Sentiment Analysis for Online Reviews with Regularized Text Logistic Regression

Peng Liu[1]
Ying Chen[1]
Chung Piaw Teo[2]

[1]Department of Statistics & Applied Probability
[2]Department of Analytics & Operations

National University of Singapore

Introduction

- Increasing user-generated reviews and feedback on product and service on the web.
- It becomes essential for executives and managers to collect and analyze text information to capture comprehensive understanding on market demand and expectation.
- Direct analysis of the unstructured text is challenging and time consuming given the ultra-high dimension and complex dependence.
- Build an efficient classifier to identify a (small) set of essential features so as to provide rapid and valuable suggestions for sentiment analysis and operational improvement.

Text Mining

Text mining is a process of deriving insights or patterns from unstructured text information through statistical learning.

- Transform free form text into structured form to facilitate further analysis, which helps select important word features.
- **Bag-of-words (BoW)** representation of text assumes the distribution of words within each document is sufficient, and linguistic features like order and grammar can be safely ignored for sentiment analysis.

Bag of Words

Corpus (a set of documents to be analyzed) is transformed into a document-term matrix, which contains a column for each word appearing in the corpus and a row for each document.

- Usually each matrix entry is the count of the number of times a particular word appears in each document, which is referred to as frequency statistic.
- Alternatively, the tf-idf statistic, short for term frequency - inverse document frequency, can also be applied to indicate the importance of a word in the corpus.

Tf-idf

Let $x = \{x_1, x_2, ..., x_{p_n}\}$ be the word features, $d = \{d_1, d_2, ...d_n\}$ the set of documents, and n_{ij} the tf-idf value of feature x_j in d_i.

The tf-idf is calculated as the product of two statistics, term frequency (TF) and inverse document frequency (IDF).

$$\text{tf-idf} = \text{TF} \times \text{IDF}$$

where TF(x_j) = (# of x_j in doc d_i) / (# of all features in doc d_i), IDF(x) = log_e(# of all docs / # of docs with feature x_j).

The word importance increases proportionally to the number of times it appears in the document but is offset by its frequency in the corpus.

Sentiment Analysis

- After choosing proper corpus representation and feature statistic, sentiment analysis often involves sentiment classification, feature selection, topic modelling, and emotion detection, etc.
- Various model-based machine learning classifiers (logistic regression, KNN, SVM) have been applied in predicting review polarity with good performance, see e.g. Pang (2002).
- A comprehensive survey of different algorithms and applications of sentiment analysis in text mining field can be found in Medhat (2014).

Logistic Regression as Classifier

Objective: accurately classify the polarity of a review as positive or negative given its textual content.

Logistic regression:

- popular discriminative classifier that models the conditional probability of the response taking a particular value;
- usually used as a baseline model in sentiment classification task, see e.g. Francis (2006).

Challenges

- Curse of dimensionality: Direct application of logistic regression would be inefficient or even infeasible when the number of covariates is large and exceeds the number of observations (large dim with $p_n >> n$).
  It suffers from the numeric problems that result in lack of convergence, overfitting, and poor predictive accuracy see Greenland (2000), Hadjicostas (2003).
- Besides building a good statistical classifier, another key challenge is identifying a set of significant and meaningful features that can be used for further inference.

Sparsity and regularization

Regularization procedures are used to perform variables selection in high dimensional regression under sparsity assumption.

- Lasso (Tibshirani, 1996), Adaptive Lasso (Zou, 2006), Elastic Net (Zou and Hastie, 2005), and Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001) for univariate selection;
- Group Lasso and group LAR (Yuan and Lin, 2006), $\ell 1/\ell q$ regularization (Zhao et al., 2008; Liu and Zhang, 2008), group SCAD (Wang et al. 2007; Huang et al. 2012) for group structures selection.
- Sparse group Lasso (Friedman et al. 2010; Zhou et al., 2010) for bi-level selection. Multivariate sparse group Lasso (Obozinski et al., 2011; Li et al., 2015) for multiresponse regression problems.

Objective

- Regularized Text Logistic (RTL) regression for ultra-high dimensional **text reviews** and derive asymptotic properties under sparsity.

$$
\begin{aligned}
Q_n(\beta_n) &= -\frac{1}{n}\Sigma_{i=1}^{n}\{y_i log\pi_i + (1-y_i) log_(1-\pi_i)\} + \Sigma_{j=1}^{p_n} p_\lambda(|\beta_j|) \\
\text{where } \pi_i &= \frac{e^{\eta_i}}{1+e^{\eta_i}}, \eta_i = \beta^\top x_i
\end{aligned}
$$

- Identify a small set of features amenable for operational improvement.
- Perform sentiment classification and compare with existing ML classifiers.

Outline

1. Introduction ✓
2. Data and corpus pre-processing
3. Regularized Text Logistic Regression Model
4. Real Data Analysis and Discussion
5. Conclusion

Data

We consider two datasets for restaurants and hotels respectively.

- **Restaurant**: 1899 TripAdvisor (English) reviews on major restaurants in a Singapore hotel from 1^{st} December, 2015 till 9^{th} September, 2016 via a scraper we developed and collect attributes such as review title, content, date and rating of each review on the webpage.
- **Hotel**: TripAdvisor reviews on hotels in different cities in the US as used in Wang (2011). We selected the first 30 (.dat) files from the original dataset, resulting in 2519 reviews from 28^{th} Mar, 2001 to 27^{th} Jan, 2009 with attributes of date, customer satisfaction rating score (1-5), title and full text.

In both datasets, a review is classified as positive if the rating is 4-5, and negative if rating is 1-2, with the rest discarded.

Sample reviews of Restaurant Dataset

+ We sat outside overlooking the promenade, superb food, wines and service. Try the traditional fish and chips and lamb chops! Lucky to get a table as we hadn't made a booking. Lovely complimentary breads and with my hungry adult males they brought out more while waiting on our mains! Nice touch! (posted on 8 Dec 2015)

− Came here on Sunday night with friends with high expectations because of the famous chef behind the brand, I could not be more disappointed in terms of food or service, I suppose coming from Europe I am use to good French bistro with good standard but here it's fine dining with Michelin star price ouch! (posted on 21 Mar 2016)

Sample reviews of Hotel Dataset

+ Quality hotel at great price. Very clean. Free breakfast with good selections. Staff friendly and most helpful. A great stay! (posted on 25 Nov 2008)

– Don't do it!! This place is run down, dirty and loud. The pictures they provide on the web do not tell the story so don't be fooled. (posted on 6 Mar 2008)

Corpus Pre-processing

- Using BoW representation, the reviews are transformed into a document-term matrix, where each entry represents the tf-idf statistic of a word feature in a document.
- Pre-processing is performed to retain meaningful words and filter out uninformative ones, since direct transformation usually results in a big matrix that brings challenges in both memory and computation.
- Pre-processing steps include: (1) transform into lowercase, (2) remove punctuation, (3) remove stop words (e.g. the, a), (4) strip while space, and (5) stem words to remove suffixes, i.e. services, server, and served are all replaced by serv.

Corpus Pre-processing

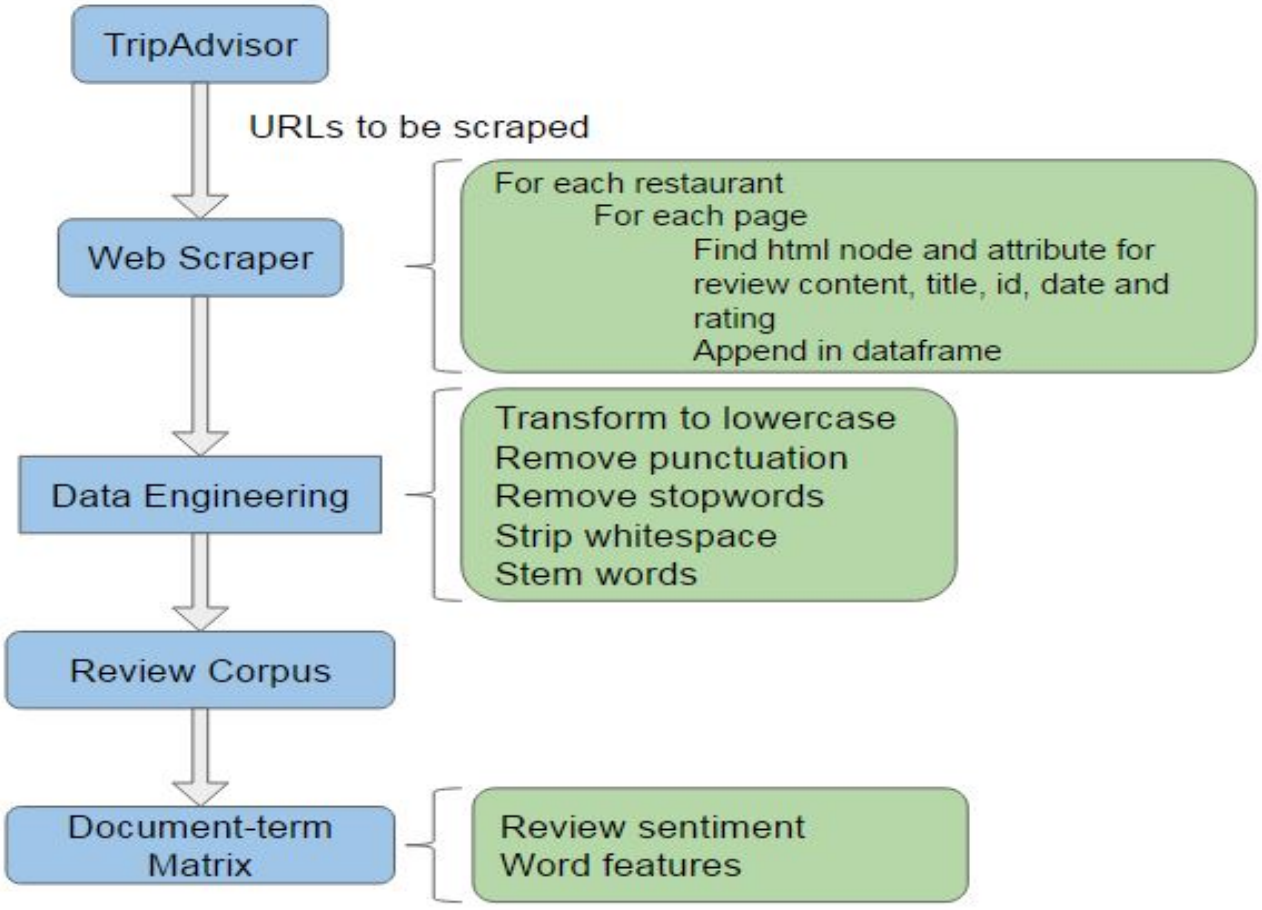

Figure 1: Flowchart of Data Collection and Processing

Statistics

Table 1: Summary statistics of both datasets

Dataset	Region	Positive	Negative	Total
Restaurant	Singapore	1525	130	**1655**
Hotel	US	1558	593	**2151**

After processing, there are 5 543 and 11 324 different features for Restaurant and Hotel datasets respectively.

