56.5
300	IG	2.0	93.0	100.0	6.0	81.0	2.2	100.0	100.0	0.0	82.0	2.2	98.0	100.0	1.8	89.0
	MI	2.0	92.5	100.0	6.5	81.0	2.2	100.0	100.0	0.0	82.0	2.2	97.5	100.0	2.3	88.5
	CHI	2.2	75.0	100.0	22.5	58.5	2.6	95.5	100.0	4.3	64.5	2.4	88.5	100.0	10.3	65.0
	LASSO	3.1	100.0	100.0	0.0	43.5	3.0	100.0	100.0	0.0	44.0	3.0	100.0	100.0	0.0	41.0
	SIS	2.2	66.0	100.0	22.3	62.5	2.3	73.5	100.0	15.0	71.5	2.3	73.5	100.0	13.3	72.5
	DC-SIS	2.1	70.0	100.0	20.0	65.0	2.3	75.0	100.0	12.0	72.0	2.5	75.0	100.0	12.0	72.5
	FR	2.1	69.5	100.0	30.5	66.0	2.2	80.0	100.0	20.0	77.0	2.4	89.5	100.0	10.5	85.5
	ST	2.0	91.5	100.0	4.3	85.0	2.0	96.0	100.0	2.0	91.5	2.1	98.5	100.0	0.8	94.0
500	IG	2.1	96.5	100.0	2.5	86.0	2.0	100.0	100.0	0.0	86.5	2.0	100.0	100.0	0.0	91.5
	MI	2.1	96.5	100.0	2.5	86.0	2.0	100.0	100.0	0.0	86.5	2.0	100.0	100.0	0.0	91.5
	CHI	2.2	93.0	100.0	5.3	77.0	2.6	100.0	100.0	0.0	63.0	2.3	99.0	100.0	1.0	70.0
	LASSO	3.1	98.5	100.0	1.5	42.5	3.3	100.0	100.0	0.0	43.5	3.0	100.0	100.0	0.0	52.5
	SIS	2.1	80.5	100.0	11.8	78.5	2.2	83.0	100.0	9.0	82.5	2.2	84.5	100.0	8.0	84.0
	DC-SIS	2.0	85.0	100.0	7.5	82.0	2.2	85.0	100.0	8.0	83.0	2.2	87.0	100.0	6.0	85.0
	FR	2.1	82.5	100.0	17.5	81.0	2.1	94.0	100.0	6.0	92.5	2.2	95.5	100.0	4.3	94.0
	ST	2.0	99.0	100.0	0.5	95.5	2.0	100.0	100.0	0.0	96.0	2.0	100.0	100.0	0.0	97.0

Simulation Results



Simulation Results

Setting 2 (Model+Back)

