Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Sequential Text-Term Selection in Vector Space Models

Feifei Wang, Jingyuan Liu & Hansheng Wang

To cite this article: Feifei Wang, Jingyuan Liu & Hansheng Wang (2021) Sequential Text-Term Selection in Vector Space Models, Journal of Business & Economic Statistics, 39:1, 82-97, DOI: 10.1080/07350015.2019.1634079

To link to this article: https://doi.org/10.1080/07350015.2019.1634079

Sequential Text-Term Selection in Vector Space Models

Feifei WANG
Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing 100872, China (*ff11161224@126.com*)

Jingyuan LIU
MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Lab of Statistics, Xiamen University, Xiamen 361005, China (*jingyuan@xmu.edu.cn*)

Hansheng WANG
Guanghua School of Management, Peking University, Beijing 100871, China (*hansheng@gsm.pku.edu.cn*)

Text mining has recently attracted a great deal of attention with the accumulation of text documents in all fields. In this article, we focus on the use of textual information to explain continuous variables in the framework of linear regressions. To handle the unstructured texts, one common practice is to structuralize the text documents via vector space models. However, using words or phrases as the basic analysis terms in vector space models is in high debate. In addition, vector space models often lead to an extremely large term set and suffer from the curse of dimensionality, which makes term selection important and necessary. Toward this end, we propose a novel term screening method for vector space models under a linear regression setup. We first split the entire term space into different subspaces according to the length of terms and then conduct term screening in a sequential manner. We prove the screening consistency of the method and assess the empirical performance of the proposed method with simulations based on a dataset of online consumer reviews for cellphones. Then, we analyze the associated real data. The results show that the sequential term selection technique can effectively detect the relevant terms by a few steps.

KEY WORDS: Screening consistency; Term selection; Text mining; Vector space models.

1. INTRODUCTION

Text mining, also known as text data mining, refers to extracting high-quality information from unstructured text documents (Tan 1999). Text mining has wide applications and has become increasingly important with the increasing accumulation of text documents in all fields. For example, text-based analysis of consumer reviews has attracted considerable attention in both academic research and managerial practice (Berger, Sorensen, and Rasmussen 2010; Lee and Bradlow 2011; Ludwig et al. 2013; Zhao et al. 2013). Researchers use text mining techniques to uncover the hidden reasons for consumer preference. Other applications include information retrieval, spam detection, sentiment analysis, and so on (Kumar and Bhatia 2013).

In this article, we confine our interests in using text documents to explain a continuous response. Specifically, each text document is paired with a univariate and continuous response variable. For example, the text document could be an online product description and its corresponding continuous response variable could be the click-through-rate (CTR). For another example, the text document could be a consumer review, and the continuous response variable is the corresponding rating score. The goal is to exploit the dependent relationship between text document and the response. Taking cellphone reviews as an example, practitioners might care about why the rating score of a certain type of cellphones is low/high. Do the battery, the after-sales service, or other extraordinary functions have an impact on the rating score after controlling for the effects of all the other factors? To answer these questions, a regression model, studying the relationship between cellphone reviews and the rating scores, is needed.

Given that textual data are highly unstructured, the first step in analyzing the text documents is to make them structured. A popular paradigm of structuralizing text documents is vector space models (Salton, Wong, and Yang 1975; Salton 1989; Belew and Rijsbergen 2000). Specifically, we refer to the basic analysis unit in documents as *terms*, which can correspond to words or phrases. All unique terms, typically with nonignorable frequencies, construct a term dictionary $\mathcal{T}$ with size p. Vector space models have been used in various domains of text mining, such as information retrieval, relevancy rankings, document clustering and classification. For example, in information retrieval, both queries and documents are represented by vector space models, and the relevance of a document to a query is given by the similarity of their term vectors

Journal of Business & Economic Statistics
January 2021, Vol. 39, No. 1
DOI: 10.1080/07350015.2019.1634079
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jbes*.
This article has been corrected with minor changes. These changes do not impact the academic content of the article.

(Salton and Buckley 1988; Hofmann 1999; Manning, Raghavan, and Schütze 2008). In applications of document clustering or classification, the unsupervised (clustering) or supervised (classification) machine learning methods are applied directly on the document-term vectors (Yang and Pedersen 1997; Li and Zha 2006; Gomez and Moens 2012; Jia et al. 2014). A more detailed discussion for vector space models can be found in Turney and Pantel (2010).

To better understand vector space models, we clarify two important concepts, "word" and "phrase," as the terms in vector space models could be either words or phrases. In linguistics, a word is the smallest element that can express semantic meaning, while a phrase is a sequence of words, often conveying an idiomatic meaning. A phrase is allowed to contain only one word. For example, "shipping" is a word, and it is also a single-word phrase, while "free shipping" is a phrase consisting of two words. When taking words as terms in vector space models and only focusing on those with relatively high frequencies, the size of the dictionary $\mathcal{T}$ is relatively small. However, when taking phrases as terms, even if only those with high frequencies are considered, the resulting cardinality of the dictionary could be much larger.

However, it is still preferred to use phrases as terms because a basic assumption of vector space models is exchangeability, which means the model always stands when we change the order of terms arbitrarily (Aldous 1985; Blei, Ng, and Jordan 2003; Batra, Bawa, and Punjab 2010). Under this assumption, when we take words as terms, the order of words is inevitably ignored by vector space models. However, word order is critical since it influences the meaning of the documents dramatically. For instance, the two sentences "free shipping but no free returns" and "free returns but no free shipping" are composed of the same words but represent completely different shipping policies. If "word" is the basic analysis unit in vector space models, the two sentences would result in the same word vectors, and the model could be misleading. Hence, a number of researchers have chosen phrases, which retain word order inherently, as terms in vector space models (Caropreso, Matwin, and Sebastiani 2000; Wu, Li, and Xu 2006; Ifrim, Bakir, and Weikum 2008; Jia et al. 2014). Therefore, due to the high dimensionality of a phrase dictionary and the sparse assumption that most phrases are redundant, noisy and irrelevant to the response, phrase selection is of fundamental importance.

Traditional term selection methods include information gain, mutual information, chi-square (CHI), relevancy score, correlation coefficients, etc. (Ng, Goh, and Low 1997; Yang and Pedersen 1997; Sebastiani 2002). These methods assess each term by calculating certain types of its "association" with response, and then select those with relatively high correlations (Ng, Goh, and Low 1997; Yang and Pedersen 1997; Liu et al. 2003; Yu and Liu 2003). These bivariate marginal methods have been widely used in text-mining related fields (Zhang, Wu, and Srihari 2004). However, they may miss other potentially important terms that jointly contribute to the response.

To address this issue, a variety of model-based methods have been developed (Kudo and Matsumoto 2004; Ifrim, Bakir, and Weikum 2008; Gomez and Moens 2012; Taddy 2013; Jia et al. 2014), as powerful complements to those traditional methods. Most model-based methods are mainly proposed and built for document classification. For instance, the regularized inverse regression (Taddy 2013) used the inverse conditional distribution for text given the response to obtain low-dimensional document scores, and applied the sparsity-inducing independent Laplace priors for simplifying the predictor sets. Another notable piece of work is concise comparative summarization (CCS) proposed by Jia et al. (2014). The authors apply sparse classification methods, LASSO (Tibshirani 1996) and L^1-penalized logistic regression (e.g., Genkin, Lewis, and Madigan 2007; Ifrim, Bakir, and Weikum 2008), to select phrases as concise summaries of classes. However, when continuous responses are of interest, such as CTR or rating score, the aforementioned methods have to base on a reasonably chosen cutoff before implementation. In addition to the challenge of determining cutoffs, the process of categorizing responses might lose considerable amounts of information.

In this article, we propose a novel term screening method for vector space models under a regression setup with continuous response. We allow for both words and phrases as terms in vector space models. In this regard, an important work is sure independence screening (SIS) proposed by Fan and Lv (2008). The basic idea of SIS is to first select variables that are marginally correlated with the response using a fast but crude method and then apply standard feature selection methods (e.g., LASSO) to further select variables relevant to the response. Since then, feature screening has attracted a great deal of attention in the statistical literature. For example, Wang (2009) applied a forward regression algorithm to select features and demonstrated its screening consistency property under an ultrahigh-dimensional setup. Fan and Song (2010) extended the method of SIS in generalized linear models by ranking the maximum marginal likelihood estimates. Li, Zhong, and Zhu (2012) proposed a model-free feature screening method based on the distance correlation, which can be directly applied to grouped predictors and multivariate responses. Liu, Li, and Wu (2014) developed a feature screening method for varying coefficient models based on conditional correlation coefficient. See Liu, Zhong, and Li (2015) for an overview.

Motivated by these works, we propose a sequential term selection method to identify significant terms in vector space models. Under a dictionary $\mathcal{T}$ with p unique terms, let $\tau_j (1 \leq j \leq p)$ denote the length of term j, that is, the number of words in this term. Assume $1 \leq \tau_j \leq q$, where q is the predefined maximum length. We split the entire term space (i.e., the dictionary $\mathcal{T}$) into different subspaces according to the length of the terms, that is, $\mathcal{T} = \bigcup_{1 \leq l \leq q} \mathcal{T}^{(l)}$, where $\mathcal{T}^{(l)}$ is the collection of terms with length l. We then select terms in $\mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(q)}$ in a sequential manner. We then collect all of the selected terms in this solution path, resulting in a screened term subspace. The screening consistency is developed to guarantee the validity of screened model for further regularization methods.

The rest of this article is organized as follows. Section 2 presents a sequential term selection method for vector space models. The screening consistency is carefully explored subsequently. Section 3 presents a number of numerical studies to demonstrate the finite sample performance of the method. Section 4 applies sequential term selection method to a cus-

tomer review dataset of cellphone online sales. Through this dataset, we find the sequential term selection method and bivariate marginal methods can complement each other very well. Section 5 concludes the article with a brief discussion.