Regularized Text Logistic Regression

- The response (review polarity) is a random variable $Y_i \sim B(n_i, \pi_i)$ with the binomial denominator $n_i = 1$ and probability π for all i.
- In the resulting document-term matrix, let $x = \{x_1, x_2, ..., x_{p_n}\}$ be the word features with tf-idf statistic as the observation, $d = \{d_1, d_2, ...d_n\}$ be the set of documents.
- The dependence is represented via logic transformation,

$$\pi_i = \frac{exp\{\beta^\top x_i\}}{1 + exp\{\beta^\top x_i\}}$$

where $\beta = (\beta_1^\top, \beta_2^\top)^\top$ is the coefficient vector and assumed to be sparse, i.e. only a subset of features $\beta_1 \in R^{k_n}$ have significant impact in determining review polarity and $\beta_2 \in R^{m_n}$ are insignificant, $k_n + m_n = p_n$.

Loss function

The estimation is achieved by maximizing the regularized log-likelihood function

$$Q_n(\beta_n) = -\frac{1}{n}\Sigma_{i=1}^{n}\{y_i log\pi_i + (1-y_i)log_(1-\pi_i)\} + \Sigma_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|) \quad (1)$$

where $p_\lambda(\beta)$ is some penalty function, and λ is tuning parameter.

Fan and Li (2001) proposed smoothly clipped absolute deviation (SCAD) penalty that, through simultaneous variable selection and coefficient estimation, reaches oracle properties (the estimator is able to perform as well as if the correct set of significant features were known in advance).

Smoothly Clipped Absolute Deviation

The SCAD penalty is defined as follows:

$$p_{\lambda,\gamma}(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{if} \quad |\beta_j| \leq \lambda; \\ -\left(\frac{|\beta_j|^2 - 2\gamma\lambda|\beta_j| + \lambda^2}{2(\gamma-1)}\right) & \text{if} \quad \lambda < |\beta_j| \leq \gamma\lambda; \\ \frac{(\gamma+1)\lambda^2}{2} & \text{if} \quad |\beta_j| > \gamma\lambda \end{cases}$$

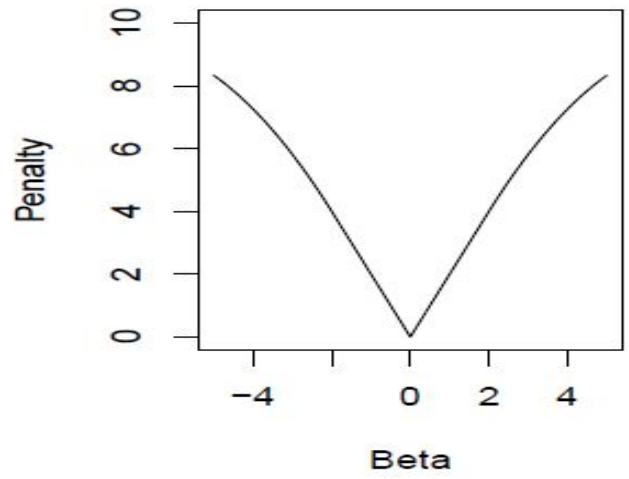

Figure 2: SCAD penalty function

SCAD

- SCAD corresponds to a quadratic spline function with knots at λ and $\gamma\lambda$.
- The first derivative for some $\gamma > 2$ and $\beta > 0$ is $p\prime_{\lambda,\gamma}(\beta) = \lambda\{I(\beta \leq \lambda) + \frac{(\gamma\lambda-\beta)_+}{(\gamma-1)\lambda} I(\beta > \lambda)\}$
- The SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$ but singular at 0 with its derivatives being 0 outside the range $[-\gamma\lambda, \gamma\lambda]$.
- This leads to small coefficients being set to 0, a few other coefficients being shrunk towards 0, while retaining large coefficients as they are, thus producing sparse solution and approximately unbiased estimates for large coefficients.

SCAD

The solution to SCAD penalty is

$$\hat{\beta}_j^{SCAD} = \begin{cases} (|\beta_j| - \lambda)_+ sign(\hat{\beta}_j) & if\, |\hat{\beta}_j| < 2\lambda; \\ \{(\gamma - 1)\hat{\beta}_j - sign(\hat{\beta}_j)\gamma\lambda\}/(\gamma - 2) & if\, 2\lambda < |\hat{\beta}_j| \leq \gamma\lambda; \\ \hat{\beta}_j & if \quad |\hat{\beta}_j| > \gamma\lambda \end{cases}$$

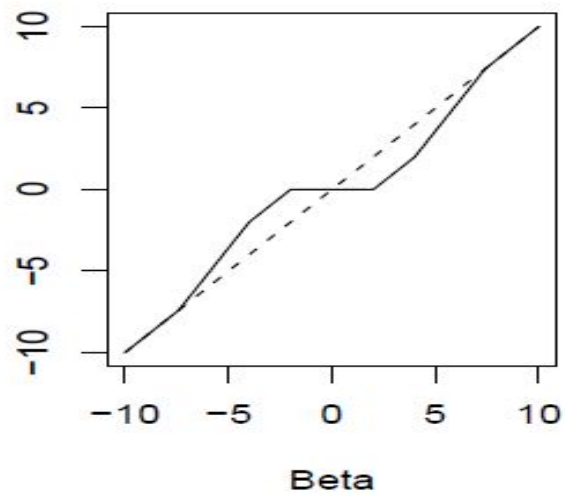

Figure 3: SCAD thresholding rule

Parameter Tuning

- The regularized logistic regression model depends on a few hyper-parameters, including the number of folds K in cross validation, the tuning parameter γ and penalty coefficient λ that jointly control the trade-off between model bias and variance.
- λ is selected by minimizing the cross-validated prediction error along the coefficient path.
- The cross-validated prediction error is defined as $CV = \frac{1}{N}\sum_{i=1}^{N} Q_n(\hat{\beta}^{-k(i)})$, where $\hat{\beta}^{-k}(x)$ is the estimated coefficients with the k^{th} observation removed for $k = 1, 2, ..., K$.

Parameter Tuning

- Having three hyper parameters that jointly control the trade-off between model bias and variance, including the number of folds k in cross validation, the tuning parameter γ and penalty coefficient λ, we apply a data-adaptive approach to allow k, γ and λ vary jointly along a 3D grid of different values in search of the minimal cross-validation error, with k ranging from 5 to 20 and γ from 2.1 to 4 with a step size of 0.1.
- k, γ and λ will vary jointly along a 3D grid of different values in search of the minimal cross-validation error.

Asymptotic Properties

We study the sampling properties of the proposed penalized likelihood estimator with diverging dimensionality and discuss the regularity required conditions to obtain global minimizer.

- Let $\beta_0 = (\beta_{10}^\top, \beta_{20}^\top)^\top$ be the true set of coefficients to be estimated, where β_{10} is a k_nx1 vector of nonzero coefficients and β_{20} is an m_nx1 vector of zero coefficients where $k_n + m_n = p_n$, the number of parameters which is allowed to grow slowly as sample size goes to infinity.
- Similarly, let $\beta_n = (\beta_{1n}^\top, \beta_{2n}^\top)^\top$ be the observed coefficients for the above mentioned cost function $Q_n(\beta_n)$, with the corresponding global minimizer $\tilde{\beta}_n$ and proposed RTL estimator $\hat{\beta}_n$.

Regularity Conditions

Denote $a_n = max_{1 \le j \le p_n}\{p\prime_{\lambda_n}(|\beta_{0j}|), \beta_{0j} \neq 0\}$ and $b_n = max_{1 \le j \le p_n}\{p\prime\prime_{\lambda_n}(|\beta_{0j}|), \beta_{0j} \neq 0\}$.

- (C.1) $\|\tilde{\beta}_n - \beta_0\|^2 = O_p(p_n/n)$.
- (C.2) $\liminf_{n\to\infty}\liminf_{\beta_n \to 0^+} p\prime_{\lambda_n}(|\beta_n|)/\lambda_n > 0$.
- (C.3) $a_n = O(n^{-1/2})$ as $n \to \infty$.
- (C.4) $b_n \to 0$ as $n \to \infty$.
- (C.5) $\lambda_n \to 0$ and $\sqrt{p_n/n}/\lambda_n \to 0$.
- (C.6) There exist constants C_3 and C_4 such that when $\beta_1, \beta_2 > C_3\lambda_n$, $|p\prime\prime_{\lambda_n}(\beta_1) - p\prime\prime_{\lambda_n}(\beta_2)| \le C_2|\beta_1 - \beta_2|$.

Regularity Conditions

- (C.7) For Fisher information matrix $I_n(\beta_n) = E[\{\frac{\partial Q_n(\beta_n)}{\partial \beta_n}\}\{\frac{\partial Q_n(\beta_n)}{\partial \beta_n}\}^\top]$, there exist constants C_1 and C_2 satisfying $0 < C_1 < \lambda_{min}\{I_n(\beta_n)\} \leq \lambda_{max}\{I_n(\beta_n)\} < C_2 < \infty$ where $\lambda_{min}\{I_n(\beta_n)\}$ and $\lambda_{max}\{I_n(\beta_n)\}$ are minimal and maximal eigenvalues of $I_n(\beta_n)$ respectively.
  And for $j, k = 1, 2, ..., p_n$,

$$E_{\beta_n}\left\{\frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}}\frac{\partial Q_n(\beta_n)}{\partial \beta_{nk}}\right\}^2 < C_3 < \infty$$

  and

$$E_{\beta_n}\left\{\frac{\partial Q_n^2(\beta_n)}{\partial \beta_{nj}\partial \beta_{nk}}\right\}^2 \leq C_4 < \infty$$

Regularity Conditions

- (C.8) There exists a large enough open subset w_n of $\Omega_n \in R^{p_n}$ which contains the true parameter point β_n, such that for almost all V_{ni} the density admits all third derivatives $\partial Q_n(V_{ni}, \beta_n)/\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{nl}$ for all $\beta_n \in \omega_n$. Furthermore, there are functions M_{njkl} such that

$$\left|\frac{\partial Q_n(V_{ni}, \beta_n)}{\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{nl}}\right| \leq M_{njkl}(V_{ni})$$

  for all $\beta_n \in \omega_n$, and

$$E_{\beta_n}\{M^2_{njkl}(V_{ni})\} < C_5 < \infty$$

- (C.9) Let $\beta_{01}, \beta_{02}, ..., \beta_{0k_n}$ be nonzero and $\beta_{0k_{n+1}}, \beta_{0k_{n+2}}, ..., \beta_{0p_n}$ be zero. Then $\beta_{01}, \beta_{02}, ..., \beta_{0k_n}$ satisfy

$$\min_{1\leq j\leq k_n} |\beta_{0j}|/\lambda_n \to \infty \qquad \text{as} \quad n \to \infty$$

Regularity Conditions

- Condition (C.1) is imposed in logistic regression setting with diverging dimensionality as in (He and Shao, 2000).
- Condition (C.2) ensures sparse solution by making $p_{\lambda_n}(\beta_n)$ singular at the origin.
- Condition (C.3) guarantees unbiasedness property of large parameters, and condition (C.4) ensures $p_{\lambda_n}(\beta_n)$ does not have much more influence than $l_n(\beta_n)$ on SCAD estimator.
- Condition (C.5), (C.8) and (C.9) is used in the proof of Oracle property, and condition (C.6) is a smoothness condition on $p_{\lambda_n}(\beta_n)$.
- Condition (C.7) assumes $l_n(\beta_n)$ to be positive definite with uniformly bounded eigenvalues.