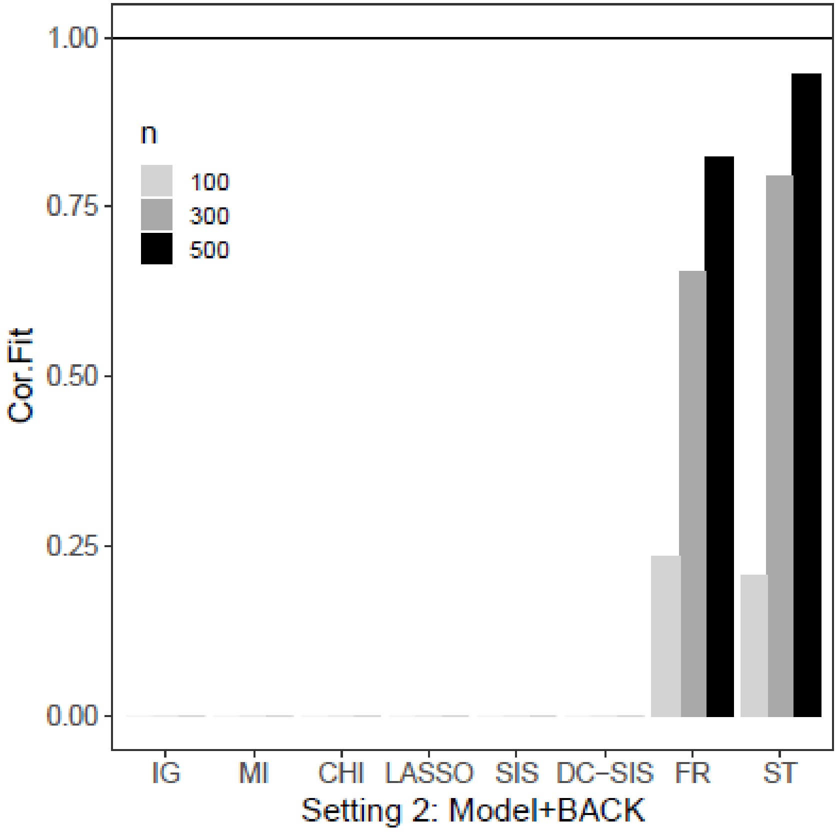
n	Model (+BACK)	Model Size	Cov-Prob	Cor-Zeros	Incor-Zeros	Cor-Fit	Model Size	Cov-Prob	Cor-Zeros	Incor-Zeros	Cor-Fit	Model Size	Cov-Prob	Cor-Zeros	Incor-Zeros	Cor-Fit
		Theoretical $R^2 = 30\%$					Theoretical $R^2 = 50\%$					Theoretical $R^2 = 70\%$				
100	IG	1.7	0.0	100.0	100.0	0.0	2.0	0.0	100.0	100.0	0.0	2.2	0.0	100.0	100.0	0.0
	MI	1.7	0.0	100.0	100.0	0.0	2.0	0.0	100.0	100.0	0.0	2.2	0.0	100.0	100.0	0.0
	CHI	1.8	0.0	100.0	100.0	0.0	2.1	0.0	100.0	100.0	0.0	2.4	0.0	100.0	100.0	0.0
	LASSO	3.1	0.0	100.0	100.0	0.0	3.3	0.0	100.0	100.0	0.0	3.5	0.0	100.0	100.0	0.0
	SIS	1.9	85.5	100.0	8.0	0.0	2.1	0.0	100.0	100.0	0.0	2.1	0.0	100.0	100.0	0.0
	DC-SIS	2.0	87.0	100.0	5.0	0.0	2.1	0.0	100.0	100.0	0.0	2.1	0.0	100.0	100.0	0.0
	FR	1.6	4.8	100.0	68.5	4.5	2.1	27.5	100.0	35.0	23.5	2.2	97.0	100.0	1.5	33.0
	ST	1.4	4.0	100.0	68.8	4.0	2.0	23.0	100.0	43.3	20.5	2.3	49.0	100.0	25.8	47.0
300	IG	2.1	0.0	100.0	100.0	0.0	2.2	0.0	100.0	100.0	0.0	3.4	0.0	100.0	100.0	0.0
	MI	2.1	0.0	100.0	100.0	0.0	2.2	0.0	100.0	100.0	0.0	3.4	0.0	100.0	100.0	0.0
	CHI	2.1	0.0	100.0	100.0	0.0	2.1	0.0	100.0	100.0	0.0	3.0	0.0	100.0	100.0	0.0
	LASSO	3.2	0.0	100.0	100.0	0.0	3.5	0.0	100.0	100.0	0.0	3.8	0.0	100.0	100.0	0.0
	SIS	2.1	0.0	100.0	100.0	0.0	2.1	0.0	100.0	100.0	0.0	2.1	0.0	100.0	100.0	0.0
	DC-SIS	2.0	0.0	100.0	100.0	0.0	2.2	0.0	100.0	100.0	0.0	2.1	0.0	100.0	100.0	0.0
	FR	2.0	52.0	100.0	24.0	51.0	2.4	98.0	100.0	1.0	65.5	2.1	83.0	100.0	8.3	75.0
	ST	2.1	59.0	100.0	20.8	52.0	2.2	86.0	100.0	7.0	79.5	2.1	97.0	100.0	1.5	94.0
500	IG	2.1	0.0	100.0	100.0	0.0	2.3	0.0	100.0	100.0	0.0	5.2	0.0	100.0	100.0	0.0
	MI	2.1	0.0	100.0	100.0	0.0	2.3	0.0	100.0	100.0	0.0	5.2	0.0	100.0	100.0	0.0
	CHI	2.2	0.0	100.0	100.0	0.0	2.2	0.0	100.0	100.0	0.0	5.4	0.0	100.0	100.0	0.0
	LASSO	3.4	0.0	100.0	100.0	0.0	3.8	0.0	100.0	100.0	0.0	4.3	0.0	100.0	100.0	0.0
	SIS	2.1	0.0	100.0	100.0	0.0	2.0	0.0	100.0	100.0	0.0	2.1	0.0	100.0	100.0	0.0
	DC-SIS	2.3	0.0	100.0	100.0	0.0	2.1	0.0	100.0	100.0	0.0	2.1	0.0	100.0	100.0	0.0
	FR	2.2	68.0	100.0	12.0	65.0	2.8	98.0	100.0	1.0	82.5	2.2	96.0	100.0	3.0	91.5
	ST	2.1	82.0	100.0	9.0	77.5	2.1	98.0	100.0	1.0	94.5	2.0	100.0	100.0	0.0	97.5

