AStA Advances in Statistical Analysis
https://doi.org/10.1007/s10182-023-00472-0

ORIGINAL PAPER

Lasso-based variable selection methods in text regression: the case of short texts

Marzia Freo[1] · **Alessandra Luati**[2,3]

Received: 11 October 2021 / Accepted: 8 February 2023

Abstract

Communication through websites is often characterised by short texts, made of few words, such as image captions or tweets. This paper explores the class of supervised learning methods for the analysis of short texts, as an alternative to unsupervised methods, widely employed to infer topics from structured texts. The aim is to assess the effectiveness of text data in social sciences, when they are used as explanatory variables in regression models. To this purpose, we compare different variable selection procedures when text regression models are fitted to real, short, text data. We discuss the results obtained by several variants of lasso, screening-based methods and randomisation-based models, such as sure independence screening and stability selection, in terms of number and importance of selected variables, assessed through goodness-of-fit measures, inclusion frequency and model class reliance. Latent Dirichlet allocation results are also considered as a term of comparison. Our perspective is primarily empirical and our starting point is the analysis of two real case studies, though bootstrap replications of each dataset are considered. The first case study aims at explaining price variations based on the information contained in the description of items on sale on e-commerce platforms. The second regards open questions in surveys on satisfaction ratings. The case studies are different in nature and representative of different kinds of short texts, as, in one case, a concise descriptive text is considered, whereas, in the other case, the text expresses an opinion.

Keywords Text mining · Lasso · Variable screening · Stability selection · Latent Dirichlet allocation

✉ Marzia Freo
marzia.freo@ec.europa.eu

[1] European Commission, Joint Research Centre (JRC), Ispra, Italy

[2] Department of Mathematics, Imperial College London, London, UK

[3] Department of Statistics, University of Bologna, Bologna, Italy

Published online: 20 March 2023

1 Introduction

Over the past decades, new technologies and the development of online social platforms have made available to researchers a large amount of text data. Consequently, many studies have been conducted with the aim of exploiting the informative content of digital text. Currently, texts are used as data in a variety of applications in favour of social and economic insights: authorship, sentiment, nowcasting, policy uncertainty, media slant, market definition and other topics, as it is witnessed by the review by Gentzkow et al. (2019). Further studies highlight the contribution of text data to different areas of human life such as politics (Jentsch et al. 2020), public administration (Hollibaugh 2019), education (Ferreira-Mello et al. 2019) and several branches of medical sciences (Luque et al. 2019).

With the focus on marketing and business, Reisenbichler and Reutterer (2018) have recently overviewed the wide range of theoretical and applied research based on text as data and have highlighted the major role played by topic modelling. The latter is a class of unsupervised learning methods developed in a probabilistic setting and capable of clustering text documents in a number of, precisely, topics. The most applied topic model is probably the latent Dirichlet allocation (LDA), referring to the Bayesian model developed by Blei et al. (2003), which, in essence, represents each document as a probability distribution over topics and, on its turn, each topic as a probability distribution over words. LDA is a model-based clustering method, related to finite mixture models. It is recognised to be a flexible and versatile tool to analyse text data, and as such, it has afterwards been extended in multiple variants.

As a matter of fact, when individual texts are very short, say in the range of words from one to thirty, as it is the case of data prevalent on websites, such as titles, image captions, questions in Q&A webpages, LDA might generate topics which are not meaningful. For example, it is recognised that LDA does not perform well when applied to short text fragments, such as microblogging, tweets, headlines and product reviews. This is essentially due to sparsity, as in these cases LDA has too few word co-occurrence information. Several strategies have been proposed to alleviate the problem of data sparsity in short texts, either by combining short documents together, or by employing external resources, such as Wikipedia, to overcome the lack of information or, also, by using alternative models better suited for short texts; see the discussion in Cheng et al. (2014), Jipeng et al. (2019), Tuan et al. (2020), Anderlucci and Viroli (2020) and the many references therein. The overall impression is that there is room to investigate whether alternative approaches are somehow more suitable for telegraphic texts.

As an alternative to the topic modelling approach, we explore the class of supervised learning methods that may provide further complementary knowledge to the widely acknowledged LDA for the analysis of short texts. More specifically, in line with the classification recently introduced by Gentzkow et al. (2019, section 3) the methods used in the paper belong to the class of text regressions.