Consistency

Theorem (Consistency)

Suppose Conditions (C.1), (C.3), (C.4) and (C.7) hold. Then there exists a local minimizer $\hat{\beta}_n = (\hat{\beta}_{1n}^\top, \hat{\beta}_{2n}^\top)^\top$ *of* $Q_n(\beta_n)$ *such that*

$$\|\hat{\beta}_n - \beta_0\| = O_p(\sqrt{p_n/n})$$

This theorem shows that under some conditions, there exists a root-(n/p_n)-consistent estimator.

Sparsity

Lemma (Sparsity)

Suppose (C.1)-(C.7) hold. Then the SCAD estimator $\hat{\beta}_n$ satisfies

$$Pr(\hat{\beta}_{2n} = 0) \to 1$$

In addition, denote

$$\Sigma_{\lambda_n} = \text{diag}\{p''_{\lambda_n}(|\beta_{01}|), ..., p''_{\lambda_n}(|\beta_{0k_n}|)\}$$

and

$$b_n = \{p'_{\lambda_n}(|\beta_{01}|)\text{sgn}(\beta_{01}), ..., p'_{\lambda_n}(|\beta_{0k_n}|)\text{sgn}(\beta_{0k_n})\}^\top$$

where diag$\{\cdot\}$ is a diagonal matrix and sng$(\cdot)$ is a sign function.

Oracle Property

Theorem (Oracle property)

Suppose $p_n^2/n \to 0$, *and the regularity conditions (C.1)-(C.7) are satisfied, then the local minimizer* $\hat{\beta}_n = (\hat{\beta}_{1n}^\top, \hat{\beta}_{2n}^\top)^\top$ *of* $Q_n(\beta_n)$ *in the theorem on consistency satisfies*
(1) Sparsity: $Pr(\hat{\beta}_{2n} = 0) \to 1$ *as* $n \to \infty$.
(2) Asymptotic normality:

$$\sqrt{n}\alpha_n^\top I_n^{-1/2}(\beta_{10})(I_n(\beta_{10}) + \Sigma_{\lambda_n})[(\hat{\beta}_{1n} - \beta_{10}) + (I_n(\beta_{10}) + \Sigma_{\lambda_n})^{-1} b_n] \xrightarrow{D} N(0, 1)$$

where α_n *is an arbitrary* k_n*x1 vector such that* $\|\alpha_n\| = 1$.

Oracle Property

This theorem ensures model sparsity and asymptotic normality when the number of parameters diverges. When n is large enough, it holds that $\Sigma_{\lambda_n} = 0$ and $b_n = 0$ for SCAD penalty, and the asymptotic normality becomes

$$\sqrt{n}\alpha_n^\top I_n^{1/2}(\beta_{10})(\hat{\beta}_{1n} - \beta_{10}) \xrightarrow{D} N(0, 1)$$

which is as efficient as the maximum likelihood estimator of β_{10} if β_{20} were known in advance.

Global Property

Theorem (Global Property)

Under conditions (C.1), (C.5) and (C.7), with probability tending to 1, the local minimizer $\hat{\beta}_n$ is the global minimizer of (1), i.e.

$$Pr(Q_n(\hat{\beta}_n) = inf_{\beta_n} Q_n(\beta_n)) \to 1$$

Sentiment Classification

We implement RTL regression on the Restaurant and Hotel datasets.

- Investigate sentiment classification performance of the proposed RTL approach.
- Compare with alternative sentiment classification approaches.
- Identify a set of meaningful features amenable for operational improvement.

Performance Measure

For the binary sentiment classification problem, there are 4 outcomes as displayed in a two-way contingency table,

Table 2: Four performance measures

	Actual positive	Actual negative
Predicted positive	a	b
Predicted negative	c	d

- To measure the performance of sentiment classifier, we consider 6 commonly used metrics.

Performance Measure

Table 3: Six metrics to measure model performance

Metric	Definition	Interpretation
True positive rate (TPR)	$\frac{a}{a+c}$	Also called recall or sensitivity, measures the proportion of positive reviews that are correctly identified, or the extent to which true positives are not missed.
True negative rate (TNR)	$\frac{d}{b+d}$	Also called specificity, measures the proportion of negatives that are correctly identified.
Positive predictive value (PPV)	$\frac{a}{a+b}$	Also called precision, measures the proportion of positive reviews that are true positive.

Performance Measure

Table 4: Six metrics to measure model performance

Metric	Definition	Interpretation
Negative predictive value (NPV)	$\frac{d}{c+d}$	Measures the proportion of negative reviews that are true negative.
Accuracy	$\frac{a+d}{a+b+c+d}$	Measures the proportion of reviews that are correctly classified.
F1 Score	$\frac{2*PPV*TPR}{PPV+TPR}$	Combines precision and recall as an overall metric

Alternative Models

We consider 4 alternative models: (unregularized) logistic regression, Naive Bayes, K-Nearest Neighbor and Support Vector Machine.

- Following the same data-adaptive approach, the hyper parameters are selected similarly via cross validation from a grid of values, for example the cost is allowed to vary between 10^{-1} and 10^{2} and γ among 0.5, 1 and 2 in SVM.
- For unregularized logistic regression, maximum likelihood estimation would face the problem of singularity when the feature space gets larger than the number of documents in the corpus, thus a range of thresholds from 80% to 99% is applied to control the sparsity of the document-term matrix so as to downsize the feature space.

Model Selection

- In SCAD regularized logistic regression, the minimal average cross validation error is obtained by searching over a 3D grid, resulting in $K = 6$, $\gamma = 3.8$ and $\lambda = 0.0243$ for Restaurant Dataset, and $K = 18$, $\gamma = 4$ and $\lambda = 0.0146$ for Hotel Dataset.
- The coefficient path obtained from optimal hyper-parameters suggests that a small value of λ would result in too many features, including many insignificant ones, while a big value of λ would pose heavy penalty on the coefficients and end up with large bias in model selection.
- Through this procedure, 115 out of 5543 (2%) features are selected for Restaurant Dataset and 357 out of 11324 (3%) for Hotel Dataset.