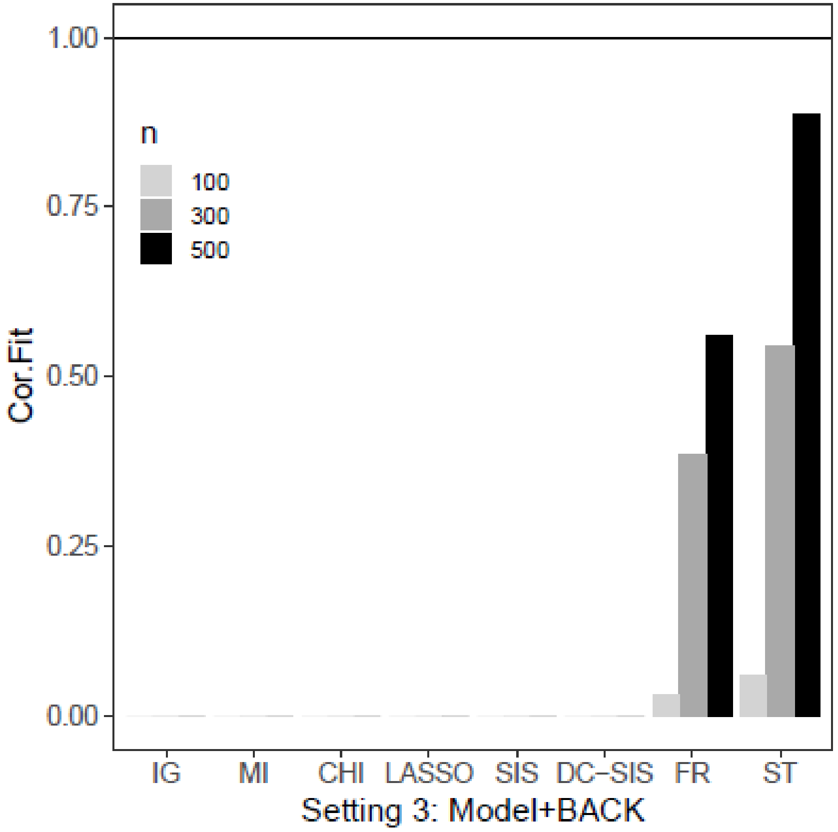
Simulation Results

Setting 3 (Model+Back)