In the context of text regression, the actual problem that applied researchers have to face is variable selection. It is known that text is inherently high-dimensional and sparse, even when telegraphic. Indeed, the set composed by the union

of words, and eventually n-grams, is generally huge. As each token, be it a single word or an n-gram, is a feature or a potential predictor, statistical analysis requires methods for high-dimensional parameter spaces. Drawing on the literature where the number of predictors, P, is either larger than, or large relative to, the number of observations, N, analyses of text data have been based upon two main ways of restricting the attention to lower-dimensional subspaces: shrinkage and dimension reduction (Friedman et al. 2008). Both methods have been applied in text regression: the lasso penalised regression and its modified version in Nowak and Smith (2017) are attributable to the class of shrinkage methods; the singular value decomposition used by Foster et al. (2013) and the latent semantic indexing in Deerwester et al. (1990) belong to the class of dimension reduction methods. Another stream of the literature has faced the problem of text regression by means of nonparametric methods, such as an extension of neural networks and recursive trees. The latter methods will not be considered in the paper, as it is often the case that they provide results mainly focused on prediction, that may be hard to interpret (see Minaee et al. 2020 and references therein). Rather, with a focus on interpretation, we investigate the potential of text regression methods when variable selection is performed by recurring to variable selection models.

The objective of the paper is to compare different variable selection procedures when regression models are fitted to real, short, text data. The focus is on lasso-based methods, as the lasso (Tibshirani 1996), along with several of its variants, is one among the most commonly applied variable selection methods, thanks to its flexibility and computational feasibility. LDA for topic modelling is also included, as a term of comparison. The relative performance is compared in terms of complexity and quality. Complexity is related to the number of selected variables and is measured through the predictive R^2 index. Quality, on the other hand, is related to the selection of relevant variables that is variables that give value and explanation of the response quantitative variable. As objective measures of relative performance in terms of quality we select the frequency of inclusion and the model class reliance. The former measures how frequently the same variable is selected over bootstrap replications, and will be accompanied by measures of variability. The latter (Fisher et al. 2018) is a core measure of variable importance related to permutation-based importance measures, applied here in the context of text regression. All in all, these three indicators concur to describe the debated concept of interpretability. Indeed, interpretability in the context of machine learning methods has been defined by Doshi-Velez and Kim (2017), based on the Merriam-Webster dictionary, as the "ability to explain or to present in understandable terms to a human". Our view is similar in spirit to the proposal expressed by Margot and Luta (2021), who claim that algorithms with high "predictivity, stability and simplicity" are interpretable in the sense of Doshi-Velez and Kim (2017). Our approach investigates similar aspects of different methods, and as a main novelty, we also focus on the ability to detect relevant variables.

Our perspective is empirical and the analysis is carried out on two real datasets, belonging to two fields of application, present in many digital contexts and whose solutions provide results of interest for a wide range of situations.

The first case study affords the issues of explaining variation of prices of goods within the same category, based on the descriptive text or label provided by producers on e-commerce platforms. Solving this task may provide new insights on hedonic evaluation and on price index measurement. Moreover, whether estimated prices were made available, they would decrease informative asymmetries in the markets, by opening information to consumers.

The second case study is concerned with how open questions inserted in a questionnaire may be informative on the overall satisfaction ratings. Datasets of this type are very common. According to the review in Reisenbichler and Reutterer (2018), they are analysed in a large portion of papers; see also Lange et al. (2022), where a bagging approach to unsupervised sentiment analysis is developed and the relation to unsupervised learning based on embedding methods and lexicons is discussed. This research focus is motivated by the growing interest in using text analytics, either of social or of traditionally collected interviews, in order to gain insights about experience or satisfaction. Here, the main rationale is twofold. Firstly, open questions may raise opinions and point of views that could have not been planned by conventional multiple-choice closed questions. Secondly, learning from words contained in open questions to explain satisfaction ratings may open the way to new methods of survey design, capable of replacing traditionally collected questionnaires with pre-defined attributes that are often expensive and require time to be prepared and filled.

The selection of applications necessarily omits many worthy areas of interest, and we do not have the ambition of exhaustively covering the wide range of possible applications. Although in this paper we do not introduce new methodologies to address the issues of short text modelling, we believe that empirical analyses may shed a great deal of light on the effectiveness of text data as explanatory variables in parametric regression models. As a matter of fact, the methods considered in the paper have often been compared and validated only over simulated data. Besides, the two case studies considered in the paper are inherently different in nature, as, in the first application, a concise descriptive text of attributes is considered, whereas, in the second case study, the short text expresses an opinion.

The paper is organised as follows: Sect. 2 overviews the literature on text regression models through shrinkage methods; in Sect. 3, we illustrate the datasets and the text pre-processing for each case study, while in Sect. 4 the design of the study is described along with few technical details. In Sect. 5, we analyse the results on each case study. Concluding remarks are provided in Sect. 6.