In-sample Performance

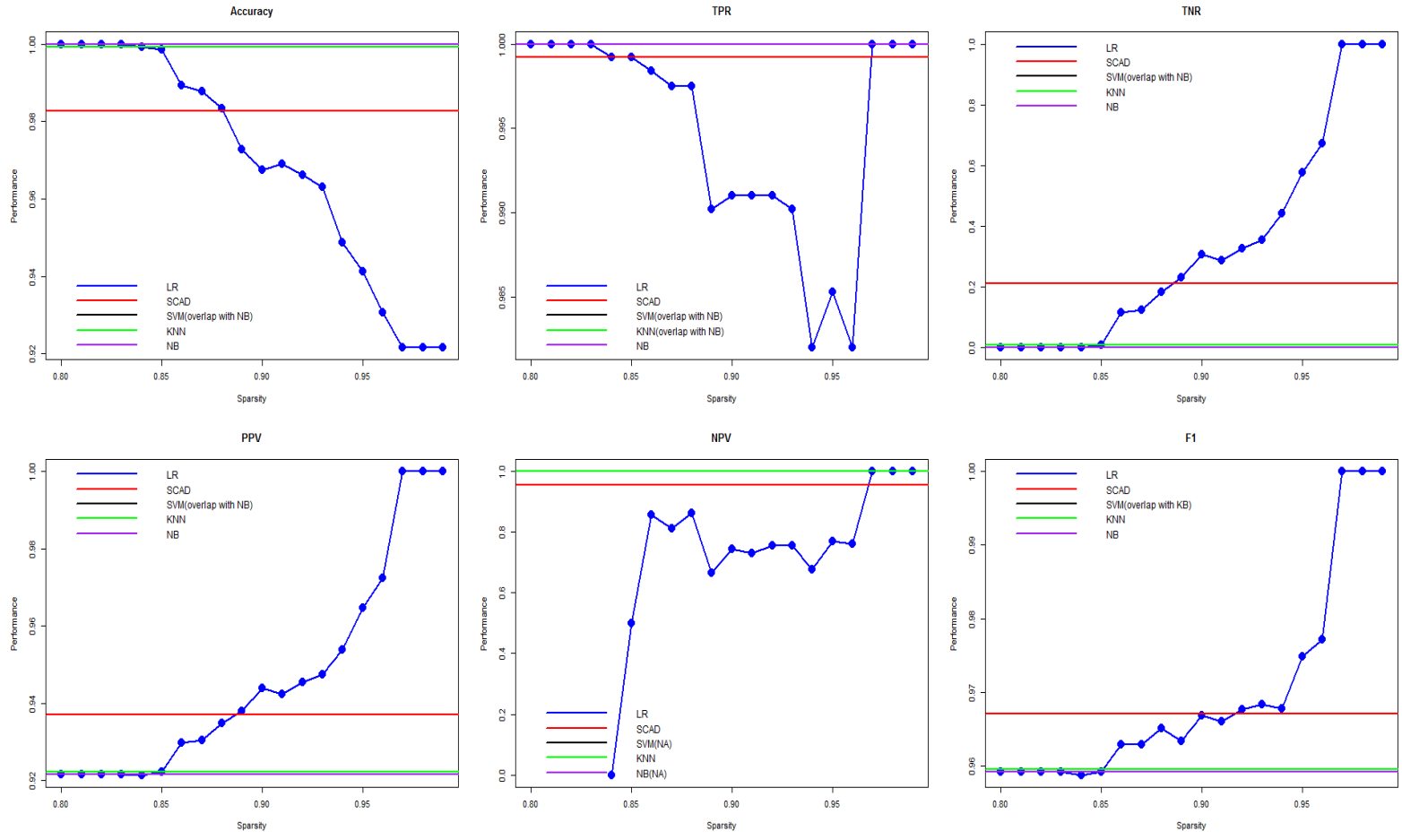

Figure 4: In-sample performance for Restaurant Dataset

In-sample Performance

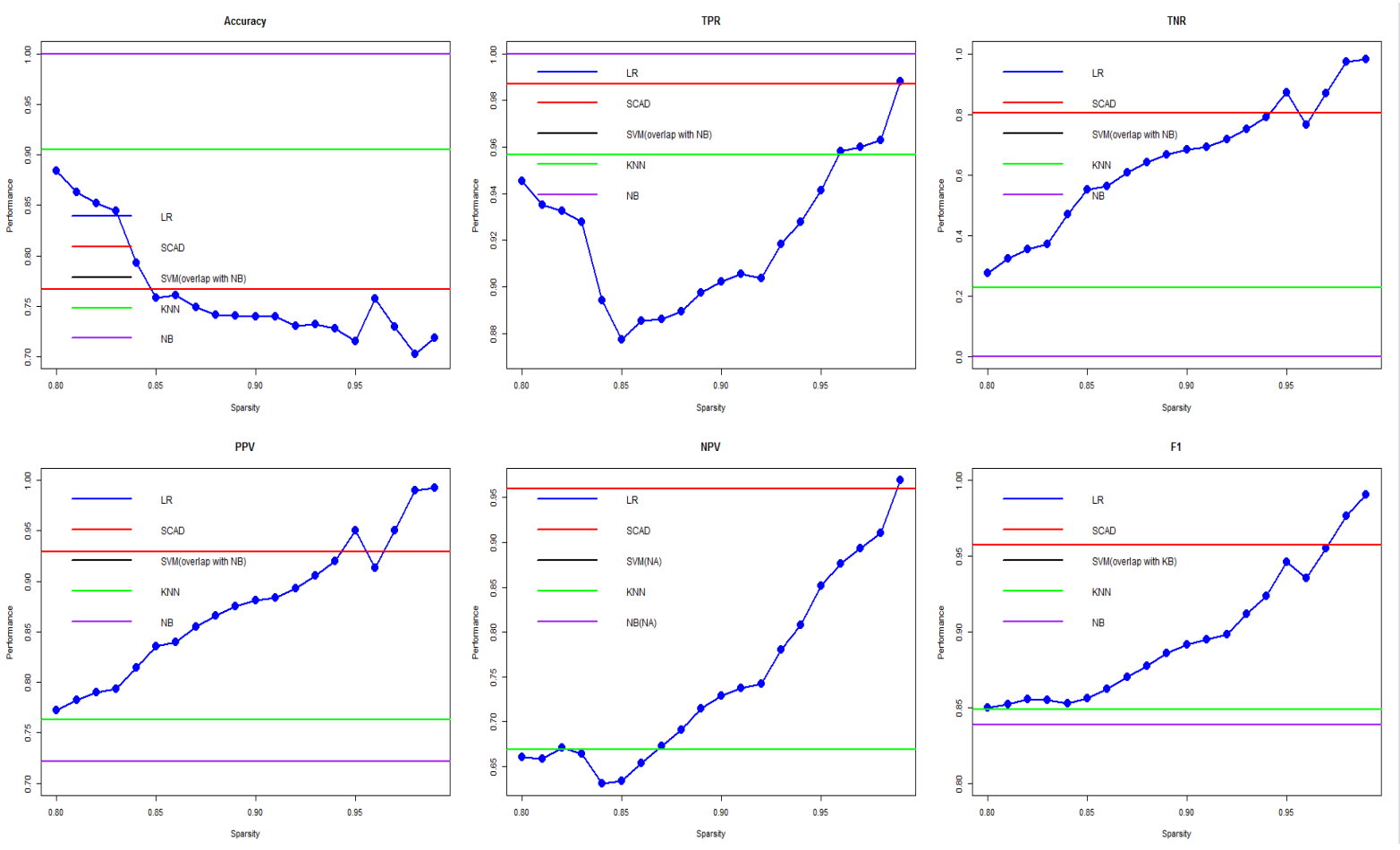

Figure 5: In-sample performance for Hotel Dataset

In-sample Performance

- Overall the SCAD regularized logistic regression model achieves good in-sample performance.
- In particular, its TNR, PPV and F1 scores all rank the highest for both datasets and NPV highest for Hotel Dataset compared with other models.

Out-of-sample Performance

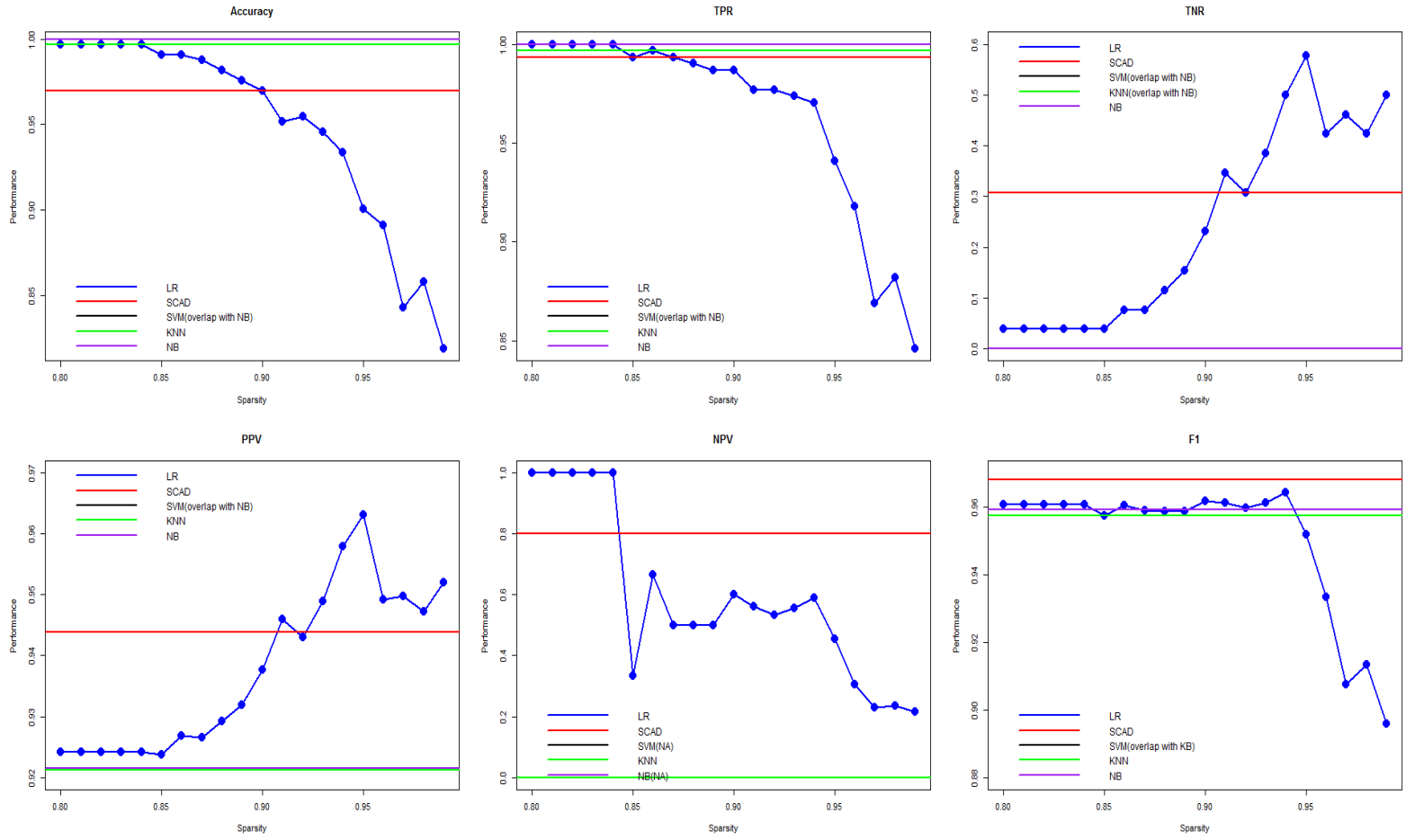

Figure 6: Out-of-sample performance on different metrics for Restaurant Dataset

Out-of-sample Performance

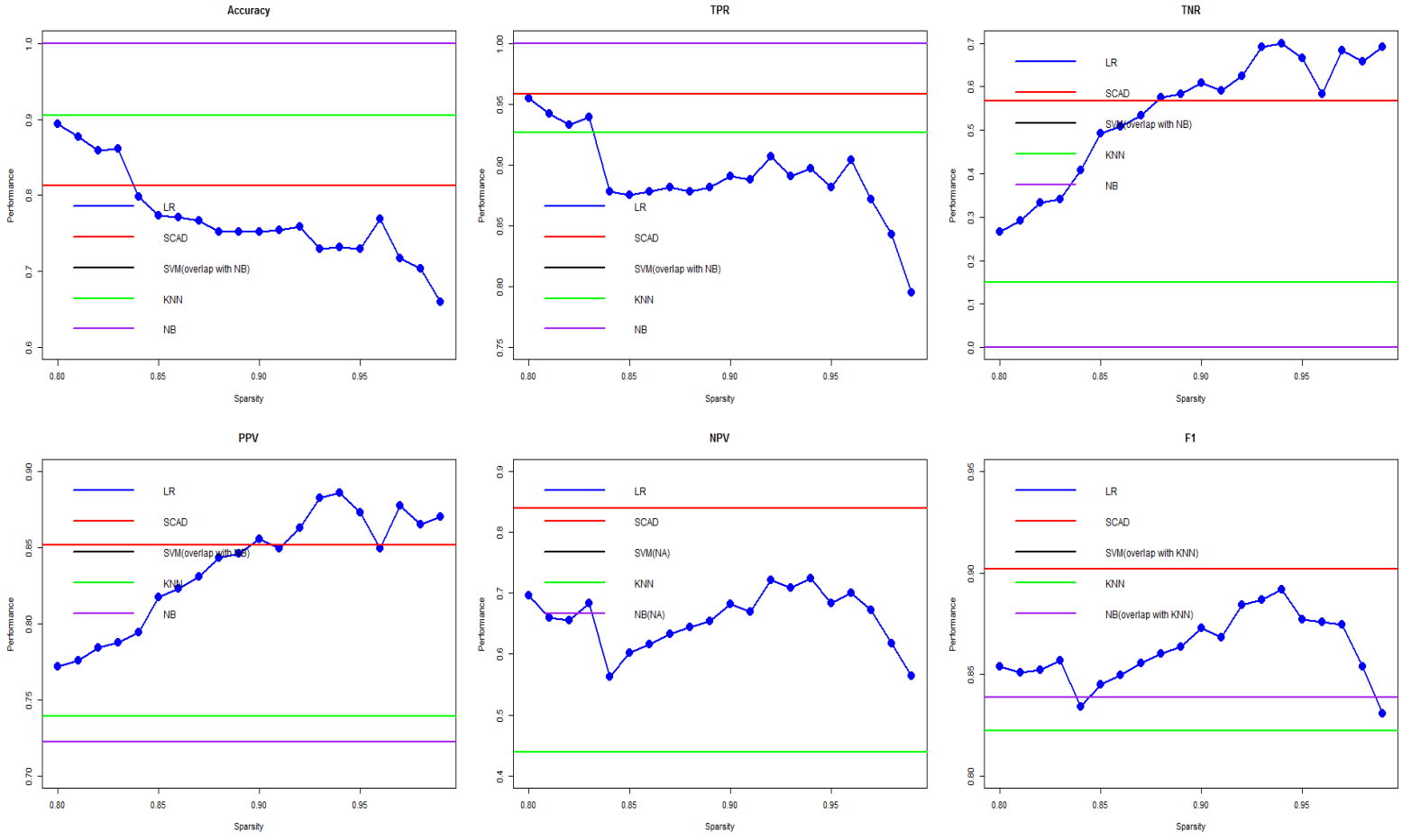

Figure 7: Out-of-sample performance on different metrics for Hotel Dataset

Out-of-sample Performance

- Overall the SCAD-regularized logistic regression model gives good performance on each metric compared with alternative models.
- In particular, its out-of-sample F1 score is better than all alternative models, showing the advantage of our data-adaptive regularization approach in striking a proper balance between model bias and variance and achieving top performance in the combined metric of precision and recall.
- The review length, namely number of words in a review added as an additional feature, has negative coefficients in both datasets, showing negative reviews tend to be long in length.