2. A SEQUENTIAL TERM SELECTION METHOD

2.1. Model and Notations

Let $\mathcal{W} = \{w_1^*, w_2^*, \ldots, w_d^*\}$ be a set (or dictionary) of d distinct words, where $d \geq 1$ and $w_k^* (1 \leq k \leq d)$ represents one particular word. For example, one can define $\mathcal{W} = \{w_1^* = \text{Tom}, w_2^* = \text{loves}, w_3^* = \text{Jerry}\}$, and hence, $d = 3$. We let $\mathcal{S}$ be a term, which can be represented by a sequence of words as $\mathcal{S} = \langle w_1 w_2 \ldots w_m \rangle$. Here, $w_h (1 \leq h \leq m)$ is one particular word, which might be included (or excluded) by $\mathcal{W}$. We define an operator $\tau(\mathcal{S}) = m$ as the length of the term. $\tau(\mathcal{S}) = 1$ if $\mathcal{S}$ represents an individual word. For completeness, we allow $\mathcal{S}$ to be an empty term with $\tau(\mathcal{S}) = 0$ and write $\mathcal{S} = \langle \emptyset \rangle$. If $w_h \in \mathcal{W}$ for any $1 \leq h \leq m$, we then say $\mathcal{S}$ is generated by $\mathcal{W}$ and write $\mathcal{S} \in \mathcal{W}$. If $\mathcal{S} = \langle \emptyset \rangle$, then $\mathcal{S} \notin \mathcal{W}$.

Considering the previous example, we define $w_1 = w_1^* = \text{Tom}$, $w_2 = w_2^* = \text{loves}$ and $w_3 = w_3^* = \text{Jerry}$; then, we have $\mathcal{S} = \langle \text{Tom loves Jerry} \rangle$ and $\mathcal{S} \in \mathcal{W}$. If $\mathcal{S} = \langle \text{Tom and Jerry} \rangle$, then $\mathcal{S} \notin \mathcal{W}$ because "and" is excluded by $\mathcal{W}$. Let $\mathcal{S}_1 = \langle w_{i_1} w_{i_2} \ldots w_{i_l} \rangle$ and $\mathcal{S}_2 = \langle w_{j_1} w_{j_2} \ldots w_{j_l} \rangle$ be two arbitrary terms. We define $\mathcal{S}_1 \cup \mathcal{S}_2 = \langle w_{i_1} \ldots w_{i_l} w_{j_1} \ldots w_{j_l} \rangle$ as a new term. For example, if $\mathcal{S}_1 = \langle \text{Tom loves Jerry} \rangle$ and $\mathcal{S}_2 = \langle \text{Jerry loves Tom} \rangle$, then $\mathcal{S}_3 = \mathcal{S}_1 \cup \mathcal{S}_2 = \langle \text{Tom loves Jerry Jerry loves Tom} \rangle$. Additionally, if two terms $\mathcal{S}_a$ and $\mathcal{S}_b$ exist, such that $\mathcal{S}_2 = \mathcal{S}_a \cup \mathcal{S}_1 \cup \mathcal{S}_b$, we then say $\mathcal{S}_1$ is included in $\mathcal{S}_2$ and write $\mathcal{S}_1 \subset \mathcal{S}_2$. It is remarkable that we allow $\mathcal{S}_a$ and/or $\mathcal{S}_b$ to be empty terms. For example, if $\mathcal{S}_1 = \langle \text{loves Jerry} \rangle$ and $\mathcal{S}_2 = \langle \text{Tom loves Jerry} \rangle$, then $\mathcal{S}_1 \subset \mathcal{S}_2 = \mathcal{S}_a \cup \mathcal{S}_1 \cup \mathcal{S}_b$, where $\mathcal{S}_a = \langle \text{Tom} \rangle$ and $\mathcal{S}_b = \langle \emptyset \rangle$.

Assume that we have a total of n documents generated by a word dictionary $\mathcal{W}$. A document is also a sequence of words and hence can be regarded as a term. Let $\mathcal{C} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n\}$ denote the collection of documents. Each document $\mathcal{S}_i$ is paired with a univariate and continuous response variable $Y_i \in \mathbb{R}^1$. The goal is to study the association between $\mathcal{S}_i$ and Y_i.

We first extract structured information from a highly unstructured $\mathcal{S}_i$. One popular method to structuralize text documents is the vector space models (Salton 1989; Belew and Rijsbergen 2000). In a traditional vector space model, each document $\mathcal{S}_i$ can be represented by a d-dimensional vector, where the kth element $(1 \leq k \leq d)$ represents whether word $w_k^* \in \mathcal{W}$ appears in $\mathcal{S}_i$ or not. By doing so, the linear regression model can be specified as

$$Y_i = \tilde{\beta}_0 + \sum_{1 \leq k \leq d} \tilde{\beta}_k \mathcal{I}(w_k^* \in \mathcal{S}_i) + \varepsilon_i, \tag{1}$$

where $\tilde{\beta}_k$ is the corresponding regression coefficient of word w_k^*, $\mathcal{I}(\cdot)$ is an indicator function representing whether the word w_k^* is in $\mathcal{S}_i$, and ε is the random noise with mean 0.

The above model maintains word exchangeability, that is, it only involves frequencies of words but not the order. However, word orders can be quite crucial. For example, the two sentences "Tom loves Jerry" and "Jerry loves Tom" result in the same word vectors, but they have a different word order and therefore express different meanings, which Equation (1) cannot discover. Therefore, when we use "word" as the basic unit in vector space model, we inevitably ignore the order of words, and the underlying model in Equation (1) becomes much less informative.

Thus, to retain the original word order in $\mathcal{S}_i$, it is preferred to use terms and the sequence of words to represent the document. We assume q is the predefined maximum length of terms. Let $\mathcal{T}_q = \{\mathcal{S}_1^*, \ldots, \mathcal{S}_p^*\}$ be a dictionary of p distinct terms, where $\mathcal{S}_j^* \in \mathcal{W}$ and $1 \leq \tau(\mathcal{S}_j^*) \leq q$ for any $1 \leq j \leq p$. It is remarkable that p is typically much larger than d since $p \propto d^q$, where "$\propto$" stands for "proportional to." In this context, each document $\mathcal{S}_i$ can be represented by a p-dimensional vector, where the jth element $(1 \leq j \leq p)$ represents whether term $\mathcal{S}_j^* \in \mathcal{T}_q$ appears in $\mathcal{S}_i$ or not. Thus, the model becomes

$$Y_i = \beta_0 + \sum_{1 \leq j \leq p} \beta_j \mathcal{I}(\mathcal{S}_j^* \subset \mathcal{S}_i) + \varepsilon_i. \tag{2}$$

Model (2) takes the word order into consideration by using terms, rather than words. Thus, it is the true structural model. For real analysis, the cardinality of $\mathcal{T}_q$ can be quite large and typically larger than the sample size n, even with reduced word dictionary $\mathcal{W}$. This motivates us to investigate an efficient method to select relevant terms to the response.

2.2. A Sequential Text-Term Selection Method

Let $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ be the response vector. We define $X_{ij} = \mathcal{I}(\mathcal{S}_j^* \subset \mathcal{S}_i)$ for any $1 \leq i \leq n$ and $1 \leq j \leq p$. Let $\boldsymbol{X}_i = (1, X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^{p+1}$, $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top = (\mathbf{1}, \mathbb{X}_{(1)}, \ldots, \mathbb{X}_{(p)}) \in \mathbb{R}^{n \times (p+1)}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^{p+1}$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$. Then, we can rewrite model (2) as

$$\mathbb{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3}$$

We define the full model as $\mathcal{F} = \{1, \ldots, p\}$, the true model as $\mathcal{F}_1 = \{j : \beta_j \neq 0\}$, and its complement as $\mathcal{F}_0 = \{j : \beta_j = 0\}$. We first split the entire term space $\mathcal{T}_q$ into subspaces of different term lengths. Specifically, we let $\mathcal{L}^{(m)} = \{j : \tau(\mathcal{S}_j^*) = m\}$ denote the indices of all terms with length m in $\mathcal{T}_q$. Then, we use $\mathcal{T}_{q(\mathcal{L}^{(m)})} = \{\mathcal{S}_j^* : j \in \mathcal{L}^{(m)}\}$ to denote the set of terms corresponding to $\mathcal{L}^{(m)}$. By doing so, the entire term space $\mathcal{T}_q$ can be split into q subspaces, that is, $\mathcal{T}_q = \bigcup_{1 \leq m \leq q} \mathcal{T}_{q(\mathcal{L}^{(m)})}$. Additionally, $\mathcal{F} = \bigcup_{1 \leq m \leq q} \mathcal{L}^{(m)}$. The terms in each subspace $\mathcal{T}_{q(\mathcal{L}^{(m)})}$ would be examined consecutively.

Let $\mathcal{F}_c^{(m)}$ collect the indices of all terms under consideration in the mth step, $\mathcal{F}_r^{(m)}$ collect the indices of relevant terms selected in the mth step, and $\mathcal{F}^{(m)}$ denote the resulting model in the mth step. Then, the detailed algorithm is given below.

Step 1 (*Initialization*). Set $\mathcal{F}_c^{(0)} = \mathcal{F}_r^{(0)} = \mathcal{L}^{(1)}$ and $\mathcal{F}^{(0)} = \emptyset$.

Step 2 (*Sequential selection*). After the $(m-1)$th step $(1 \leq m \leq q)$, we obtain $\mathcal{F}_c^{(m-1)}$, $\mathcal{F}_r^{(m-1)}$ and $\mathcal{F}^{(m-1)}$. We then proceed to the next step as follows.