2 Modelling high-dimensional sparse text data

Variable selection in regression analysis is an age-old problem in statistics, which currently encountered a renewed interest due to the increasing availability of high-dimensional data. In sparse settings, the focus is to disentangle few meaningful variables, playing a major role for interpretation purposes, from the redundant and noisy remaining ones. A large variety of methods have been developed for sparse high-dimensional regression, with the majority of applications dealing with research in genomics. Alternative methodologies may be grouped in three main classes, with

several mixed proposals: penalty-based, screening-based and randomisation-based methods.

Penalty-based procedures encourage sparsity by imposing a penalisation on parameters at the estimation stage. Different penalties, on the norm or concave, and different shapes of the penalty give rise to alternative specifications: for instance, lasso (least absolute shrinkage and selection operator, Tibshirani 1996), Ridge (Hoerl and Kennard 1970) and elastic net (Zou and Hastie 2005) impose penalty on the norm with different shape, while non-negative garrote (Breiman 1995) and SCAD (Fan and Li 2001) impose different concave penalties. One of the most commonly applied methods is certainly the lasso, both for its computational feasibility and for its predictive performance. The lasso is not without limitations as it may exhibit poor variable selection results (Bach 2008). Meinshausen and Bühlmann (2006) find that it tends to select noisy variables when the penalty parameter is chosen to optimise prediction and suggest to recur to alternative criteria rather than cross-validation to identify causal predictors. A principle approach to model selection is provided by information criteria, which search for a balance between the maximised likelihood function and the model complexity, by adding a penalty term related to the dimension of the parameter space. Traditional information criteria are the Akaike information criterion (AIC, Akaike 1974), which guarantees predictive performance, and the Schwarz information criterion (BIC, Schwarz 1978), derived under a Bayesian approach and proved to be consistent in a number of circumstances. Unfortunately, when the parameter space is large, BIC has been observed to be too liberal (Bogdan et al. 2004; Broman and Speed 2002). Thus, Berger (unpublished) suggests the generalised information criterion, which refines the choice of prior distribution, and Meinshausen and Bühlmann (2006) propose a data adaptive tuning parameter procedure. Alternatively, the extended BIC (Chen and Chen 2008) adjusts the prior probabilities by adding a further tuning parameter which accounts for the cardinality of models when the number of covariates increase. Compared to the previously developed criteria, it retains high simplicity.

Screening procedures for variable selection rank predictors by relevance; subsequently, the research will reduce dimensionality by selecting the highly ranked predictors and running the standard variable selection lasso (or its variants) over selected variables. Candes and Tao (2007) propose the Dantzig selector, which is the solution to an ℓ_1-regularisation problem and achieves the ideal risk. A further method is the sure independence screening (SIS) by Fan and Lv (2008). The sure screening is a property according to which all important variables survive to the variable screening with probability tending to one, and obviously, it is desirable that each selection model does possess it. In this method, the variable screening relies to the correlation learning, by ranking predictors according to their marginal correlation with the response variable, and filters out those predictors that have low marginal correlation with the response variable. Predictors are considered one by one, independently. As a result, the method reduces the space of predictors. After that, the original problem of estimating the model may be solved with classical estimators.

Another way to tackle high dimensionality is to resort to methods inspired to the idea of randomisation or consensus combination. The rationale of randomisation is

to execute the lasso, or other variable selection algorithms, over repeated samples of the original data, generated by bootstrap or resampling methods, and to average over multiple results, so that the instability of running a selection algorithm only once can be overcome. Several randomisations algorithms have been proposed: Bolasso (Bach 2008) lets lasso select variables over bootstrap replications and keeps the intersection of variables selected over replications; stability selection (Meinshausen and Bühlmann 2010) chooses all the variables that occur in a large section of the resulting selected set. Complementary pairs stability selection (Shah and Samworth 2013) is a variant of stability selection based on a modified bootstrap scheme, which yields improved error bounds and therefore favours the applicability of the methodology. These methods have good consistency properties in terms of variable selection. As a drawback, they are more computationally intensive.

The aim of this study is to identify which of these methods are more effective in selecting the relevant variables. To this purpose, we compare some specifications of the previously presented methods in terms of number of selected variables, by a measure of fit that is the predictive R^2, and importance of the selected variables, in terms of frequency of inclusion and model class reliance. To the best of our knowledge, these alternative variable selection methods have never been compared over regression models based on text as data.