Significant features of Restaurant Dataset

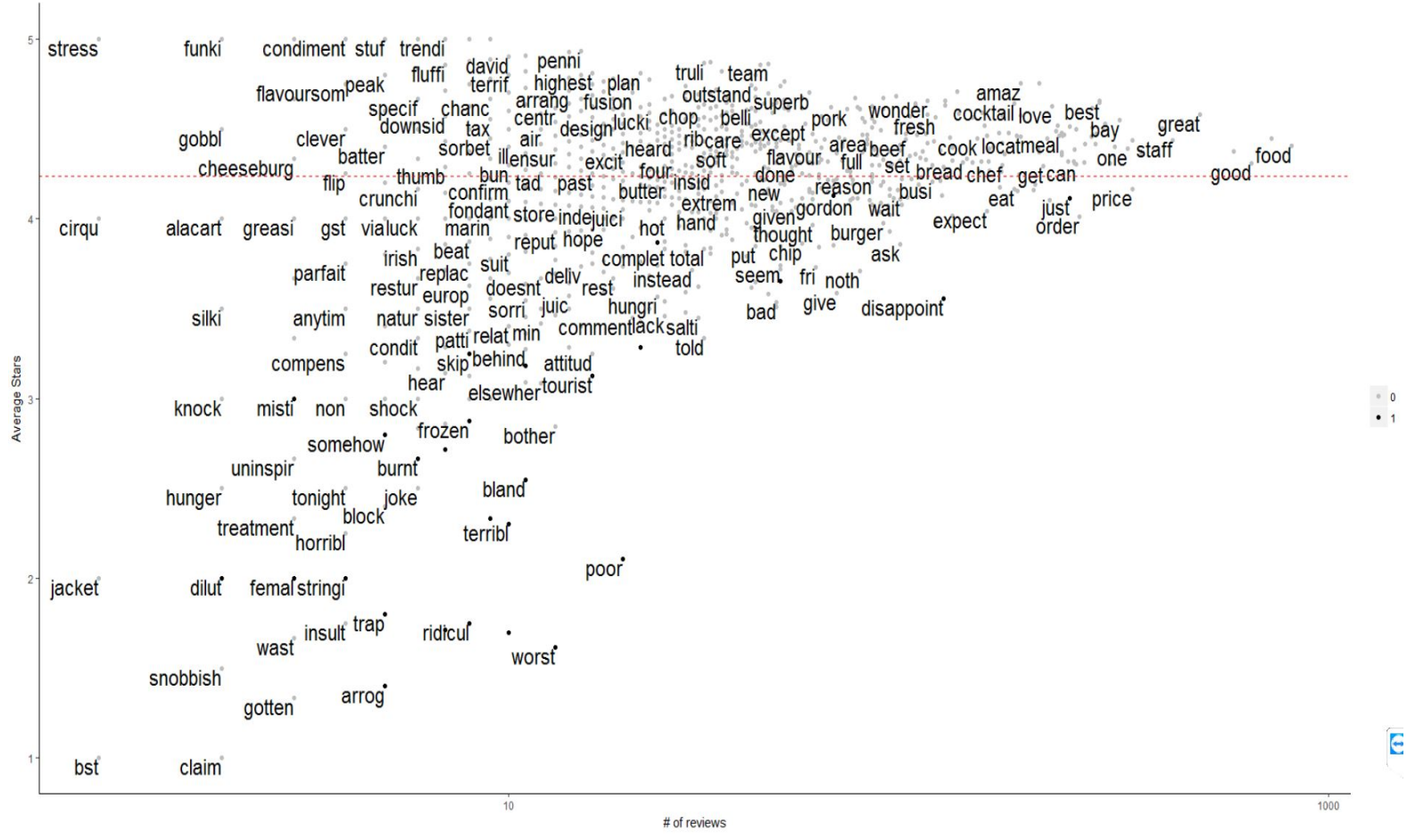

Figure 8: Significant features of Restaurant Dataset using RTL

Selected Features in Restaurant Dataset

Table 5: Selected Features in Restaurant Dataset

Features	Coefficients	No. review	Avg. rating
snobbish	-20.8842	2	1.5
poor	-18.6196	19	2.1053
knife	-17.7071	3	3
worst	-16.9393	13	1.6154
somewher	-12.2143	7	2.8571
disappoint	-9.3179	115	3.5565
dilut	-7.9327	2	2
sorri	-6.8963	11	3.5455
tasteless	-6.8474	10	1.7
shame	-6.14	8	3
stringi	-4.7924	4	2
terribl	-4.1641	10	2.3
soso	-4.0705	3	2
frozen	-3.7069	8	2.875
overpr	-2.7885	21	3.2857
tourist	-2.6612	16	3.125
embarrass	-1.4255	4	2.25
caucasian	-1.32	4	2.75
averag	-1.1767	46	3.6522
inattent	-1.0966	3	3.3333
heartburn	-1.0508	1	2
shock	-0.6174	6	3
stomach	-0.5772	11	3.1818
michelin	-0.3775	23	3.8696

Significant features of Hotel Dataset

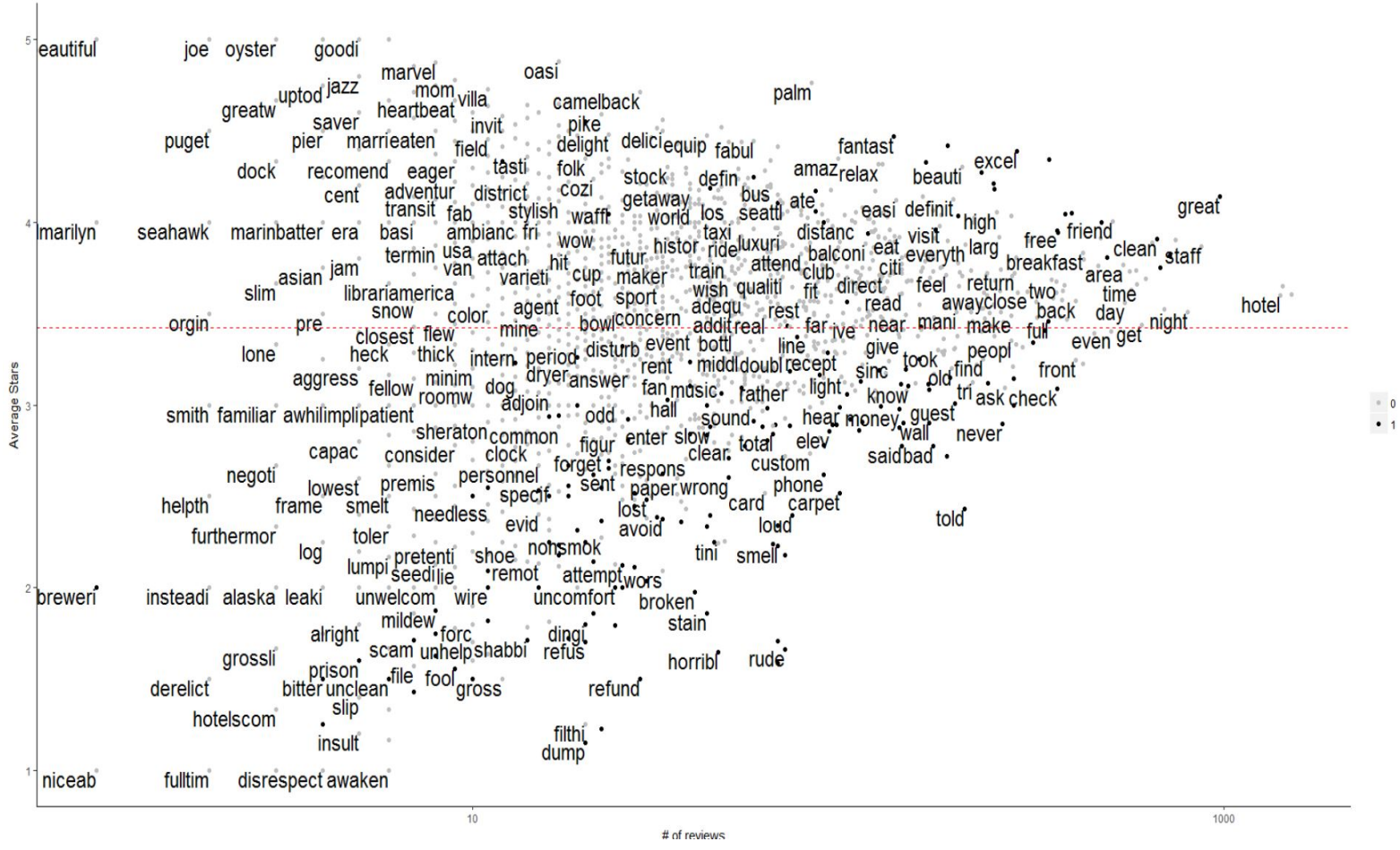

Figure 9: Significant features of Hotel Dataset using RTL

Selected Features in Hotel Dataset

Table 6: Selected Features in Hotel Dataset

Features	Coefficients	No. review	Avg. rating
mcdonald	-41.7725	6	3
told	-25.9445	204	2.4314
bare	-24.9097	25	2
adjac	-19.8972	25	3.32
uncomfort	-18.8143	24	2
dirti	-18.2363	65	1.7077
advertis	-16.5059	31	2.3871
final	-16.4063	63	2.2381
cigarett	-13.6851	18	1.7222
never	-12.0323	257	2.8988
knock	-11.7202	19	2.3158
loud	-11.4796	71	2.3944
toilet	-11.1226	44	2.25
refund	-9.1969	28	1.5
carpet	-8.1589	95	2.5158
check	-6.4943	359	3.0864
smell	-4.3192	65	2.2308
noisi	-3.1037	93	2.8925
elev	-3.0784	89	2.8539
modern	0.4574	86	4
airport	0.5277	171	3.9591
free	1.2791	359	3.9554
metro	1.7182	43	4.186
shop	2.7443	196	4.0357
store	3.6455	65	4.1077
spacious	4.8664	161	4.3292

Discussion

- From a managerial perspective, such models could help managers rapidly identify the words that significantly influence the review polarity and their respective weights in classifying customer opinions.
- Features selected via RTL are generally more informative and useful in helping hoteliers quickly drill down specific aspects for operational improvement.
- For example, features like "stringi" and "dilut" in Restaurant Dataset or "cigarett" and "circuit" in Hotel Dataset are both significant in determining the review rating and also self-explanatory of the key aspects covered in these reviews, thus effectively highlighting the challenges which require attention.