(2.1) (*Consideration set*). In the mth step, we first construct $\mathcal{F}_c^{(m)}$ as follows. We define an operator "$\otimes$" acting on two sets of relevant terms $\mathcal{F}_r^{(g)}$ and

$\mathcal{F}_r^{(h)}$. We let $\mathcal{F}_r^{(g)} \otimes \mathcal{F}_r^{(h)} = \{j : \mathcal{S}_j^* = \mathcal{S}_{j_1}^* \cup \mathcal{S}_{j_2}^*, j_1 \in \mathcal{F}_r^{(g)}, j_2 \in \mathcal{F}_r^{(h)}\}$ be the indices of terms that right join the term in $\mathcal{F}_r^{(g)}$ by the term in $\mathcal{F}_r^{(h)}$. Then, the consideration set $\mathcal{F}_c^{(m)}$ is given by

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

(2.2) (*Term selection*). For each j in $\mathcal{F}_c^{(m)}$, compute

$$\hat{\omega}_j = \frac{(\mathbb{X}_{(j)} - \overline{\mathbb{X}}_{(j)})^\top (\mathbb{Y} - \overline{\mathbb{Y}})}{\sqrt{(\mathbb{X}_{(j)} - \overline{\mathbb{X}}_{(j)})^\top (\mathbb{X}_{(j)} - \overline{\mathbb{X}}_{(j)})(\mathbb{Y} - \overline{\mathbb{Y}})^\top (\mathbb{Y} - \overline{\mathbb{Y}})}}, \quad (4)$$

where $\overline{\mathbb{X}}_{(j)}$ and $\overline{\mathbb{Y}}$ are the sample means of $\mathbb{X}_{(j)}$ and $\mathbb{Y}$, respectively. Then, define the screened index set $\mathcal{F}_r^{(m)} = \{j : |\hat{\omega}_j| \text{ ranks among the top } d^{(m)}\}$. Update $\mathcal{F}^{(m)} = \mathcal{F}^{(m-1)} \bigcup \mathcal{F}_r^{(m)}$.

Step 3 (*Solution path*). Iterate Step (2) for q times, which results in a total of q candidate models, $\mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(q)}$. We then collect these models by a solution path $\mathbb{F} = \{\mathcal{F}^{(m)}, 1 \le m \le q\}$, with $\mathcal{F}^{(m)} = \bigcup_{1 \le i \le m} \mathcal{F}_r^{(m)}$.

It is remarkable that in the mth step, we do not consider all terms in $\mathcal{L}^{(m)}$ but construct a consideration set $\mathcal{F}_c^{(m)}$. Given that the size of $\mathcal{L}^{(m)}$ could still be large, we only consider those terms "most likely" to be relevant with the response in the mth step. Intuitively, if a new term is constructed from two relevant terms, then this new term is also likely to be relevant and thus needs consideration. Therefore, we use the relevant terms already selected in previous steps to construct the consideration set in Step (2.1). Furthermore, as to the choice of $d^{(m)}$ in each iteration, we might follow some hard threshold rules, such as $d^{(m)} = [n/\log(n)]$, according to Fan and Lv (2008), or $d^{(m)} = [(n-1)/q]$ for a more conservative size.

Another thing worthy to mention is that $\mathcal{F}^{(m)}$ is not the final model that we use to analyze and make statistical inferences in real applications. We often need to apply a further variable selection method, such as the backward elimination, to refine the screened model $\mathcal{F}^{(m)}$. Model $\mathcal{F}^{(m)}$ mainly serves as a "quick-and-dirty" way to roughly rule out the unimportant terms. Its purpose is to reduce the model size to a moderate scale, typically less than the sample size n, so that the traditional statistical tools can be applied. Therefore, we are safe as long as all of the "truly important information" is still in $\mathcal{F}^{(m)}$ and the final sparse model is recovered by the variable selection technique.

To gain the theoretical insights about the algorithm, we impose the following regularity conditions.

(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 $c_1 > 0$, $\kappa > 0$, $\min_{j \in \mathcal{F}_1} |\omega_j| \ge 2c_1 n^{-\kappa}$.

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

Both of the conditions are typically and frequently used in screening-related literature. Based on these conditions, we explore the screening consistency of the proposed algorithm. Note that it is unrealistic to expect $\mathcal{F}_1 \in \mathbb{F}$, which means that the true model is selected accurately by the solution path $\mathbb{F}$. However, it is possible to have $\mathcal{F}_1 \subset \mathcal{F}^{(m)}$ for some $1 \le m \le q$, which means all of the relevant terms are selected by the solution path. Therefore, we define the solution path $\mathbb{F}$ to be *screening consistent* if

$$\Pr\left(\mathcal{F}_1 \subset \mathcal{F}^{(m)} \in \mathbb{F}, \text{for some } 1 \le m \le q\right) \to 1. \quad (5)$$

Theorem 1. Under conditions (A1) and (A2), the proposed algorithm possesses the screening consistency defined by (5) for $p = \mathcal{O}(\exp(n^a))$ and $|\mathcal{F}_1| = o(n)$ for some $a > 0$ and $b > 0$, where $|\mathcal{F}_1|$ is the cardinality of $\mathcal{F}_1$.

Theorem 1 guarantees that under some mild conditions, the proposed algorithm would not miss any truly important terms with an overwhelming probability. This builds up a solid basis for further term selection based on the screened model. The screening consistency is empirically verified by the following simulation studies.

3. SIMULATION STUDY

3.1. Simulation Design

The simulation studies are based upon a real textual dataset—consumer reviews for cellphones, collected from Jingdong, one of the largest B2C online retailers in China. The detailed description of the data can be seen in Section 4. Since the consumer reviews are feedback about consumer experiences with products and services, their content can help to understand consumer preference.

In the simulation studies, we extract four high-frequency words—w_1 = battery, w_2 = logistics, w_3 = durable and w_4 = fast. Consumer reviews containing at least one of the four words are selected as the collection of documents $\mathcal{C}$, which results in a total of 22,487 documents. After data preprocessing steps (details shown in Section 4), the remaining words construct a dictionary $\mathcal{W}$ with size $d = 15{,}325$. The term dictionary $\mathcal{T}_q$ is generated from $\mathcal{W}$ with the size $p = d + d^2 + \cdots + d^q$. Here, we set $q = 3$, and hence, p is approximately 10^{12}. Without a loss of generality, we let the first six terms in $\mathcal{T}_q$ be $s_1 = \langle\text{battery}\rangle$, $s_2 = \langle\text{logistics}\rangle$, $s_3 = \langle\text{durable}\rangle$, $s_4 = \langle\text{fast}\rangle$, $s_5 = \langle\text{battery durable}\rangle$ and $s_6 = \langle\text{logistics fast}\rangle$. Then, we consider the following three simulation settings to evaluate the performance of the sequential term selection method.

Setting 1. $$Y_i = -1.5 + \mathcal{I}(s_1 \subset \mathcal{S}_i) + \mathcal{I}(s_2 \subset \mathcal{S}_i) + \varepsilon_i \quad (6)$$

Setting 2. $$Y_i = -2 + 1.5\mathcal{I}(s_5 \subset \mathcal{S}_i) + 1.5\mathcal{I}(s_6 \subset \mathcal{S}_i) + \varepsilon_i \quad (7)$$

Setting 3. $$Y_i = -1 + 0.5\mathcal{I}(s_1 \subset \mathcal{S}_i) + 0.5\mathcal{I}(s_3 \subset \mathcal{S}_i) + \mathcal{I}(s_5 \subset \mathcal{S}_i) + \varepsilon_i \quad (8)$$

Simulation *Setting 1* contains only two terms of length one. *Setting 2* is more complicated by involving longer terms; and

the relevant terms s_5 and s_6 could be highly correlated with irrelevant ones (i.e., s_1–s_4). In *Setting 3*, terms of different lengths are considered. Since s_1 and s_3 are subterms of s_5, there exists high correlations between s_5 and the other two terms.

In the three simulation settings, $\mathcal{S}_i$ is the ith consumer review, Y_i is a continuous variable associated with $\mathcal{S}_i$, and $\varepsilon_i \sim N(0, \sigma^2)$. σ^2 is chosen to achieve different signal-to-noise ratios represented by the theoretical R^2—30%, 50%, and 70%. For each simulation setting, we create $T = 200$ simulated datasets from $\mathcal{C}$. For the tth ($1 \le t \le T$) dataset, we first sample $n = 100, 300, 500$ consumer reviews from $\mathcal{C}$, and calculate $Y_{i(t)} (1 \le i \le n)$.

We conduct the sequential term selection method on each simulated dataset with $q = 3$. For the choice of $d^{(m)}$, we try $d^{(m)} = \kappa * [n/\log(n)]$ for various positive integers κ, and the results are quite similar. Thus, we only report on $\kappa = 1$ to save space. After obtaining the solution path, we can further conduct the backward regression to recover the final sparse model. In this article, the model is further selected by using a backward regression with extended BIC criterion (Chen and Chen 2008),

$$\text{BIC}(M) = \log\{\hat{\sigma}^2_{(M)}\} + n^{-1}|M|(\log n + 2\log|M|), \quad (9)$$

where M represents a model and $|M|$ is its cardinality.