3 Data description and pre-processing

3.1 Case study 1

The first case study focuses on the task of pricing items available on producers' e-commerce platforms. Our specific aim is to model regular fashion prices within a cross-sectional comparison, as we are not primarily interested in price dynamics. Indeed, fashion goods are seasonal products sold over a finite season. In these markets, retailers often use dynamic markdown policies in which an initial retail price is set at the beginning of the season; then, the price is subsequently marked down as the season progresses, in order to minimise the stock-out risk (Soysal and Krishnamurthi 2012). We use data scraped from UK e-commerce websites of four fashion producer brands. The scraping operation collects records available on the four websites, during the week of one season. The brands are chosen to be highly heterogeneous with respect to the average price of items on sales: two of them are large fast fashion chains selling at very low prices and the remaining two brands find their business on the design of enhanced fashion item. Fast fashion and enhanced design brands are comparatively discussed in Cachon and Swinney (2011), and we argue that our analysis can provide material for a further insights on enhanced design versus fast fashion. Each record collects information on price, category, brand, which we have voluntarily excluded from the specification, and a field of description. The

Table 1 Sample from the dataset of case study 1

Description	Brand	Price (GBP)
Knitwear		
Checked Flared Trousers	d	25,99
90's Zip Through Cardigan	c	15
Patch Striped Cotton Jumper	b	175
Colour Block Silk Cashmere Roll-neck Sweater	a	420
Dresses		
Striped Cotton Silk Jumpsuit	a	850
Stretch Denim Shirtdress	b	199
*Black Backless Halter Neck Fishtail Maxi Dress by Club L	c	35
Checked Dress with Floral Embroidery	d	39,99

analyses are carried out for the knitwear and dresses categories. A sample from the dataset is reported in Table 1.[1]

3.1.1 Pre-processing

Before affording the issue of sparse modelling, we process some preliminary steps to reduce the dimensionality and to map raw text into a numerical matrix, the document term matrix (DTM), whose *ij*th element indicates the count of the *j*th word or token in the *i*th document. Some text preparation operations are required in order to process the data and reduce meaningless dimensionality. First, we cancel out non-words elements (like numbers, punctuations and proper names); then, words in a standard English stop-words list are automatically removed, and further contractions, such as don't or it's, or misspellings are excluded manually; finally, words are replaced with their roots through stemming. Eventually, stems displaying sparsity higher than 99% are deleted.

After the pre-processing step, the DTM matrix for the knitwear dataset contains $N = 382$ and $P = 229$ words, the dresses dataset $N = 1110$ and $P = 402$ columns. Note that DTM is high-dimensional in column even if not in the row dimension. In both the knitwear and the dresses datasets, each document is composed, in median, by only three words, and each word appears in less than 1% of documents. As a whole, DTMs are highly sparse, with about 99% of empty cells. Some descriptive statistics are shown in Table 2. The response variable, the price, displays a high standard deviation as compared to the average level and shows an asymmetric distribution. In order to assess the robustness of the results, the analysis has been carried out both on the prices and on the log prices. As the results, in terms of average price variability explained by using the log-transformed data, do not significantly differ from the ones obtained based on the original prices and as interpretation of the results is at the core of the hedonic evaluation, we choose to present the results

[1] The complete datasets are available as supplementary materials.

Table 2 Descriptive statistics for case study 1

	Mean	Median	SD	Min	Max	Most frequent words			
Knitwear									
n. words by document	3.5	3.0	1.1	0.0	9.0	1	sweater	6	knit
%word presence over docs	1.5	0.5	3.5	0.3	38.7	2	jumper	7	cardigan
Price	188.5	49.0	278.8	5.6	1610.0	3	top	8	crop
ln(Price)	4.3	3.9	1.4	1.7	7.4	4	cashmere	9	stripe
Sparsity	98.5					5	cotton	10	cable
Dresses									
n. words by document	3.0	3.0	1.2	0.7	7.0	1	print	6	lace
%word presence over docs	0.7	0.2	1.5	0.1	10.8	2	floral	7	mini
Price	138.2	39	336.2	5.6	3090.0	3	midi	8	petite
ln(Price)	4	3.7	1.1	1.7	8.0	4	maxi	9	club
Sparsity	99.3					5	glamorous	10	bodycon

of the analysis on the prices on their natural scale. It is true that the overall variability is smaller when the logarithmic transform of the data is taken. On the other side, relying on the original prices allows us to express a priori judgments on those that may be relevant influential variables.

3.2 Case study 2

The second case study is related to a section of the Tech Company Employee reviews dataset, downloaded from www.kaggle.com.[2] Data were scraped from www.glassdoor.com. Glassdoor is a website which allows current and former employees to anonymously review companies and also to anonymously submit and view salaries, as well as to search and apply for jobs on the same platform. The analyses are carried out for reviews about a worldwide tech anonymous company. Each record collects information on company, location, date of the review, job, position. Moreover, it collects evaluations on a 1 to 5 scale on overall-ratings and other aspects concerning the job, along with two further open questions on pros and cons. Words deriving by the positive field, pros, are presented preceded by the prefix p while words deriving by the negative significant, cons, are preceded by the prefix c. Our aim is to explain the overall-rating using only the text contained in pros and cons open questions fields as data. A sample from the dataset is reported in Table 3.