Managerial Insight

Table 7: Reviews containing the root stringi in Restaurant Dataset

Rating	Content
2	The views were wonderful but really not worth the price...then a steak which was tough, **stringy**, and bland...
2	...Steak had no sear and was very **stringy**...nothing more than overpriced pub food
2	The food is overpriced ...beef was **stringy**...They screwed up.
2	We went here with high expectations...there was red **stringy** veins through it...

Managerial Insight

Table 8: Reviews containing the root dilut in Restaurant Dataset

Rating	Content
2	...did not have much taste **diluted**...
2	...they look beautiful but are generally a little **diluted**...

Managerial Insight

Table 9: Reviews containing the root cigarett in Hotel Dataset

Rating	Content
2	...Room was clean but was non-smoking and smelled like **cigarettes**. Wouldn't stay there again.
2	...One nice touch was the fact that the bedroom window actually opened, which was a Godsend when the 'smoke free' hotel A/C system was pouring **cigarette** smoke into our room (ugh!)
1	...**Cigarette** stains in the carpets. Dark and dingy...

Managerial Insight

Table 10: Reviews containing the root circuit in Hotel Dataset

Rating	Content
2	...it was very noisy and I wouldn't have been able to sleep, had to call maintenance to turn off via a **circuit** breaker...
2	...but I didn't have a noise problem, other than the deafeningly loud fan that I couldn't figure out how to shut off until I called the front desk the next day (hint: it was the giant switch in the **circuit** breaker)...

Working efficiency

- Another aspect we're interested in is how fast the useful information on operational improvement can be delivered to hoteliers, i.e. if the set of features can identify a relatively small set of reviews that should be paid attention to.
- Two metrics are used to measure the efficiency of getting useful information from all the reviews - the total number of reviews which contain the significant features, and the total length of characters in these reviews at log scale.
- Results show that reviews that require further perusal from SCAD regularized model are significantly less compared with those from unregularized models, hoteliers can thus focus on only a small set of relevant reviews instead of browsing through a large but not necessarily informative set of reviews.

Efficiency Comparison

Table 11: Distributional statistics on the number of reviews required by different model for Restaurant Dataset

Model	Threshold	Mean	Median	Standard deviation
LR	0.8	491	462	111
LR	0.81	491	462	111
LR	0.82	429	415	129
LR	0.83	468	462	144
LR	0.84	386	346	134
LR	0.85	359	300	123
LR	0.86	328	276	117
LR	0.87	328	276	117
LR	0.88	314	270	117
LR	0.89	313	270	118
LR	0.9	301	268	113
LR	0.91	289	265	118
LR	0.92	269	246	120
LR	0.93	253	233	129
LR	0.94	229	187	131
LR	0.95	200	160	128
LR	0.96	157	113	114
LR	0.97	0	0	NA
LR	0.98	0	0	NA
LR	0.99	0	0	NA
RTL	NA	6	1	14

Efficiency Comparison

Table 12: Distributional statistics on the length of reviews required by different model for Restaurant Dataset

Model	Threshold	Mean	Median	Standard deviation
LR	0.8	46267	44309	10796
LR	0.81	46267	44309	10796
LR	0.82	41625	39370	11031
LR	0.83	45574	44309	11777
LR	0.84	37620	33940	11866
LR	0.85	35120	32553	11096
LR	0.86	33802	31369	9815
LR	0.87	33802	31369	9815
LR	0.88	32757	31073	9649
LR	0.89	32699	31073	9708
LR	0.9	31704	30881	9462
LR	0.91	30618	28492	9918
LR	0.92	29045	25839	9948
LR	0.93	27016	24935	10672
LR	0.94	24518	22727	11285
LR	0.95	22204	19273	11126
LR	0.96	18069	14495	10808
LR	0.97	0	0	NA
LR	0.98	0	0	NA
LR	0.99	0	0	NA
RTL	NA	857	238	1954

Efficiency Comparison

Table 13: Distributional statistics on the number of reviews required by different model for Hotel Dataset

Model	Threshold	Mean	Median	Standard deviation
LR	0.8	674	617	267
LR	0.81	641	584	256
LR	0.82	617	563	253
LR	0.83	605	532	260
LR	0.84	564	493	259
LR	0.85	549	471	262
LR	0.86	507	467	172
LR	0.87	474	407	179
LR	0.88	471	382	254
LR	0.89	438	359	260
LR	0.9	433	359	263
LR	0.91	391	291	256
LR	0.92	374	274	272
LR	0.93	336	240	257
LR	0.94	296	204	224
LR	0.95	239	177	160
LR	0.96	218	156	184
LR	0.97	179	122	170
LR	0.98	141	88	153
LR	0.99	99	54	132
RTL	NA	47	4	128

Efficiency Comparison

Table 14: Distributional statistics on the length of reviews required by different model for Hotel Dataset

Model	Threshold	Mean	Median	Standard deviation
LR	0.8	143739	133666	46834
LR	0.81	139264	131300	45401
LR	0.82	134108	130672	44425
LR	0.83	131625	128287	45947
LR	0.84	125971	117016	47145
LR	0.85	121050	110539	47767
LR	0.86	113321	109030	31174
LR	0.87	106061	100030	33872
LR	0.88	106332	99357	46598
LR	0.89	99196	81197	47268
LR	0.9	97287	79588	48397
LR	0.91	88465	73485	47273
LR	0.92	85223	69890	51287
LR	0.93	77416	60835	50025
LR	0.94	69345	52437	44595
LR	0.95	57415	44325	32188
LR	0.96	53925	40614	38387
LR	0.97	44935	31821	36089
LR	0.98	35722	24336	33288
LR	0.99	25414	14947	29346
RTL	NA	11396	1220	26971

Conclusion

- We proposed regularized text logistic (RTL) regression model that achieves steadily good performance in various metrics compared with alternative models.
- The RTL regression model selected a set of key useful features which could help hoteliers make informed decisions for operational improvement in an efficient and effective manner.