n	Model (+BACK)	Model Size	Cov-Prob	Cor-Zeros	Incor-Zeros	Cor-Fit	Model Size	Cov-Prob	Cor-Zeros	Incor-Zeros	Cor-Fit	Model Size	Cov-Prob	Cor-Zeros	Incor-Zeros	Cor-Fit
		Theoretical $R^2 = 30\%$					Theoretical $R^2 = 50\%$					Theoretical $R^2 = 70\%$				
100	IG	1.3	0.0	100.0	79.2	0.0	1.6	0.0	100.0	77.0	0.0	2.5	0.0	100.0	60.7	0.0
	MI	1.3	0.0	100.0	79.2	0.0	1.6	0.0	100.0	77.0	0.0	2.5	0.0	100.0	60.7	0.0
	CHI	1.6	0.0	100.0	81.2	0.0	2.1	0.0	100.0	76.7	0.0	3.1	0.0	100.0	57.3	0.0
	LASSO	3.3	0.0	100.0	84.3	0.0	4.1	0.0	100.0	68.8	0.0	5.3	0.0	100.0	42.7	0.0
	SIS	1.5	0.0	100.0	85.7	0.0	1.9	0.0	100.0	76.0	0.0	2.9	0.0	100.0	44.3	0.0
	DC-SIS	1.6	0.0	100.0	82.0	0.0	1.9	0.0	100.0	70.0	0.0	2.9	0.0	100.0	41.2	0.0
	FR	1.2	0.0	100.0	67.5	0.0	1.5	3.5	100.0	58.3	3.0	2.8	47.5	100.0	29.2	33.5
	ST	1.4	4.0	100.0	68.3	1.5	1.7	8.0	100.0	54.0	6.0	2.7	63.0	100.0	15.5	56.5
300	IG	1.8	0.0	100.0	80.0	0.0	2.8	0.0	100.0	61.7	0.0	4.3	0.0	100.0	36.0	0.0
	MI	1.8	0.0	100.0	80.0	0.0	2.8	0.0	100.0	61.7	0.0	4.3	0.0	100.0	36.0	0.0
	CHI	1.9	0.0	100.0	83.0	0.0	3.2	0.0	100.0	64.5	0.0	4.8	0.0	100.0	36.7	0.0
	LASSO	3.7	0.0	100.0	62.8	0.0	4.9	0.0	100.0	39.2	0.0	5.5	0.0	100.0	33.3	0.0
	SIS	1.9	0.0	100.0	74.5	0.0	3.0	0.0	100.0	42.8	0.0	3.6	0.0	100.0	33.3	0.0
	DC-SIS	1.9	0.0	100.0	71.0	0.0	2.9	0.0	100.0	41.0	0.0	3.4	0.0	100.0	30.0	0.0
	FR	1.6	6.0	100.0	59.2	6.0	2.7	47.5	100.0	28.2	38.5	3.8	89.5	100.0	3.5	52.5
	ST	1.6	4.5	100.0	51.8	4.5	2.6	60.5	100.0	16.0	54.5	3.0	100.0	100.0	0.0	96.0
500	IG	2.2	0.0	100.0	76.3	0.0	3.6	0.0	100.0	44.7	0.0	5.2	0.0	100.0	33.3	0.0
	MI	2.2	0.0	100.0	76.3	0.0	3.6	0.0	100.0	44.7	0.0	5.2	0.0	100.0	33.3	0.0
	CHI	2.3	0.0	100.0	75.5	0.0	3.9	0.0	100.0	43.3	0.0	5.1	0.0	100.0	33.7	0.0
	LASSO	4.6	0.0	100.0	44.3	0.0	5.4	0.0	100.0	33.5	0.0	5.9	0.0	100.0	33.3	0.0
	SIS	2.4	0.0	100.0	93.5	0.0	3.3	0.0	100.0	33.8	0.0	3.8	0.0	100.0	33.3	0.0
	DC-SIS	2.3	0.0	100.0	91.0	0.0	3.1	0.0	100.0	31.0	0.0	3.8	0.0	100.0	29.0	0.0
	FR	2.0	18.0	100.0	46.7	15.0	3.3	79.5	100.0	8.8	56.0	4.1	90.0	100.0	3.3	83.0
	ST	2.1	26.0	100.0	33.7	23.5	3.0	92.5	100.0	2.5	88.5	3.0	99.5	100.0	0.2	97.0

Simulation Results





4. Real Data Analysis

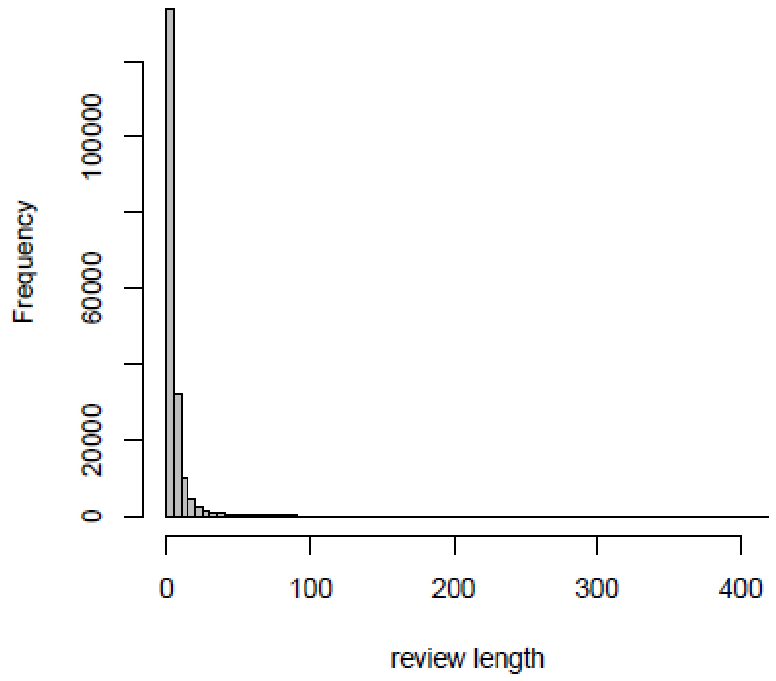
Data Description

- We collected 188,107 consumer reviews for 297 cellphones on Jingdong (www.JD.com).