3.2. Evaluation Criteria

Let $\hat{\boldsymbol{\beta}}_{(t)} = (\hat{\beta}_{0(t)}, \hat{\beta}_{1(t)}, \ldots, \hat{\beta}_{p(t)})^\top \in \mathbb{R}^{p+1}$ be the estimated coefficients obtained from the tth simulation run. The selected model is denoted by $\hat{\mathcal{F}}_{(t)} = \{j : |\hat{\beta}_{j(t)}| > 0\}$ with size $|\hat{\mathcal{F}}_{(t)}|$. To evaluate the screening consistency, we define

$$\text{Coverage probability} = T^{-1}\sum_{t=1}^{T} \mathcal{I}(\mathcal{F}_1 \subset \hat{\mathcal{F}}_{(t)}). \quad (10)$$

Furthermore, the model selection consistency, meaning that the selected model perfectly recovers the true model, is also evaluated by

$$\text{Percentage of correctly fit} = T^{-1}\sum_{t=1}^{T} \mathcal{I}(\mathcal{F}_1 = \hat{\mathcal{F}}_{(t)}). \quad (11)$$

The correct and incorrect zero rates are also computed.

$$\text{Percentage of correct zeros} = \frac{1}{T(p-p_1)}\sum_{t=1}^{T}\sum_{j=1}^{p}\left\{I(\hat{\beta}_{j(t)} = 0) \times I(\beta_j = 0)\right\}, \quad (12)$$

$$\text{Percentage of incorrect zeros} = \frac{1}{Tp_1}\sum_{t=1}^{T}\sum_{j=1}^{p}\left\{I(\hat{\beta}_{j(t)} = 0) \times I(\beta_j \neq 0)\right\}, \quad (13)$$

where p_1 is the number of relevant terms in the true model.

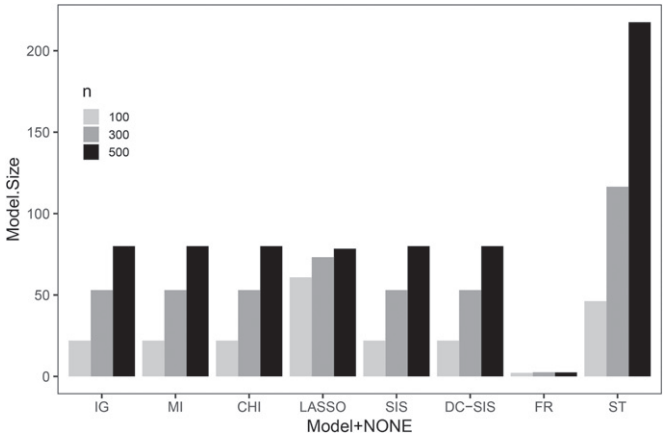

Figure 1. The average model sizes obtained by different methods for "Model+NONE" under *Setting 1* and $R^2 = 50\%$.

3.3. Simulation Results

We compare the following methods: the newly proposed sequential term selection method (ST for short), information gain (IG; Kent 1983), mutual information (MI; Hutter 2001), CHI (Sebastiani 2002), LASSO (Tibshirani 1996), SIS (Fan and Lv 2008), screening based on the distance correlation (DC-SIS; Li, Zhong, and Zhu 2012), and forward regression (FR; Wang 2009). For a given screening method, we use "NONE" to represent the model without using further variable selection and "BACK" to represent the model further selected by the backward elimination. The cardinality of selected models by IG, MI, CHI, SIS, and DC-SIS is fixed to be $[n/\log(n)]$. The tuning parameter λ in LASSO is selected by using 10-fold cross-validation. The extended BIC criterion is applied in FR for model selection from the solution path (Wang 2009).

The detailed simulation results under *Setting 1* are shown in Appendix B to save space. For illustration purpose, we take the theoretical $R^2 = 50\%$ for instance, and compare four key evaluation criteria in Figures 1 and 4. Figure 1 presents the average model sizes achieved by different methods under "Model+NONE." As the sample size n increases, the numbers of terms selected by all methods also increase. For traditional feature selection methods (IG, MI, CHI) and SIS-based selection methods (SIS and DC-SIS), the average model sizes are bounded by $[n/\log(n)]$. LASSO+NONE obtains similar model sizes as the aforementioned methods. As for FR+NONE, since it utilizes the extended BIC as the stopping rule (Wang 2009), the resulted average model size is much smaller than others. Finally, ST+NONE has the largest average model size, since it keeps $[n/\log(n)]$ terms at each term length.

Figure 2 shows the average model sizes under "Model+BACK." As shown in Figure 2, after using the backward elimination, the average model sizes of all methods, except for FR, dramatically decrease. In addition, our proposed method, ST+BACK, is closest to the true model size (the black line in Figure 2). It is also notable that, the average model sizes obtained by FR+BACK are similar with those of FR+NONE. This is because FR+NONE has already obtained a relatively concise term set, and further selection would not help much.

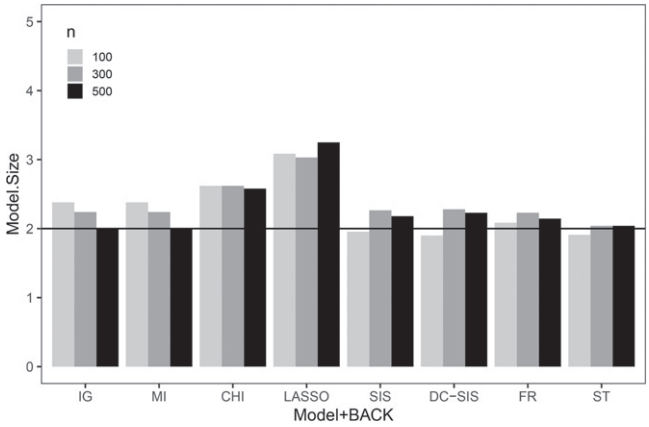

Figure 2. The average model sizes obtained by different methods for "Model+BACK" under *Setting 1* and $R^2 = 50\%$. The black line indicates the true model size, which is equal to 2.

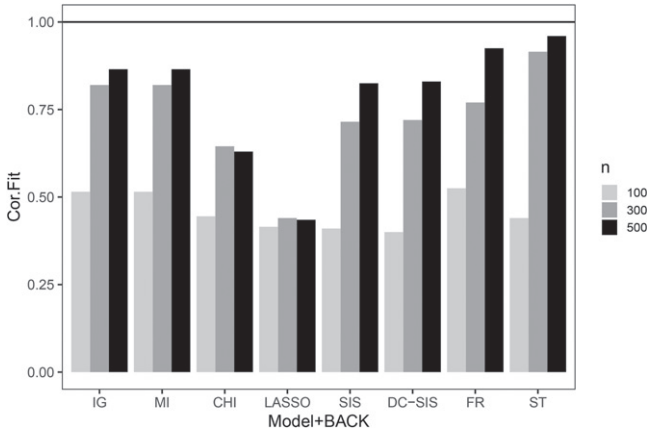

Figure 4. The percentages of correctly fit obtained by different methods for "Model+BACK" under *Setting 1* and $R^2 = 50\%$.

Figure 3 compares the coverage probabilities among different methods for "Model+BACK." Although the average model sizes under "Model+BACK" have dramatically dropped, the coverage probabilities of all methods are not dropped much. As the sample size increases, the coverage probabilities of all methods, especially IG, MI, CHI, LASSO, and ST, approach one, implying the screening consistency. These results guarantee valid candidate model sets for further backward selection.

The detailed simulation results about the average model sizes and coverage probabilities under *Setting 2* and *Setting 3* are provided in Appendix B. The main results generally replicate those obtained from *Setting 1*.

Figure 4 shows the percentages of correctly fit obtained by different "Model+BACK" methods. As discussed, *Setting 1* is a toy setting, containing only two length-one terms. Therefore, all methods can achieve relatively high correct-fit rates, indicating the model selection consistency. Even though, ST+BACK is still the winner among all methods.

Settings 2 and *3*, the correct-fit rates of which depicted by Figure 5, are designed more complicated by involving terms with longer lengths. Thus the model selection consistency is challenging, and the correct-fit rates are quite different among methods. Only FR+BACK and ST+BACK achieve nonzero correct-fit rates, and ST outperforms FR significantly.

In summary, the screening consistency of the proposed sequential term selection method, along with the model selection consistency of the backward regression based on the screened model, is verified empirically in the three simulation settings.

4. REAL DATA ANALYSIS

We demonstrate the proposed method on the following real dataset. We collected 188,107 consumer reviews for 297 cell-

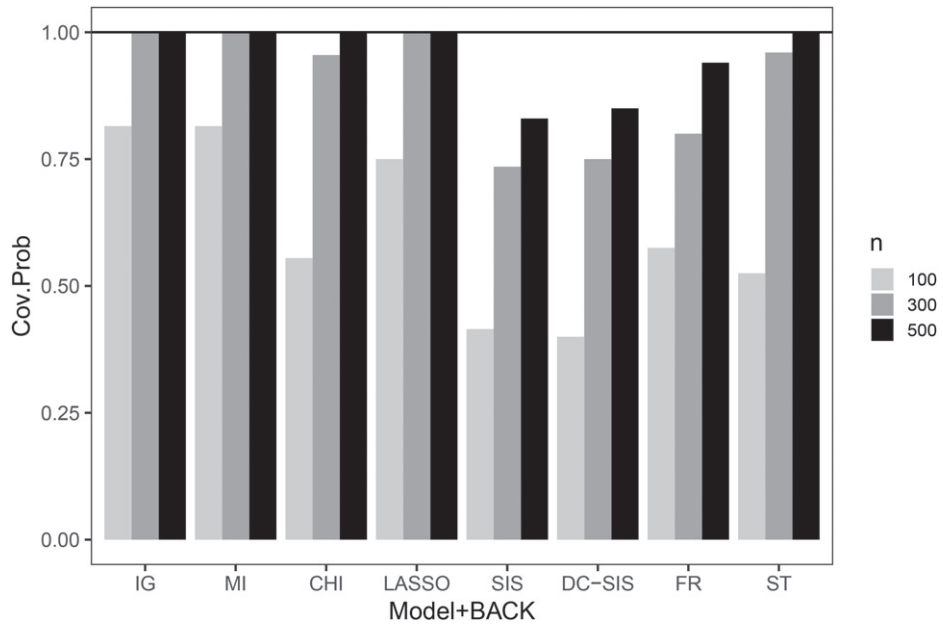

Figure 3. The coverage probabilities obtained by different methods for "Model+BACK" under *Setting 1* and $R^2 = 50\%$.