Coefficient Path

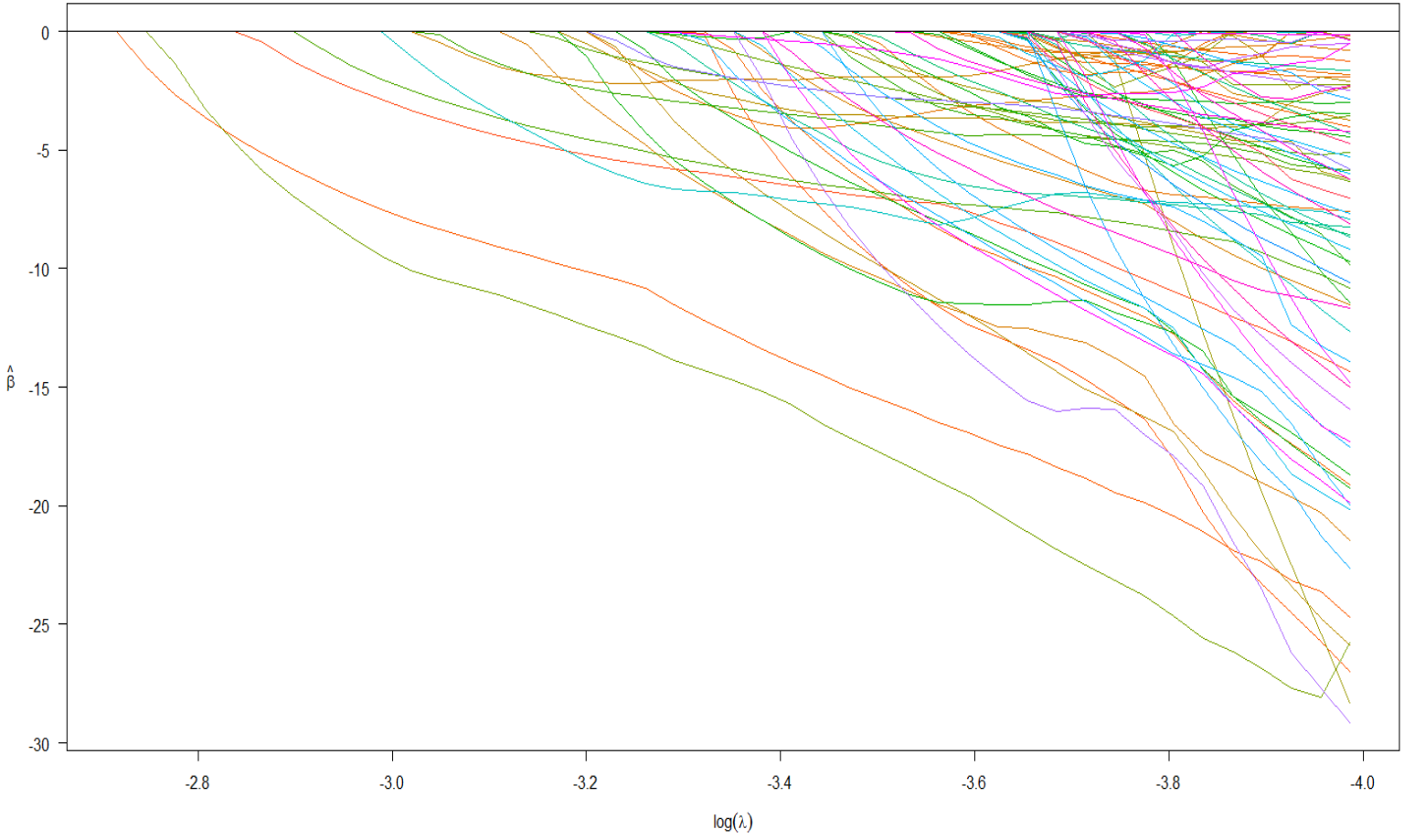

Figure 10: Coefficient path using optimal parameters for Restaurant Dataset

Coefficient Path

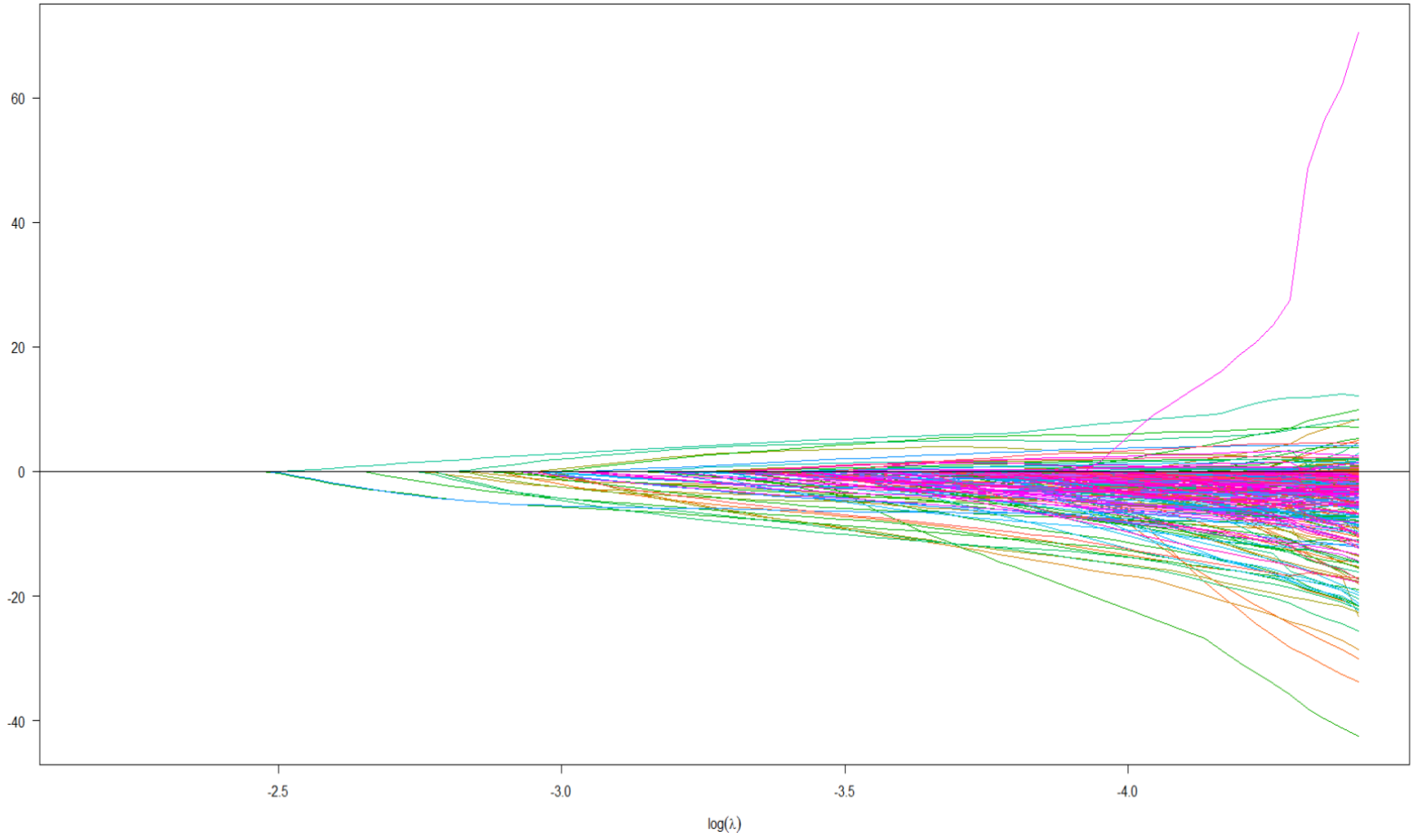

Figure 11: Coefficient path using optimal parameters for Hotel Dataset

Proof of Asymptotic Consistency

The proof essentially follows (Fan and Peng, 2004), although in this case we prove existence and consistency of the SCAD estimator via minimization of the objective function. Let $\alpha_n = \sqrt{p_n/n}$. It's enough to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$Pr\{\inf_{\|u\|=C} Q_n(\beta_0 + \alpha_n u) > Q_n(\beta_0)\} \geq 1 - \epsilon$$

This implies that with probability tending to 1 there is a local minimum in the ball $\{\beta_0 + \alpha_n u : \|u\| \leq C\}$ such that $\|\hat{\beta}_n - \beta_0\| = O_p(\sqrt{p_n/n})$.

Proof of Asymptotic Consistency

Denote $D_n(u) = Q_n(\beta_0 + \alpha_n u) - Q_n(\beta_0)$. Since $p_{\lambda_n}(0) = 0$,

$$\begin{aligned} D_n(u) &= L_n(\beta_0 + \alpha_n u) - L_n(\beta_0) + \Sigma_{j=1}^{p_n}[p_{\lambda_n}(\beta_{0j} + \alpha_n u_j) - p_{\lambda_n}(|\beta_{0j}|)] \\ &\geq L_n(\beta_0 + \alpha_n u) - L_n(\beta_0) + \Sigma_{j=1}^{k_n}[p_{\lambda_n}(\beta_{0j} + \alpha_n u_j) - p_{\lambda_n}(|\beta_{0j}|)] \\ &\triangleq (C.9) + (II) \end{aligned}$$

Using Taylor's expansion, we have

$$\begin{aligned} (C.9) &= \alpha_n \nabla^\top L_n(\beta_0) u + \frac{1}{2} u^\top \nabla^2 L_n(\beta_0) u \alpha_n^2 + \frac{1}{6} \nabla^\top \{u^\top \nabla^2 L_n(\beta_n^*) u\} u \alpha_n^3 \\ &\triangleq I_1 + I_2 + I_3 \end{aligned}$$

where β_n^* lies between β_0 and $\beta_0 + \alpha_n u$, and

$$\begin{aligned} (II) &= \Sigma_{j=1}^{k_n}[\alpha_n p\prime_{\lambda_n}(|\beta_{0j}|)\mathrm{sgn}(\beta_{0j}) u_j + \alpha_n^2 p\prime\prime_{\lambda_n}(|\beta_{0j}|) u_j^2 \{1 + o(1)\}] \\ &\triangleq I_4 + I_5 \end{aligned}$$

Proof of Asymptotic Consistency

According to Condition (C.3) and (C.7),

$$|I_1| = |\alpha_n \nabla^\top L_n(\beta_0) u| \leq \alpha_n \|\nabla^\top L_n(\beta_0)\| \|u\|$$
$$= O_p(\alpha_n/\sqrt{n})\|u\| = O_p(\alpha_n^2/n)\|u\|$$

Now we consider I_2. Using Chebyshev's inequality, for any ϵ, we have

$$P(\|\frac{1}{n}\nabla^2 L_n(\beta_0) - I_n(\beta_0)\| \geq \frac{\epsilon}{p_n})$$
$$\leq \frac{p_n^2}{n^2\epsilon^2} E \sum_{i,j=1}^{p_n} \{\frac{\partial L_n(\beta_0)}{\partial \beta_i \beta_j} - E\frac{\partial L_n(\beta_0)}{\partial \beta_i \beta_j}\}^2$$
$$= \frac{p_n^4}{n} = o(1)$$

which results in $\|\frac{1}{n}\nabla^2 L_n(\beta_0) - I_n(\beta_0)\| = o_p(\frac{1}{p_n})$.

Proof of Asymptotic Consistency

Thus we have

$$I_2 = \frac{1}{2}u^\top[\frac{1}{n}\{\nabla^2 L_n(\beta_0) - E\nabla^2 L_n(\beta_0)\}]un\alpha_n^2 + \frac{1}{2}u^\top I_n(\beta_0)un\alpha_n^2$$
$$= \frac{n\alpha_n^2}{2}u^\top I_n(\beta_0)u + o_p(1)n\alpha_n^2\|u^2\|$$

Based on Condition (C.8) and Cauchy-Schwarz inequality, we have

$$|I_3| = |\frac{1}{6}\sum_{i,j,k=1}^{p_n}\frac{\partial L_n(\beta_n^*)}{\partial\beta_{ni}\partial\beta_{nj}\partial\beta_{nk}}u_i u_j u_k\alpha_n^3|$$
$$\leq \frac{1}{6}\sum_{l=1}^{n}\{\sum_{i,j,k=1}^{p_n}M_{nijk}^2(V_{nl})\}^{1/2}\|u\|^3\alpha_n^3$$

Proof of Asymptotic Consistency

Since $p_n^4/n \to 0$ and $p_n^2\alpha_n \to 0$ as $n \to \infty$, we have

$$\frac{1}{6}\sum_{l=1}^{n}\{\sum_{i,j,k=1}^{p_n} M_{nijk}^2(V_{nl})\}^{1/2}\|u\|^3\alpha_n^3$$

$$= O_p(p_n^{3/2}\alpha_n)n\alpha_n^2\|u\|^2 = o_p(n\alpha_n^2)\|u\|^2$$

Thus $I_3 = o_p(n\alpha_n^2)\|u\|^2$. In addition,

$$|I_4| = \Sigma_{j=1}^{k_n}|\alpha_n p\prime_{\lambda_n}(|\beta_{0j}|)\mathrm{sgn}(\beta_{0j})u_j| \le \sqrt{k_n}n\alpha_n\alpha_n\|u\| \le n\alpha_n^2\|u\|$$

and

$$|I_5| = \Sigma_{j=1}^{k_n}\alpha_n^2 p\prime\prime_{\lambda_n}(|\beta_{0j}|)u_j^2\{1+o(1)\} \le 2\max_{1\le j\le k_n} p\prime\prime_{\lambda_n}(|\beta_{0j}|)\alpha_n^2\|u\|^2$$

By choosing a sufficiently large C, I_1, I_3, I_4 and I_5 are all dominated by I_2, which is positive. This completes the proof of this theorem.

Proof of Sparsity

It is sufficient to show that with probability tending to 1 as $n \to 0$, for any β_{1n} satisfying $\|\beta_{1n} - \beta_{10}\| = O_p(\sqrt{p_n/n})$, for some $\epsilon_n = C\sqrt{p_n/n}$ and $j = k_n + 1, ..., p_n$,

$$\begin{cases} \frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} > 0 & \text{for} \quad 0 < \beta_{nj} < \epsilon_n, \\ \frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} < 0 & \text{for} \quad -\epsilon_n < \beta_{nj} < 0 \end{cases}$$

Proof of Sparsity

By Taylor's expansion,

$$\begin{aligned}
\frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} &= \frac{\partial L_n(\beta_n)}{\partial \beta_{nj}} - p\prime\lambda_n(|\beta_{nj}|)\mathrm{sgn}(\beta_{nj}) \\
&= \frac{\partial L_n(\beta_0)}{\partial \beta_{nj}} + \sum_{l=1}^{p_n} \frac{\partial^2 L_n(\beta_0)}{\partial \beta_{nj}\partial \beta_{nl}}(\beta_{nl} - \beta_{0l}) \\
&+ \sum_{l,k=1}^{p_n} \frac{\partial^3 L_n(\beta_n^*)}{\partial \beta_{nj}\partial \beta_{nl}\partial \beta_{nk}}(\beta_{nl} - \beta_{0l})(\beta_{nk} - \beta_{0k}) \\
&\quad - p\prime\lambda_n(|\beta_{nj}|)\mathrm{sgn}(\beta_{nj}) \\
&\triangleq I_1 + I_2 + I_3 + I_4
\end{aligned}$$

where β_n^* lies between β_n and β_0.