We shall treat ratings as a response variable in text regression. Though more specific methods can be applied to interval scales and ordinal variables (see, for instance, Hastie et al. 2009, ch. 14), for sake of interpretation and comparison, and in order to

[2] Specifically, data were retrieved, and processed, on February 2019 from https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews, no longer available on the web. The datasets are available as supplementary material.

Table 3 Sample from the dataset of case study 2

Pros	Cons	Rating
Ratings		
It was great and I loved it	Not enough free snack especially chips	5
Many great people to work with and learn from	Lack of accountability for management	2
Overall, great perks.	You are always expected to perform highly,.	4

They take care of their employees. You get paid top of market which makes it difficult to consider working anywhere else. Flexibility with your schedule. No set start time or end time. Ease of working from home. Transparency of work being done at every level. You get to have insight on a lot. Ability to give and receive feedback and it is a part of the culture

They create that environment of uneasiness. You cannot be human and have bad days or go through personal issues that sometimes affect you, you know, the ones that happen in real life. You literally eat and breathe.cname. Work life balance is based on your team culture and your boss. At the end of the day what your boss says, goes. HR does not do a great job at managing team energy. They do whatever the boss in that team thinks is right. Which is very anti.cname. culture

Table 4 Descriptive statistics for case study 2

Ratings	Mean	Median	SD	Min	Max	Most frequent words			
n. words by document	19.5	15.0	16.7	0.0	139.0	1	c.people	6	p.company
%word presence over docs	1.7	1.1	2.4	0.3	24.9	2	c.company.name	7	p.free
Rating	3.4	4.0	1.4	1.0	5.0	3	p.company.name	8	c.management
ln(Rating)	1.1	1.4	0.5	0.0	1.6	4	p.pay	9	c.time
Sparsity	98.3					5	p.people	10	c.company

effectively perform variable selection, we shall investigate the marginal contribution of attributes by estimating linear regression models. The motivation for a regression analysis on rating is twofold. First, variable selection methods for categorical data are not as developed as methods for continuous variables and thus often do not allow for homogeneous comparisons. Second, ratings are most analysed as quantitative variables, which makes the results of the analysis in the paper to be useful in several related applications.

After a similar text pre-processing to the one described in section 3.1.1, the DTM for the Employees dataset is a $N = 808$ by $P = 1135$ matrix. In this case, we also considered unordered pairs of words with sparsity lower than 99%. In fact, no pair is resulted among the most frequently selected variables; see Table 4. Each document is composed in median by 15 words. Each word, in mean, is present in the 1.7% of documents, and the sparsity of the DTM reaches the 98.3%.

4 Design of the study

4.1 Bootstrap

To evaluate to which extent relevant variables have been detected is a challenging task, as the true model is unknown but in simulations. In practical applications, only one dataset at a time is given and one does not know which variables are truly influential. To mimic the availability of several datasets, we recur to bootstrap replications and resample (five hundred times) by each same unique dataset. Indeed, the bootstrap may yield desirable perturbations similar to those of multiple data sets (Efron and Tibshirani 1998). The analyses have been carried out for each case study.

Bootstrap is performed using the classical approach of resampling with replacement. Each replication of the original dataset is divided into training and hold-out datasets. We expect that duplicated observations may somehow overestimate the performance evaluation of the specifications at hand, but in a manner which, in the same way, we expect to be uniform across methods. Indeed, we have verified that the order of performance across specifications does not change if simple cross-validation without repetition is used, which, in addition, reduces the size of training and hold-out datasets. For each bootstrap dataset, we first select variables over the training dataset, by using the pool of models described in section 4.3. Then, a linear regression model with the selected variables as predictors is estimated over the hold-out dataset by the method of ordinary least squares.

4.2 Criteria

We assess the relative performance of alternative selection models on the basis of the following indicators: the predictive R^2, the inclusion frequency and the model class reliance.

To compare the models in terms of complexity, i.e. of number of variables selected, we consider the predictive R^2. We remark here that we compute the predictive R^2, not because we aim to evaluate the usefulness of selected variables for out-of-sample forecasting, which is out of our scope, but rather to evaluate how relevant selected words are in explaining the response variable when new potentially similar datasets are considered. As a matter of fact, the datasets we analyse are renewed either at any season (in the case study 1 of e-commerce data) or at any further session of Human Resources evaluation (in case study 2 of employees' ratings data) and the goal of our study is to identify which methods allows us to retrieve the relevant drivers of either item prices or employees satisfaction, in any further dataset of this type.

The overall quality of each model is assessed through the inclusion frequency and the model class reliance that are essentially indicators of variable importance, both computed at the variable level and averaged at the specification level.