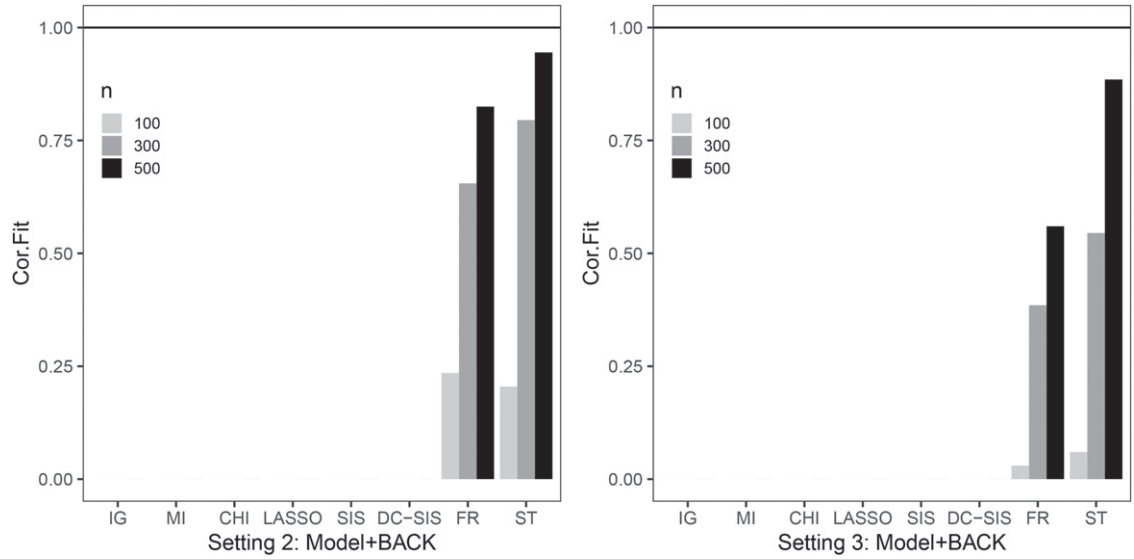

Figure 5. The percentages of correctly fit for "Model+BACK" under *Setting 2* and *Setting 3* for $R^2 = 50\%$.

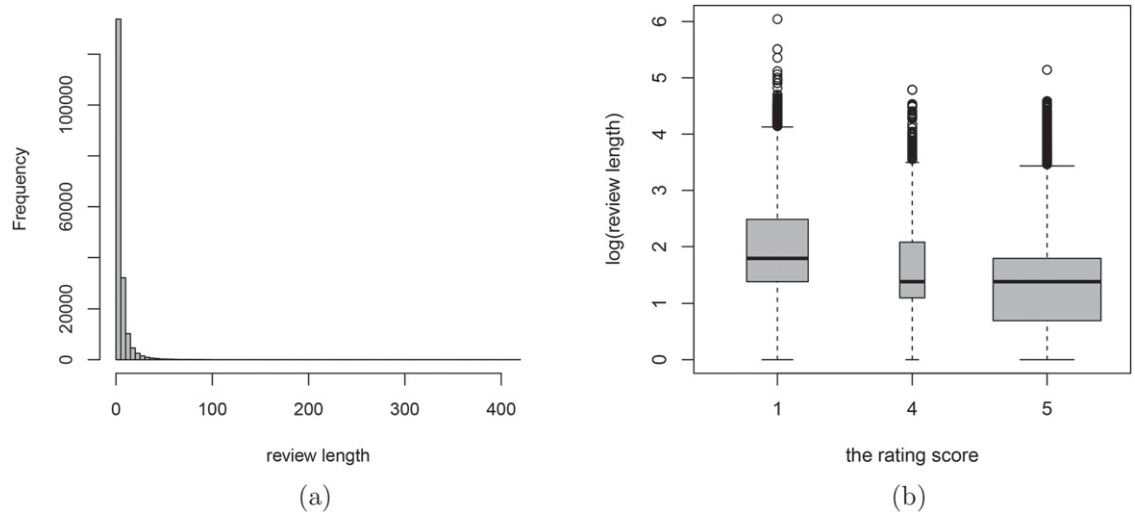

Figure 6. The histogram of review length (a) and the boxplot of review length (in logarithm) under different rating scores (b).

phones on the Jingdong website (*www.JD.com*), which is one of the largest B2C online retailers in China. Each consumer review contained the posting time, the rating score (on a 5-point scale from 1 = awful to 5 = excellent), the rated cellphone, and the full textual content. The user-generated reviews describe the true feelings of consumers concerning the products and services. Therefore, the objective of this study is to discover factors that are correlated with consumers' evaluations on cellphones.

Following common practice in text mining, we first preprocessed the consumer reviews by removing numbers and punctuation. Since the raw text is mainly written in Chinese, we then performed word segmentation by using an open source package *Jieba*. Finally, we remove stop words, which are commonly used but have little semantic meaning on most occasions, such as "the" and "is." After preprocessing, there are 17,088 unique words appearing in all consumer reviews.

We first provide some descriptive analysis concerning the review corpus. The histogram of the review length (i.e., the number of words in each review) is displayed in Figure 6(a). The distribution of review length is highly right-skewed, and the length of most reviews is smaller than 10. Figure 6(b) illustrates distributions of review length under different rating scores. It is clear that the lower the rating score, the longer the reviews. A possible explanation is that consumers who are not satisfied with the products are more willing to describe their shopping experience in detail. It offers insights for discovering crucial factors that are correlated with customers' evaluations from review contents.

To explore the relationship between textual reviews and customer preference, consider the average rating scores of each cellphone as dependent variable, ranging from 1.000 to 5.000 with mean 4.178 and standard deviation 0.668. All terms with length 1, 2, and 3 are included as independent variables. To detect terms that are highly correlated with rating scores, a direct way is to apply the bivariate marginal method by evaluating each term's correlation with the response. Then, we pick up terms whose absolute correlation coefficients are higher than 0.5. This leads to a total of 26 terms, shown in

Table 1. Term selection results by using the bivariate method

Category	Number	Selected terms
Sentiment terms	6	⟨bad⟩, ⟨disappointed⟩, ⟨rubbish⟩, ⟨good⟩, ⟨not good⟩, ⟨too bad⟩,
Product-related terms	9	⟨broken⟩, ⟨signal⟩, ⟨stuck⟩, ⟨display⟩, ⟨system-crash⟩, ⟨starting-up⟩, ⟨bad quality⟩, ⟨bad signal⟩, ⟨open system crash⟩
Service-related terms	11	⟨refund⟩, ⟨after-sale service⟩, ⟨customer-service⟩, ⟨return⟩, ⟨apply⟩, ⟨change⟩, ⟨shopping experience⟩, ⟨rubbish customer-service⟩, ⟨rubbish after-sales⟩, ⟨bad shopping experience⟩, ⟨satisfied buy again⟩

Table 1. We find that the selection results consist of three types of terms: (1) sentiment terms that clearly indicate preferences of customers, such as "bad," "disappointed," and "good"; (2) product-related terms, such as "signal," "display," and "system-crash"; (3) service-related terms, including "refund" and "after-sale service." These empirical results can reflect customers' real concerns.

Then we apply our proposed sequential term selection method on this dataset, and try $q = 1, 2, 3$. For each q, we set the number of terms selected in each step no larger than 100. The results are shown in Table 2. We see clearly that, after using the backward elimination, the number of selected terms are largely reduced. By comparing with the selection results shown in Table 1, we find some common terms, such as "return," "broken," "stuck," and "system-crash." However, the sequential term selection method detect other terms, which are also worthwhile considering. For example, "fake" is a factor that can definitely disappoint customers, and "fast logistics" is a good factor that makes customers satisfied. These findings suggest that the bivariate marginal method and our proposed method can serve as meaningful complements to each other.

Lastly, to further investigate the influence of selected terms, we establish a regression model using the terms selected under $q = 3$. The standardized regression coefficients are present in Table 3. All terms are at 1% level of significance. Furthermore, the exact values of standardized coefficients could help us understand the direction and degree of customer preference. For example, all product-related factors, that is, ⟨mainboard⟩, ⟨battery change⟩, ⟨screen scratch change⟩, are negatively related to the rating scores, which need the most urgent improvement. Among the service-related factors, ⟨logistics fast satisfactory⟩ has positive estimated coefficient, suggesting consumers are quite satisfied with the current logistics. However, the negative estimated coefficient associated with ⟨return⟩ suggests the service of product return needs improvement. In summary, these findings indicate directions for cellphone manufacturers to modify their products or for online retailers to improve their service.