Proof of Sparsity

For I_1, we have

$$I_1 = O_p(\sqrt{n}) = O_p(\sqrt{np_n})$$

For I_2, we have

$$I_2 = \sum_{l=1}^{p_n} \{\frac{\partial^2 L_n(\beta_0)}{\partial\beta_{nj}\partial\beta_{nl}} - E\frac{\partial^2 L_n(\beta_0)}{\partial\beta_{nj}\partial\beta_{nl}}\}(\beta_{nl} - \beta_{0l})$$
$$+ \sum_{l=1}^{p_n} \frac{\partial^2 L_n(\beta_0)}{\partial\beta_{nj}\partial\beta_{nl}}(\beta_{nl} - \beta_{0l}) \triangleq K_1 + K_2$$

Using Cauchy-Schwarz inequality and $\|\beta_{1n} - \beta_{10}\| = O_p(\sqrt{p_n/n})$, we have

$$|K_2| = |n\sum_{l=1}^{p_n} I_n(\beta_0)(j,l)(\beta_{nl} - \beta_{0l})| \leq nO_p(\sqrt{\frac{p_n}{n}})\{\sum_{l=1}^{p_n} I_n^2(\beta_0)(j,l)\}^{1/2}$$

Proof of Sparsity

Since the eigenvalues of the Fisher information matrix are bounded according to Condition (C.7), we have
$\sum_{l=1}^{p_n} I_n^2(\beta_0)(j,l) = O(1)$. Thus $K_2 = O_p(\sqrt{np_n})$
For K_1, by Cauchy-Schwarz inequality we have

$$|K_1| \leq \|\beta_{nl} - \beta_{0l}\| \left[\sum_{l=1}^{p_n} \left\{ \frac{\partial^2 L_n(\beta_0)}{\partial \beta_{nj} \partial \beta_{nl}} - E \frac{\partial^2 L_n(\beta_0)}{\partial \beta_{nj} \partial \beta_{nl}} \right\}^2 \right]^{1/2}$$

From Condition (C.7), we have

$$\left[\sum_{l=1}^{p_n} \left\{ \frac{\partial^2 L_n(\beta_0)}{\partial \beta_{nj} \partial \beta_{nl}} - E \frac{\partial^2 L_n(\beta_0)}{\partial \beta_{nj} \partial \beta_{nl}} \right\}^2 \right]^{1/2} = O_p(\sqrt{np_n})$$

Thus $K = O_p(\sqrt{np_n})$, and $I_2 = O_p(\sqrt{np_n})$.

Proof of Sparsity

For I_3, we can write it as:

$$I_3 = \sum_{l,k=1}^{p_n} \left\{ \frac{\partial^3 L_n(\beta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nk}} - E\frac{\partial^3 L_n(\beta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nk}} \right\}(\beta_{nl} - \beta_{0l})(\beta_{nk} - \beta_{0k})$$

$$+ \sum_{l,k=1}^{p_n} E\frac{\partial^3 L_n(\beta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nk}}(\beta_{nl} - \beta_{0l})(\beta_{nk} - \beta_{0k})$$

$$\triangleq K_3 + K_4$$

For K_3, by Cauchy-Schwarz inequality,

$$K_3^2 \leq \sum_{l,k=1}^{p_n} \left\{ \frac{\partial^3 L_n(\beta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nk}} - E\frac{\partial^3 L_n(\beta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nk}} \right\}\|\beta_n - \beta_0\|^4$$

Proof of Sparsity

Under Condition (C.8) and (C.9), we have

$$K_3 = O_p\Big\{(np_n^2\frac{p_n^2}{n^2})^{1/2}\Big\} = o_p(\sqrt{np_n})$$

For K_4, by Condition (C.8),

$$|K_4| \leq C_5^{1/2} np_n \|\beta_n - \beta_0\|^2 = O_p(p_n^2) = o_p(\sqrt{np_n})$$

From above analysis, we have

$$I_1 + I_2 + I_3 = O_p(\sqrt{np_n})$$

Proof of Sparsity

Since $\sqrt{p_n/n}/\lambda_n \to 0$ and $\liminf_{n\to\infty}\liminf_{\beta_n\to 0^+} p'_{\lambda_n}(|\beta_n|)/\lambda_n > 0$, from

$$\frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}} = n\lambda_n\Big\{ - \frac{p'_{\lambda_n}(|\beta_{nj}|)}{\lambda_n}\mathrm{sgn}(\beta_{nj}) + O_p\Big(\sqrt{\frac{p_n}{n}}/\lambda_n\Big)\Big\}$$

It is easy to see that the sign of β_{nj} completely determines the sign of $\partial Q_n(\beta_n)/\partial\beta_{nj}$. This completes the proof.

Proof of Oracle Property

As shown in the theorem on consistency, there is a root-(n/p_n)-cons$^\top$*istentlocalminimizer*$_n$ of $Q_n(\beta_n)$. By the lemma on sparsity, part (1) holds. We only need to prove part (2), the asymptotic normality of the penalized estimator $\hat{\beta}_{1n}$. (Fan and Peng, 2004) showed that

$$(I_n(\beta_{10}) + \Sigma_{\lambda_n})(\hat{\beta}_{1n} - \beta_{10}) + b_n = \frac{1}{n}\nabla L_n(\beta_{10}) + o_p(n^{-1/2})$$

Based on this result, we focus on its asymptotic distribution towards standard normal distribution. It is easy to see that

$$\sqrt{n}\alpha_n^\top I_n^{-1/2}(\beta_{10})(I_n(\beta_{10}) + \Sigma_{\lambda_n})[(\hat{\beta}_{1n} - \beta_{10}) + (I_n(\beta_{10}) + \Sigma_{\lambda_n})^{-1} b_n]$$
$$= \frac{1}{\sqrt{n}}\alpha_n^\top I_n^{-1/2}(\beta_{10})\nabla L_n(\beta_{10}) + o_p(\alpha_n^\top I_n^{-1/2}(\beta_{10}))$$

Proof of Oracle Property

Given the conditions in Theorem on Oracle property, the last term is equivalent to $o_p(1)$. Let

$$Y_{in} = \frac{1}{\sqrt{n}} \alpha_n^\top I_n^{-1/2}(\beta_{10}) \nabla L_{ni}(\beta_{10}), i = 1, 2, ..., n$$

We consider if Y_{ni} meets the conditions of Lindeberg-Feller central limit theorem. It follows that for any ϵ,

$$\begin{aligned} \sum_{i=1}^{n} E\|Y_{in}\|^2 \mathbb{1}\{\|Y_{in}\| > \epsilon\} &= nE\|Y_{1n}\|^2 \mathbb{1}\{\|Y_{in}\| > \epsilon\} \\ &\leq n\{E\|Y_{1n}\|^4\}^{1/2}\{P(\|Y_{1n}\| > \epsilon)\}^{1/2} \end{aligned}$$

Proof of Oracle Property

By Condition (C.7) and since α_n is an arbitrary k_nx1 vector such that $||\alpha_n|| = 1$, we have

$$P(\|Y_{1n}\| > \epsilon)) \leq \frac{E\|\alpha_n^\top I_n^{-1/2}(\beta_{10}) \triangledown L_{1n}(\beta_{10})\|^2}{n\epsilon} = O(n^{-1})$$

and

$$E\|Y_{1n}\|^4 = \frac{1}{n^2} E\|\alpha_n^\top I_n^{-1/2}(\beta_{10}) \triangledown L_{1n}(\beta_{10})\|^4$$

$$\leq \frac{1}{n^2} \lambda_{max}(I_n(\beta_{10})) E\|\triangledown^\top L_{1n}(\beta_{10}) \triangledown L_{1n}(\beta_{10})\|^2$$

$$= O(\frac{p_n^2}{n^2})$$

Proof of Oracle Property

Thus we have

$$\sum_{i=1}^{n} E\|Y_{in}\|^2 \mathbb{1}\{\|Y_{in}\| > \epsilon\} = O(n\frac{p_n}{n}\frac{1}{\sqrt{n}}) = o_p(1)$$

On the other hand, we have

$$\sum_{i=1}^{n} \text{cov}(Y_{in}) = n\text{cov}(Y_{1n}) = \text{cov}\{\alpha_n^\top I_n^{-1/2}(\beta_{10})\nabla L_{1n}(\beta_{10})\} \to 1$$

Thus $1/\sqrt{n}\alpha_n^\top I_n^{-1/2}(\beta_{10})\nabla L_{1n}(\beta_{10})$ has an asymptotic standard normal distribution. This completes the proof.

Proof of Global Property

We refer to the proof given by (Wang et. al, 2016), in which quadratic approximation of the loss function is used, so that

$$Q_n(\beta_n) = \frac{1}{2}(\beta_n - \tilde{\beta}_n)^\top \hat{\Omega}(\beta_n - \tilde{\beta}_n) + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|)$$

where $\hat{\Omega}$ is an estimate of Σ^{-1}. The asymptotic covariance matrix Σ and its inverse matrix Ω are further decomposed into the following block matrix forms respectively according to the sparsity property in the lemma on sparsity:

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}$$

Proof of Global Property

It can be verified that

$$I_n = \Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}$$

$$\begin{bmatrix} \Sigma_{(11)}^{-1} & -\Sigma_{(11)}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ = -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{(11)}^{-1}\Sigma_{22}^{-1} & \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{(11)}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix}$$

where $\Sigma_{(11)}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Similarly, $\hat{\Omega}$ can also be partitioned as

$$\begin{bmatrix} \hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & \hat{\Omega}_{22} \end{bmatrix}$$

Proof of Global Property

The existence of global minimum indicates $Pr(Q_n(\hat{\beta}_n) \leq \inf_{\beta_n} Q_n(\beta_n)) \to 1$, which is proved in (Wang et. al, 2016) through $\hat{\beta}_{1n} = \tilde{\beta}_{1n} + \hat{\Omega}_{11}^{-1}\hat{\Omega}_{12}\tilde{\beta}_{2n}$ and we'll not further discuss here. It should be noted that a stronger condition for ensuring global minimum is obtained is discussed in (Breheny and Huang, 2011), in which the objective function $Q_n(\beta_n)$ is convex with respect to β_n despite of the nonconvex penalty component provided that $c_*(\beta_n) > 1/(\gamma - 1)$, where $c_*(\beta_n)$ is the minimum eigenvalue of $n^{-1}X^\top WX$, W is a diagonal matrix of weights with elements $w_i = \pi_i(1 - \pi_i)$, π is the predicted probability based on most recent coordinate update, and γ is a hyperparameter in the penalty function.

References


Fan, J., Li, R.
Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties
Journal of the American Statistical Association, 96, 456, 2001.


Francis L.
Taming Text: An Introduction to Text Mining
Casualty Actuarial Society Forum, 51-88, 2006.


Genkin, Alexander., Lewis, D.D, and Madigan, D.
Large-Scale Bayesian Logistic Regression for Text Categorization
American Statistical Association and the American Society for Quality, 49, 2007.

References


Greenland S, Schwartzbaum, J.A., and Finkle, W.D.
Problems From Small Samples and Sparse Data in Conditional Logistic Regression Analysis
American Journal of Epidemiology, 151, 531-539, 2000.


Hadjicostas, P.
Consistency of Logistic Regression Coefficient Estimates Calculated From a Training Sample
Statistics and Probability Letters, 62, 293-303, 2003.


Pang, B., Lee, L., Vaithyanathan S.
Thumbs up? Sentiment Classification using Machine Learning Techniques
Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10, 79-86, 2002.

References


Tibshirani, R.
Regression Shrinkage and Selection via the Lasso
Journal of the Royal Statistical Society, 267-288, 1996.


Wang, H., Lu, Y., and Zhai, C.
Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach
The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, p783-792, 2010.


Hongning, Wang
Learning Online Discussion Structures by Conditional Random Fields
The 34th Annual International ACM SIGIR Conference, 435-444, 2011.