The inclusion frequency measures how often variables are selected, over bootstrap replications. Variables that present a high number of repetitions have been selected in several bootstrap samples, implying that the model is robust to perturbations of the data set.

The model class reliance (Fisher et al. 2018) measures the extent to which well-performing model within a pre-specified class may rely on a variable of interest for prediction accuracy. Within the class, model reliance is a core measure of variable importance, in that it tells how much an individual prediction model relies on explanatory variables of interest for its accuracy. For each model m, the model reliance of each variable j, denoted as $\mathrm{MR}_{j,m}$, is computed as the ratio between: the loss function associated with the specification evaluated based on the DTM with permuted jth variable (numerator) and the same loss function evaluated based on the original DTM (denominator). By permuting the elements of the jth variable, it is possible to assess the amount of the loss, when the variable itself is rendered uninformative; see section 3 of Fisher et al. (2018) for further details. As a loss function, we consider the residuals standard error. At each bootstrap replication, the jth variable in the training set is permuted and the model reliance is computed. We then obtain the empirical bootstrap model reliance of each jth variable as the average over the bootstrap replications. Eventually, we obtain the highest model class reliance (MCR) that is the upper extreme of the interval which defines the MCR. Note that the MCR is a measure of variable importance in a given dataset.

In summary, we evaluate the performance of alternative variable selection methods on the basis of: (1) their explanatory power out of the training sample, measured in terms of predictive R^2; (2) the bootstrap inclusion frequency of selected variables; and (3) the ability to select important variables for the specific dataset.

4.3 Models

Several variants of lasso are considered that we shall discuss with few more details in section 4.4. Nevertheless, we prefer to introduce them all here to provide a full overview of the methods and models used in the analysis. First, to tune the lasso parameter, we recur to standard criteria used to optimise the predictive performance. We call `lasso-min`[3] the model attained by minimising the cross-validation error, which is recognised to optimise predictive performance. Secondly, we optimise the tuning parameter by recurring to BIC variants: `lasso-bic` is attained by minimising the BIC, while `lasso-ebic05` and `lasso-ebic10` are attained by minimising the extended BIC, with moderate and high model complexity regulated by the tuning parameter $\gamma = 0.5$ and $\gamma = 1.0$, respectively. To evaluate the performance of randomisation, we choose the stability selection method in the variant proposed by Shah et al. (2013), by imposing different thresholds of the selection probability. The results are very similar by changing the selection probability, and here, we present those attained

[3] Note that the `monospace font` refers to the estimated model in the paper. Otherwise, the general class of models of that type is considered. For instance, LDA and lasso denote the class of methods while `lda` and `lasso` denote the specification in the analysis.

with the thresholds 0.7, named `ssmb`. As a screening method, we run the SIS algorithm and consider the performance produced by different number of selected predictors, from 1 to 25, labelled from `sis-k1` to `sis-k25`. Finally, we evaluate the performance of the simple lasso obtained by imposing different number of selected predictors, from 1 to 25, labelled from `lasso-k1` to `lasso-k25`.

The results obtained with lasso-based methods are compared with those attained through the standard LDA analysis, used here as a benchmark and, as such, applied as follows. Over the randomised training sets, we select solutions corresponding to an a priori fixed number of topics, from 1 to 25, labelled from `lda-k1` to `lda-k25`, with hyperparameters set to default values, as discussed in Sect. 4.5.

Lasso-based specifications with an average number of selected variables (over bootstrap replications) are compared to LDA specifications with the same number of components.

4.4 Methodological and computational aspects

We specify the linear regression model

$$Y = X\beta + \varepsilon$$

where $Y \in \mathbb{R}^N$ denotes the response variable, $X \in \mathbb{R}^{N\times P}$ is the DTM matrix, $\beta \in \mathbb{R}^P$ is the vector of regression coefficients and $\varepsilon \in \mathbb{R}^N$ is an independent and identically distributed error term.

At each bootstrap replication, the DTM matrix is partitioned accordingly in $X_1 \in \mathbb{R}^{N_1\times P}$ and $X_2 \in \mathbb{R}^{N_2\times P}$, where N_1 and $N_2 = N - N_1$ denote the number of observations in the training set and in the test set, respectively. In our applications, we have fixed $N_1 = 2/3N$. Note that we have m bootstrap replications of the pair (Y_h, X_h), $h = 1, 2$, i.e. $(Y_h^{(m)}, X_h^{(m)}), m = 1, \dots M$ and $M = 500$, but for ease of notation we drop the superscripts.

In the training set, lasso (Tibshirani 1996) solves the penalised least squares problem

$$\hat{\beta}_\lambda^{lasso} = \arg\min_{\beta\in\mathbb{R}^P} \left\{ \| Y_1 - X_1\beta \|^2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\} \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\lambda > 0$ is a tuning parameter which shrinks to zero some coefficients and, consequently, makes the corresponding variables irrelevant.