Table 3. Standardized regression results for using selected terms under $q = 3$

Terms	Coefficient	SE	p-value
⟨rubbish⟩	−0.30	0.04	<0.01
⟨return⟩	−0.22	0.03	<0.01
⟨broken⟩	−0.20	0.03	<0.01
⟨good⟩	0.15	0.03	<0.01
⟨mainboard⟩	−0.07	0.03	<0.01
⟨fake⟩	−0.13	0.03	<0.01
⟨deceive⟩	−0.22	0.02	<0.01
⟨battery change⟩	−0.12	0.03	<0.01
⟨screen scratch change⟩	−0.10	0.02	<0.01
⟨logistics fast satisfactory⟩	0.13	0.02	<0.01

5. CONCLUDING REMARKS

In this article, we explore the association between text documents and some continuous response variables. Given that text documents are highly unstructured, vector space models are commonly used to structuralize the textual data. However, these models often lead to a large dictionary, especially when using phrases as the basic analysis unit, and thus is a high-dimensional problem. We therefore investigate a sequential text-term selection method. We first split the whole term space into different subspaces according to the length of terms and then conduct term screening in a sequential manner. That is, in each subspace, only terms that are selected from previous steps will be taken into consideration in the current step. The screening consistency is proven. The empirical performances are illustrated via simulations as well as real data analysis, both

Table 2. Term selection results by using sequential term selection method

	Model size		Terms selected by using backward regression
	NONE	BACK	
$q = 1$	100	9	⟨rubbish⟩, ⟨bad⟩, ⟨return⟩, ⟨broken⟩, ⟨system-crash⟩, ⟨good⟩, ⟨**fake**⟩, ⟨**deceive**⟩, ⟨**clear**⟩
$q = 2$	171	12	⟨rubbish⟩, ⟨bad⟩, ⟨**repair**⟩, ⟨apply⟩, ⟨good⟩, ⟨**fake**⟩, ⟨stuck⟩, ⟨**clear**⟩, ⟨**workmanship**⟩, ⟨**test broken**⟩, ⟨**fast logistics**⟩, ⟨**certified-goods**⟩
$q = 3$	240	10	⟨rubbish⟩, ⟨bad⟩, ⟨broken⟩, ⟨good⟩, ⟨**mainboard**⟩, ⟨**fake**⟩, ⟨**deceive**⟩, ⟨**battery change**⟩, ⟨**screen scratch change**⟩, ⟨**logistics fast satisfactory**⟩

NOTE: Terms only selected by this method are in bold.

based on a dataset of online consumer reviews for cellphones. The results show that the sequential term selection method can select the relevant terms by a few steps. Additionally, by using the backward strategy for further term selection, it can largely reduce the number of screened terms and capture the true ones.

ACKNOWLEDGMENTS

Jingyuan Liu is corresponding author. The authors also want to thank Mr. Xiang Li in school of economics, Xiamen University for his technical support.

FUNDING

This work was supported by funds for building world-class universities (disciplines) of the Renmin University of China, the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 18XNLG02), Basic Scientific Center Project 71988101 of National Science Foundation of China, the National Natural Science Foundation of China (No. 11771361, 11871409, 11671334, 11831008, 11525101, 71532001), JAS14007, and China's National Key Research Special Program (No. 2016YFC0207704).

[Received January 2018. Revised April 2019.]

REFERENCES

Aldous, D. J. (1985), "Exchangeability and Related Topics," in *Ëcole d'Ëtë de Probabilitës de Saint-Flour XIII*, Berlin, Heidelberg: Springer. [83]

Batra, S., Bawa, S., and Punjab, P. (2010), "Using LSI and Its Variants in Text Classification," in *Advanced Techniques in Computing Sciences and Software Engineering*, Dordrecht: Springer, pp. 313–316. [83]

Belew, R. K., and Rijsbergen, C. J. V. (2000), *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*, New York: Cambridge University Press. [82,84]

Berger, J., Sorensen, A. T., and Rasmussen, S. J. (2010), "Positive Effects of Negative Publicity: When Negative Reviews Increase Sales," *Marketing Science*, 29, 815–827. [82]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [83]

Caropreso, M. F., Matwin, S., and Sebastiani, F. (2000), "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07-2000, Pisa, Italy. [83]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection With Large Model Spaces," *Biometrika*, 95, 759–771. [86]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [83,85,86]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With np-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [83]

Genkin, A., Lewis, D. D., and Madigan, D. (2007), "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, 49, 291–304. [83]

Gomez, J. C., and Moens, M. F. (2012), "PCA Document Reconstruction for Email Classification," *Computational Statistics and Data Analysis*, 56, 741–751. [83]

Hofmann, T. (1999), "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57. [83]

Hutter, M. (2001), "Distribution of Mutual Information," *Advances in Neural Information Processing Systems*, 14, 399–406. [86]

Ifrim, G., Bakir, G., and Weikum, G. (2008), "Fast Logistic Regression for Text Categorization With Variable-Length n-Grams," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 354–362. [83]

Jia, J., Miratrix, L., Yu, B., Gawalt, B., El Ghaoui, L., Barnesmoore, L., and Clavier, S. (2014), "Concise Comparative Summaries (CCS) of Large Text Corpora With a Human Experiment," *The Annals of Applied Statistics*, 8, 499–529. [83]

Kent, J. T. (1983), "Information Gain and a General Measure of Correlation," *Biometrika*, 70, 163–173. [86]

Kudo, T., and Matsumoto, Y. (2004), "A Boosting Algorithm for Classification of Semi-Structured Text," in *Conference on Empirical Methods in Natural Language Processing*, pp. 301–308. [83]

Kumar, L., and Bhatia, P. K. (2013), "Text Mining: Concepts, Process and Applications," *Journal of Global Research in Computer Science*, 4, 36–39. [82]

Lee, T. Y., and Bradlow, E. T. (2011), "Automated Marketing Research Using Online Customer Reviews," *Journal of Marketing Research*, 48, 881–894. [82]

Li, J., and Zha, H. (2006), "Two-Way Poisson Mixture Models for Simultaneous Document Classification and Word Clustering," *Computational Statistics and Data Analysis*, 50, 163–180. [83]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [83,86]

Liu, J., Li, R., and Wu, R. (2014), "Feature Selection for Varying Coefficient Models With Ultrahigh Dimensional Covariates," *Journal of the American Statistical Association*, 109, 266–274. [83]

Liu, J., Zhong, W., and Li, R. (2015), "A Selective Overview of Feature Screening for Ultrahigh-Dimensional Data," *Science China Mathematics*, 58, 1–22. [83]

Liu, T., Liu, S., Chen, Z., and Ma, W. Y. (2003), "An Evaluation on Feature Selection for Text Clustering," in *Proceedings of the 20th International Conference on Machine Learning*, pp. 488–495. [83]

Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., and Pfann, G. (2013), "More Than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates," *Journal of Marketing*, 77, 87–103. [82]

Manning, C. D., Raghavan, P., and Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge: Cambridge University Press. [83]

Ng, H. T., Goh, W. B., and Low, K. L. (1997), "Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization," in *Proceeding of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Vol. 31), pp. 67–73. [83]

Salton, G. (1989), *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Boston, MA: Addison-Wesley Longman Publishing Co., Inc. [82,84]

Salton, G., and Buckley, C. (1988), "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing & Management*, 24, 513–523. [83]

Salton, G., Wong, A., and Yang, C. S. (1975), "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, 18, 613–620. [82]

Sebastiani, F. (2002), "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, 34, 1–47. [83,86]

Taddy, M. (2013), "Multinomial Inverse Regression for Text Analysis," *Journal of the American Statistical Association*, 108, 771–772. [83]

Tan, A. H. (1999), "Text Mining: The State of the Art and the Challenges," in *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases* (Vol. 8), pp. 65–70. [82]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [83,86]

Turney, P. D., and Pantel, P. (2010), "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research*, 37, 141–188. [83]

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [83,86]

Wu, S. T., Li, Y., and Xu, Y. (2006), "Deploying Approaches for Pattern Refinement in Text Mining," in *Proceedings of the 6th International Conference on Data Mining*, pp. 1157–1161. [83]

Yang, Y., and Pedersen, J. O. (1997), "A Comparative Study on Feature Selection in Text Categorization," in *Proceedings of the 14th International Conference on Machine Learning*, pp. 412–420. [83]

Zheng, Z., Wu, X., and Srihari, R. (2004), "Feature Selection for Text Categorization on Imbalanced Data," *ACM SIGKDD Explorations Newsletter*, 6, 80–89. [83]

Zhao, Y., Yang, S., Narayan, V., and Zhao, Y. (2013), "Modeling Consumer Learning From Online Product Reviews," *Marketing Science*, 32, 153–169. [82]

Yu, L., and Liu, H. (2003), "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in *Proceedings of the 20th International Conference on Machine Learning*, pp. 856–863. [83]

APPENDIX A: PROOF OF THEOREM 1

Before the proof of Theorem 1, let us introduce some useful lemmas that technically facilitate the proof.

Lemma 1 (Hoeffding's inequality). Suppose that an independent random sample $\{X_i,\ i = 1,\dots,n\}$ satisfies $P(X_i \in [a_i, b_i]) = 1$ for some a_i and b_i, for all $i = 1,\dots,n$. Then, for any $\varepsilon > 0$, we have

$$P(|\overline{X} - E(\overline{X})| \geq \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \tag{A.1}$$

where $\overline{X} = (X_1 + \cdots + X_n)/n$.

Lemma 2. For a random variable X with $Ee^{a|X|} < \infty$ for some $a > 0$, there exist $b > 0$ and $c > 0$ such that for any positive number M, let $P(|X| \geq M) \leq be^{-cM}$.

Lemma 3. Let $\hat{A}$ and $\hat{B}$ be estimates of $a, b \in (-\infty, \infty)$, respectively, based on a sample with size n. Suppose that for any $\varepsilon \in (0,1)$ and some $\xi > 0$,

$$P(|\hat{A} - a| \geq \varepsilon) \leq C_A \exp(-\varepsilon n^{\xi}),\ \ P(|\hat{B} - b| \geq \varepsilon) \leq C_B \exp(-\varepsilon n^{\xi})$$

for some positive constant C_A and C_B. Let z be ab, a^2, $a - b$, a/b, and $\sqrt{b}$, respectively, and $\hat{Z}$ be its corresponding estimates, $\hat{A}\hat{B}$, $\hat{A}^2$, $\hat{A} - \hat{B}$, $\hat{A}/\hat{B}$, or $\sqrt{\hat{B}}$. Then, if z is well defined, we have

$$P\left(|\hat{Z} - z| \geq \varepsilon\right) \leq C_Z \exp\left(-\frac{\varepsilon n^{\xi}}{C_Z}\right)$$

for some $C_Z > 0$.