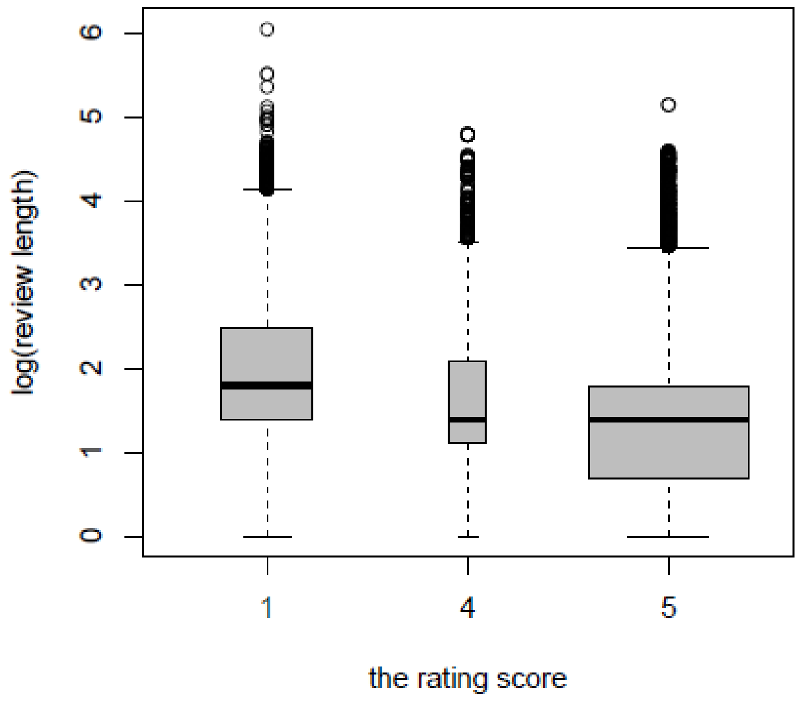
- Dependent variable: rating scores

- The user-generated reviews describe the true feelings of consumers with products and services. Therefore, the objective of this study is to discover factors that can influence consumers' evaluations on cellphones.

Descriptive Analysis



The histogram of review length



The boxplot of review length (in logarithm) under different rating scores

Descriptive Analysis

Top 100 words with highest frequencies in good reviews

Top 100 words with highest frequencies in bad reviews

Sequential Term Selection

	Model Size NONE	Model Size BACK	Selected terms
q=1	100	9	<垃圾>,<差>,<退>,<坏>,<死机>,<不错>,<假货>,<欺骗>,<清晰>
q=2	171	12	<垃圾>,<差>,<维修>,<客服>,<不错>,<假货>,<卡>,<清晰>,<做工>,<检测 坏>,<物流 很快>,<正品>
q=3	240	10	<垃圾>,<退货>,<坏>,<不错>,<主板>,<假货>,<欺骗>,<换 电池>,`<屏幕 划痕 换货>,<物流 不错 满意>

Sequential Term Selection

	Estimate	Std. Error	t.value	p.value	
(Intercept)	0.00	0.02	0.00	1.000	
垃圾	-0.30	0.04	-7.97	0.000	***
退货	-0.22	0.03	-7.19	0.000	***
坏	-0.20	0.03	-6.46	0.000	***
不错	0.15	0.03	5.59	0.000	***
主板	-0.07	0.03	-2.66	0.008	**
假货	-0.13	0.03	-4.90	0.000	***
欺骗	-0.22	0.02	-8.62	0.000	***
<电池 换>	-0.12	0.03	-4.41	0.000	***
<屏幕 划痕 换货>	-0.10	0.02	-4.20	0.000	***
<不错 物流 满意>	0.13	0.02	5.54	0.265	***

5. Conclusion

Conclusion

- We explore the association between text documents and some continuous response variable.

- Given that text documents are highly unstructured, vector space models are commonly used to structuralize the textual data.

- We propose a novel term selection method to address the high-dimensional problem.

- Results of simulations as well as real data analysis show that the sequential term selection method can select the relevant terms by a few steps.

END