In our applications, the tuning parameter is selected based on several methods. One criterion consists in minimising the K-fold cross-validation (Stone 1974), error, with $K = 10$,

$$\mathrm{CV}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i\in k-th} (Y_{1i} - X_{1(i)}\hat{\beta}_{\lambda,-k})^2$$

where Y_{1i} is the generic element of $Y_1 \in \mathbb{R}^{N_1}$, $X_{1(i)} \in \mathbb{R}^P$ denotes the ith row of X_1, the index $k : \{1, \dots, N_1\} \to \{1, \dots, K\}$ indicates the partitions to which each ith

observation is allocated by the randomisation in the kth fold, and $\hat{\beta}_{\lambda,-k}$ is the estimate of β obtained by lasso, without the contribution of the observations in the kth fold.

The Bayesian information criterion (BIC) by Schwarz (1978) is based on minimisation of the following objective function,

$$\text{BIC}(\lambda) = N_1 \log \hat{\sigma}_\lambda^2 + df(\lambda) \log(N_1)$$

where $\hat{\sigma}_\lambda^2 = \frac{1}{N_1} \sum_{i=1}^{N} (Y_{1i} - X_{1(i)} \hat{\beta}_\lambda)^2$ and $df(\lambda)$ is the effective degrees of freedom parameter for which an unbiased and consistent estimator is the number of nonzero coefficients (Zou et al. 2007).

The extended BIC (eBIC) by Chen and Chen (2008) adds an extra penalty term $\gamma \in (0, 1)$ that accounts for the model complexity, summarised by the term $\tau_j = \binom{P}{j}$, where j is the number of covariates considered in the model,

$$\text{eBIC}_\gamma(\lambda) = N_1 \log \hat{\sigma}_\lambda^2 + df(\lambda) \log(N_1) + 2\gamma df(\lambda) \log(\tau_j).$$

Stability selection based on lasso is discussed in detail in Meinshausen and Bühlmann (2010), section 2.2. The key concept is the stability path, given by the probability of each variable to be selected when randomly resampling from the data, over all the values of the regularisation parameter. Specifically, lasso provides estimates of the set of nonzero coefficients as $\hat{S}_\lambda = \{j : \hat{\beta}_{\lambda,j} \neq 0\}$, where $\hat{\beta}_{\lambda,j}$ is an element of $\hat{\beta}_\lambda$ in equation (1). Let I be a random subsample of $\{1, \ldots, N\}$ drawn without replacement. For every set $J \subseteq \{1, \ldots, P\}$, the probability of being in the selected set $\hat{S}_\lambda(I)$ is $\pi_J^\lambda = \mathbb{P}\{J \subseteq S_\lambda(I)\}$. For every variable $j = 1, ..., P$, the stability path is given by the selection probabilities π_j^λ across λ. For a cut-off $\pi_0 \in (0, 1)$ and a set of regularisation parameters Λ, the set of stable variables is defined as $S^{stable} = \{j : \max_{\lambda \in \Lambda} \hat{\pi_j^\lambda} \geq \pi_0\}$. Here we apply the complementary pairs version of stability selection by Shah and Samworth (2013), which improves error control.

As a further criterion, we select the value of λ associated with at least k nonzero coefficients, i.e. we estimate $\hat{\lambda}_k = \arg\min_{\lambda \in \Lambda} \text{ card } \{\hat{S}_\lambda\} \geq k, k = 1, \ldots, 25$.

The sure independence screening (Fan and Lv 2008) is based on correlation learning which filters out the variables that have weak correlation with the response. Let $S_* = \{j : \beta_j \neq 0\}$ denote the true model. SIS selects $S_\xi = \{j : |\omega_j| \text{is among the first} [\xi N] \textit{largest}\}$ where $[\xi N]$ denotes the integer part of ξN, $\xi \in (0, 1)$ and $\omega = X'Y$. SIS enjoys the sure screening properties, i.e. $\mathbb{P}(S_* \subset S_\xi) \to 1$ for $N \to \infty$.

Once the relevant variables have been selected, they form the new DTM matrix $X_2 \in \mathbb{R}^{N_2 \times P^*}$, where P^* denotes the number of relevant variables eventually selected by each procedure, and the regression coefficients are estimated by ordinary least squares as the solution of the system

$$X_2' X_2 \hat{\beta}^* = X_2' Y_2.$$

For the case of the lasso, Belloni and Chernozhukov (2013) discuss some additional assumptions to show that the post-estimation OLS, also referred to as post-lasso, performs at least as well as the lasso itself.