Proof of Theorem 1. We divide the proof into the following two steps.

(1) Prove the concentration inequality of $\hat{\omega}_j$.

Recall that $\hat{\omega}_j$ defined in (4) can be written as

$$\hat{\omega}_j = \frac{\overline{\mathbb{X}_j\mathbb{Y}} - \overline{\mathbb{X}}_j\overline{\mathbb{Y}}}{\sqrt{\left(\overline{\mathbb{X}_j^2} - \overline{\mathbb{X}}_j^2\right)\left(\overline{\mathbb{Y}^2} - \overline{\mathbb{Y}}^2\right)}}, \tag{A.2}$$

where

$$\overline{\mathbb{X}_j\mathbb{Y}} = \frac{1}{n}\sum_{i=1}^n X_{ij}Y_i, \quad \overline{\mathbb{X}_j^2} = \frac{1}{n}\sum_{i=1}^n X_{ij}^2, \quad \text{and} \quad \overline{\mathbb{Y}^2} = \frac{1}{n}\sum_{i=1}^n Y_i^2.$$

First, obtain the concentration inequality of $\overline{\mathbb{X}_j\mathbb{Y}}$. For any $\delta > 0$ and any $M > 0$,

$$\Pr(|\overline{\mathbb{X}_j\mathbb{Y}} - E(X_jY)| > \delta) = I_1 + I_2,$$

where $I_1 = \Pr(|\overline{\mathbb{X}_j\mathbb{Y}} - E(X_jY)| > \delta, \max_{1\leq i\leq n}|X_{ij}Y_i| \leq M)$, and $I_2 = \Pr(|\overline{\mathbb{X}_j\mathbb{Y}} - E(X_jY)| > \delta, \max_{1\leq i\leq n}|X_{ij}Y_i| > M)$. By Hoeffding's inequality in Lemma 1, $I_1 \leq 2\exp\{-\delta^2 n/(2M^2)\}$. Moreover, since $X_{ij} = \mathcal{I}(\mathcal{S}_j^* \subset \mathcal{S}_i) \in \{0,1\}$,

$$\begin{aligned} I_2 &\leq \Pr(\max_{1\leq i\leq n}|X_{ij}Y_i| > M) \leq \Pr(\max_{1\leq i\leq n}|X_{ij}| \cdot \max_{1\leq i\leq n}|Y_i| > M) \\ &\leq \Pr(\max_{1\leq i\leq n}|Y_i| > M) = \Pr(|Y_i| > M \text{ for some } i = 1,\dots,n) \\ &\leq n\Pr(|Y_i| > M) \leq nb_2\exp(-c_2M). \end{aligned}$$

The last inequality is due to condition (A2) and Lemma 2. Therefore,

$$\begin{aligned} \Pr(|\overline{\mathbb{X}_j\mathbb{Y}} - E(X_jY)| > \delta) &\leq 2\exp\{-\delta^2 n/(2M^2)\} \\ &+ b_2\exp\{-(c_2M - \log n)\} \leq 2\exp\left\{-\delta\cdot\frac{\delta n}{2M^2}\right\} \\ &+ b_2\exp\left\{-\delta\cdot\frac{c_2M - \log n}{\delta}\right\}. \end{aligned}$$

Take $M = \mathcal{O}(n^{\gamma})$, where $0 < \gamma < 1/2$; then, for all $j = 1,\dots,p$,

$$\begin{aligned} \Pr(|\overline{\mathbb{X}_j\mathbb{Y}} - E(X_jY)| > \delta) &\leq 2\exp\left\{-\delta n^{1-2\gamma}\right\} \\ &+ b_2\exp\left\{-\delta n^{\gamma}\right\} = \mathcal{O}(\exp\{-\delta n^{\xi}\}), \end{aligned}$$

where $\xi = \min(\gamma, 1 - 2\gamma)$. Therefore,

$$\max_{j=1,\dots,p}\Pr(|\overline{\mathbb{X}_j\mathbb{Y}} - E(X_jY)| > \delta) \leq \mathcal{O}(\exp\{-\delta n^{\xi}\}).$$

In the same fashion, we can derive the concentration inequalities for $\overline{\mathbb{X}}_j$, $\overline{\mathbb{X}_j^2}$, $\overline{\mathbb{Y}}$, and $\overline{\mathbb{Y}^2}$. By Lemma 3 and (A.2), for some $c > 0$, we have

$$\max_{j=1,\dots,p}\Pr(|\hat{\omega}_j - \omega_j| > \delta) \leq \mathcal{O}(\exp\{-\delta n^{\xi}/c\}). \tag{A.3}$$

(2) Prove the screening consistency.

We decompose the true model $\mathcal{F}_1$ by $\mathcal{F}_1 = \bigcup_{1\leq m\leq q}\mathcal{F}_1^{(m)}$, where $\mathcal{F}_1^{(m)} = \{j : j \in \mathcal{F}_1, \tau(\mathcal{S}_j^*) = m\} = \mathcal{F}_1 \cap \mathcal{L}^{(m)}$ denotes the true term index set with term length m. Then, the coverage probability

$$\begin{aligned} &\Pr\left(\mathcal{F}_1 \subset \mathcal{F}^{(m)} \in \mathbb{F}, \text{for some } 1 \leq m \leq q\right) \\ &\quad = \Pr\left(\mathcal{F}_1^{(m)} \subset \mathcal{F}_r^{(m)}, \text{ for all } m = 1,\dots,q\right). \end{aligned}$$

By the law of total probability,

$$\begin{aligned} \Pr\left(\mathcal{F}_1^{(m)} \subset \mathcal{F}_r^{(m)}, m = 1,\dots,q\right) &= \Pr\left(\mathcal{F}_1^{(1)} \subset \mathcal{F}_r^{(1)}\right) \\ &\cdot \Pr\left(\mathcal{F}_1^{(2)} \subset \mathcal{F}_r^{(2)} | \mathcal{F}_1^{(1)} \subset \mathcal{F}_r^{(1)}\right) \\ &\cdot \Pr\left(\mathcal{F}_1^{(3)} \subset \mathcal{F}_r^{(3)} | \mathcal{F}_1^{(1)} \subset \mathcal{F}_r^{(1)}, \mathcal{F}_1^{(2)} \subset \mathcal{F}_r^{(2)}\right) \cdots \\ &\cdot \Pr\left(\mathcal{F}_1^{(q)} \subset \mathcal{F}_r^{(q)} | \mathcal{F}_1^{(m)} \subset \mathcal{F}_r^{(m)}, m = 1,\dots,q-1\right). \end{aligned} \tag{A.4}$$

First, consider $\Pr\left(\mathcal{F}_1^{(1)} \subset \mathcal{F}_r^{(1)}\right)$. To facilitate the proof, we redefine the screening criterion as $\{j : |\hat{\omega}_j| > \alpha_n\}$, for some $\alpha_n = c_1 n^{-\kappa}$ and $\kappa < \xi$. Then,

$$\begin{aligned} \Pr\left(\mathcal{F}_1^{(1)} \subset \mathcal{F}_r^{(1)}\right) &= \Pr(\min_{j\in\mathcal{F}_1^{(1)}}|\hat{\omega}_j| > \alpha_n) \\ &= \Pr(\min_{j\in\mathcal{F}_1^{(1)}}|\omega_j| - \min_{j\in\mathcal{F}_1^{(1)}}|\hat{\omega}_j| < \min_{j\in\mathcal{F}_1^{(1)}}|\omega_j| - \alpha_n) \\ &\geq \Pr(\max_{j\in\mathcal{F}_1^{(1)}}|\hat{\omega}_j - \omega_j| < c_1 n^{-\kappa}). \end{aligned} \tag{A.5}$$

The inequality is due to the definition of α_n and condition (A1).

By the concentration inequality (A.3), (A.5) is bounded below by

$$\begin{aligned} \Pr\left(\mathcal{F}_1^{(1)} \subset \mathcal{F}_r^{(1)}\right) &= 1 - \Pr(\max_{j\in\mathcal{F}_1^{(1)}}|\hat{\omega}_j - \omega_j| > c_1 n^{-\kappa}) \\ &\geq 1 - |\mathcal{F}_1^{(1)}| \max_{j=1,\dots,p}\Pr(|\hat{\omega}_j - \omega_j| > c_1 n^{-\kappa}) \\ &\geq 1 - |\mathcal{F}_1| \cdot \mathcal{O}(\exp\{-c_2 n^{\xi-\kappa}\}) \\ &\geq 1 - \mathcal{O}(\exp(-c_2 n^{\xi-\kappa})), \end{aligned} \tag{A.6}$$

where $c_2 = c_1/c$, and the last inequality holds as $|\mathcal{F}_1| = o(n)$. Remember that $\kappa < \xi$; hence, the right-hand side of (A.6) converges to 1 as $n \to \infty$.

Since the algorithm is conducted using marginal information, each conditional probability in (A.4) can be proven in the same fashion. Thus, for a fixed q,

$$\Pr\left(\mathcal{F}_1^{(m)} \subset \mathcal{F}_r^{(m)}, m = 1,\dots,q\right) \geq [1 - \mathcal{O}\{\exp(-c_2 n^{\xi-\kappa})\}]^q \to 1.$$

□

APPENDIX B: THE DETAILED SIMULATION RESULTS

Table B1. Simulation results for "Model+NONE" under *Setting 1*

n	Model (+NONE)	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	22.0	69.5	99.9	10.3	0.0	22.0	95.0	99.9	2.5	0.0	22.0	96.5	99.9	1.8	0.0
	MI	22.0	69.5	99.9	10.3	0.0	22.0	95.0	99.9	2.5	0.0	22.0	96.5	99.9	1.8	0.0
	CHI	22.0	34.5	99.9	33.5	0.0	22.0	60.5	99.9	19.8	0.0	22.0	70.0	99.9	15.0	0.0
	LASSO	60.4	72.0	99.8	14.0	0.0	60.9	72.1	99.8	13.9	0.0	81.8	81.1	99.8	9.5	0.0
	SIS	22.0	18.0	99.9	48.8	0.0	22.0	43.5	99.9	31.3	0.0	22.0	57.0	99.9	23.5	0.0
	DC-SIS	22.0	18.5	99.9	50.3	0.0	22.0	44.0	99.9	32.8	0.0	22.0	64.5	99.9	20.8	0.0
	FR	1.6	42.0	100.0	57.0	38.5	2.2	57.5	100.0	42.5	47.5	2.8	58.0	100.0	42.0	48.0
	ST	44.9	30.0	99.8	38.8	0.0	46.2	52.5	99.8	26.3	0.0	47.2	60.5	99.8	22.0	0.0
300	IG	53.0	97.0	100.0	1.5	0.0	53.0	99.5	100.0	0.3	0.0	53.0	100.0	100.0	0.0	0.0
	MI	53.0	97.0	100.0	1.5	0.0	53.0	99.5	100.0	0.3	0.0	53.0	100.0	100.0	0.0	0.0
	CHI	53.0	90.5	100.0	4.8	0.0	53.0	97.5	100.0	1.3	0.0	53.0	99.0	100.0	0.5	0.0
	LASSO	69.8	100.0	100.0	0.0	0.0	73.2	100.0	100.0	0.0	0.0	81.5	100.0	100.0	0.0	0.0
	SIS	53.0	66.0	100.0	17.8	0.0	53.0	73.5	100.0	13.8	0.0	53.0	73.5	100.0	13.3	0.0
	DC-SIS	53.0	92.0	100.0	4.3	0.0	53.0	98.5	100.0	0.8	0.0	53.0	99.5	100.0	0.3	0.0
	FR	2.3	69.5	100.0	30.5	58.5	2.5	80.0	100.0	20.0	65.5	2.9	89.5	100.0	10.5	71.0
	ST	105.8	91.5	100.0	4.3	0.0	116.5	96.0	100.0	2.0	0.0	124.4	98.5	100.0	0.8	0.0
500	IG	80.0	98.5	100.0	0.8	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	MI	80.0	98.5	100.0	0.8	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	CHI	80.0	96.0	100.0	2.0	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	LASSO	73.3	100.0	100.0	0.0	0.0	78.5	100.0	100.0	0.0	0.0	91.3	100.0	100.0	0.0	0.0
	SIS	80.0	80.5	100.0	10.5	0.0	80.0	83.0	100.0	9.0	0.0	80.0	84.5	100.0	8.0	0.0
	DC-SIS	80.0	99.5	100.0	0.3	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	FR	2.4	82.5	100.0	17.5	71.5	2.5	94.0	100.0	6.0	77.0	2.4	95.5	100.0	4.3	83.5
	ST	194.3	99.0	100.0	0.5	0.0	217.5	100.0	100.0	0.0	0.0	235.8	100.0	100.0	0.0	0.0

NOTE: The average model size, coverage probability (Cov-Prob), percentage of correct zeros (Cor-Zeros), incorrect zeros (Incor-Zeros), and correctly fit (Cor-Fit) are shown. All of the percentages are reported omitting.

Table B2. Simulation results for "Model+BACK" under *Setting 1*

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

NOTE: The average model size, coverage probability (Cov-Prob), percentage of correct zeros (Cor-Zeros), incorrect zeros (Incor-Zeros), and correctly fit (Cor-Fit) are shown. All of the percentages are reported omitting.

Table B3. Simulation results for "Model+NONE" under *Setting 2*

n	Model (+NONE)	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	22.0	98.5	99.9	0.8	0.0	22.0	100.0	99.9	0.0	0.0	22.0	98.5	99.9	0.8	0.0
	MI	22.0	98.5	99.9	0.8	0.0	22.0	100.0	99.9	0.0	0.0	22.0	98.5	99.9	0.8	0.0
	CHI	22.0	88.0	99.9	6.0	0.0	22.0	99.5	99.9	0.3	0.0	22.0	100.0	99.9	0.0	0.0
	LASSO	39.9	90.5	99.9	5.0	2.5	41.6	99.5	99.9	0.3	0.0	45.5	100.0	99.9	0.0	0.5
	SIS	22.0	85.5	99.9	8.5	0.0	22.0	98.0	99.9	1.3	0.0	22.0	100.0	99.9	0.0	0.0
	DC-SIS	22.0	84.5	99.9	9.3	0.0	22.0	98.0	99.9	1.0	0.0	22.0	99.5	99.9	0.3	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	42.4	4.0	99.8	68.8	0.0	54.0	23.0	99.8	43.3	0.0	62.0	49.0	99.8	25.8	0.0
300	IG	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0
	MI	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0
	CHI	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0
	LASSO	41.4	96.0	100.0	2.0	0.0	43.8	98.0	100.0	1.0	0.5	47.7	100.0	100.0	0.0	0.0
	SIS	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0
	DC-SIS	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0
	FR	2.3	52.0	100.0	24.0	49.0	2.4	98.0	100.0	1.0	65.5	2.1	83.0	100.0	8.3	75.0
	ST	125.3	59.0	100.0	20.8	0.0	151.5	86.0	100.0	7.0	0.0	158.3	97.0	100.0	1.5	0.0
500	IG	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	MI	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	CHI	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	LASSO	43.6	95.0	100.0	2.5	0.0	49.0	98.5	100.0	0.8	0.5	55.4	99.5	100.0	0.3	0.0
	SIS	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	DC-SIS	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	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	225.5	82.0	100.0	9.0	0.0	240.9	98.0	100.0	1.0	0.0	240.0	100.0	100.0	0.0	0.0

NOTE: The average model size, coverage probability (Cov-Prob), percentage of correct zeros (Cor-Zeros), incorrect zeros (Incor-Zeros), and correctly fit (Cor-Fit) are shown. All of the percentages are reported omitting.

Table B4. Simulation results for "Model+BACK" under *Setting 2*

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

NOTE: The average model size, coverage probability (Cov-Prob), percentage of correct zeros (Cor-Zeros), incorrect zeros (Incor-Zeros), and correctly fit (Cor-Fit) are shown. All of the percentages are reported omitting.

Table B5. Simulation results for "Model+NONE" under *Setting 3*

n	Model (+NONE)	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	22.0	82.0	99.9	6.3	0.0	22.0	99.0	99.9	0.3	0.0	22.0	98.0	99.9	0.7	0.0
	MI	22.0	82.0	99.9	6.3	0.0	22.0	99.0	99.9	0.3	0.0	22.0	98.0	99.9	0.7	0.0
	CHI	22.0	81.5	99.9	6.3	0.0	22.0	100.0	99.9	0.0	0.0	22.0	100.0	99.9	0.0	0.0
	LASSO	37.3	38.0	99.9	28.2	0.0	44.0	70.5	99.9	11.0	0.5	52.2	96.0	99.9	1.3	0.5
	SIS	22.0	65.0	99.9	13.5	0.0	22.0	96.0	99.9	1.3	0.0	22.0	99.5	99.9	0.2	0.0
	DC-SIS	22.0	74.5	99.9	9.3	0.0	22.0	100.0	99.9	0.0	0.0	22.0	100.0	99.9	0.0	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	44.5	55.5	99.8	30.8	0.0	48.1	89.0	99.8	7.3	0.0	51.4	98.5	99.8	1.0	0.0
300	IG	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0
	MI	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0
	CHI	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0
	LASSO	41.0	74.0	100.0	9.8	0.0	48.5	96.0	100.0	1.3	0.0	58.4	100.0	100.0	0.0	0.0
	SIS	53.0	99.5	100.0	0.2	0.0	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0
	DC-SIS	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0	53.0	100.0	100.0	0.0	0.0
	FR	1.6	6.5	100.0	59.2	6.5	2.7	47.5	100.0	28.2	38.5	3.8	89.5	100.0	3.5	52.5
	ST	102.5	99.5	100.0	0.3	0.0	118.1	100.0	100.0	0.0	0.0	133.4	100.0	100.0	0.0	0.0
500	IG	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	MI	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	CHI	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	LASSO	43.7	90.5	100.0	3.2	0.5	50.6	98.5	100.0	0.5	0.0	64.7	99.5	100.0	0.2	0.0
	SIS	53.0	99.5	100.0	0.2	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	0.0
	DC-SIS	53.0	99.5	100.0	0.2	0.0	80.0	100.0	100.0	0.0	0.0	80.0	100.0	100.0	0.0	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	188.3	100.0	100.0	0.0	0.0	216.7	100.0	100.0	0.0	0.0	240.0	100.0	100.0	0.0	0.0

NOTE: The average model size, coverage probability (Cov-Prob), percentage of correct zeros (Cor-Zeros), incorrect zeros (Incor-Zeros), and correctly fit (Cor-Fit) are shown. All of the percentages are reported omitting.

Table B6. Simulation results for "Model+BACK" under *Setting 3*

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

NOTE: The average model size, coverage probability (Cov-Prob), percentage of correct zeros (Cor-Zeros), incorrect zeros (Incor-Zeros), and correctly fit (Cor-Fit) are shown. All of the percentages are reported omitting.