4.5 Comparison with LDA

The results obtained by text regression are compared with those obtained by the unsupervised generative model LDA. We do not exploit the many variants of LDA for short text neither the supervised LDA, as LDA is presented here only as a general benchmark. LDA assumes that each document in the corpus can be described as a probabilistic mixture of T topics, and as an output, LDA provides the probability of document d belonging to topic t, $\mathbb{P}(t|d)$, where $d = 1,...,D$ indicates the number of documents and $t = 1,...T$ indicates the number of topics. In turn, each topic is defined by a probabilistic distribution over the vocabulary of size P; for each topic, the word probability vector $\mathbb{P}(v|t)$ where $v=1,...,P$ indicates the number of words describes how likely it is observing a word conditional on a topic. LDA proceeds through posterior inference of the latent topics given the observed words, and as a conjugate prior to the multinomial distribution, LDA uses Dirichlet priors.

In this analysis, the Dirichlet prior hyperparameters are set as default values ($\alpha = 0.1, \beta = 0.05$) and the model is estimated using collapsed Gibbs sampling, as described in (Jones 2019).

As a next step, in each comparison of LDA to lasso-based specifications, the number of LDA topics, T, is fixed equal to the average number of selected variables (over bootstrap replications), P^*. We generate indicator variables assuming values equal to one in correspondence to each document where the topic displays topic probability, $\mathbb{P}(t|d)$, larger than the average topic probability, $1/T \sum_{t=1...T} \mathbb{P}(t|d)$, slightly modifying the procedure by Schwarz (2018), based on the largest topic probability. The number of indicator variables equals the number of topics, T, which, in its turn, in each comparison, equals the average number of selected variables, $T = P^*$. In that case, X_2 is replaced by $X_2^* \in \mathbb{R}^{N_2 \times P^*}$, which collects the P^* indicator variables. Eventually, the response variable (either ratings o prices) is regressed over the indicator variables derived by topics and collected in X_2^* and the performance is evaluated as for the lasso-based methods.

As far as the selected words comparison is considered, the most frequently selected words over bootstrap replications are considered for Lasso-based specifications. As regard as `lda-k`, the most frequently top P^* terms of each topic in each bootstrap replication are considered after discarding duplicates. Indeed the same word may appear within the top P^* terms of more than a component.

On the one side, we expect that in terms of predictive R^2 the comparison a priori favours LDA specifications, as k components in LDA embed more information than k words alone. On the other side, the comparison of selected words extracted by shrinkage methods to top-terms of LDA components, which certainly sounds more

Table 5 Prices. Number of selected variables and predictive R^2

	Case study 1							
	Knitwear				Dresses			
	N. vars		Pred. R^2		N. vars		Pred. R^2	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
`lasso-min`	53.6	23.7	0.773	0.095	79.1	39.8	0.694	0.139
`lasso-bic`	24.7	6.6	0.698	0.101	28.6	7.4	0.642	0.139
`lasso-ebic05`	19.1	6.1	0.661	0.108	23.6	6.4	0.629	0.143
`lasso-ebic10`	14.2	6.1	0.613	0.120	19.4	5.7	0.610	0.145
`lasso-k k=ebic10`	14.0	0.9	0.624	0.107	18.9	0.8	0.601	0.145
`sis-k k=ebic10`	14.0	0.0	0.625	0.099	19.0	0.0	0.603	0.149
`ssmb`	6.3	1.5	0.499	0.114	9.0	1.7	0.404	0.136
`lasso-k k=ssmb`	6.0	0.7	0.493	0.115	9.0	0.6	0.501	0.16
`sis-k k=ssmb`	6.0	0.0	0.493	0.111	9.0	0.0	0.500	0.159
`lda-k k=ssmb`	6.0	0.0	0.158	0.073	9.0	0.0	0.081	0.030
`lda-k k=ebic10`	14.0	0.0	0.366	0.090	19.0	0.0	0.177	0.048

Averages over bootstrap replications

artificial, may notwithstanding provide useful information on relative ability to detect relevant words.

4.6 Computational details

All the computations are carried out based on the `R` software. In particular, lasso is implemented by the `glmnet` package by Friedman et al. (2010). Stability selection is run using the `ssmb` package by Hofner and Hothorn (2017) and Hofner et al. (2015). Sure independence screening is carried out through the package `SIS` (Saldana and Feng 2018). LDA topic model is fit by using `textmineR` (Jones 2019).

5 Results

5.1 Case study 1: prices

5.1.1 Predictive R^2

The main results of case study 1, in terms of number of selected variables and predictive R^2, are displayed in Table 5 and summarised in Fig. 1.

As expected, `lasso-min` always selects, on average, a very high number of predictors in both categories, to reach the highest adjusted and predictive R^2 and

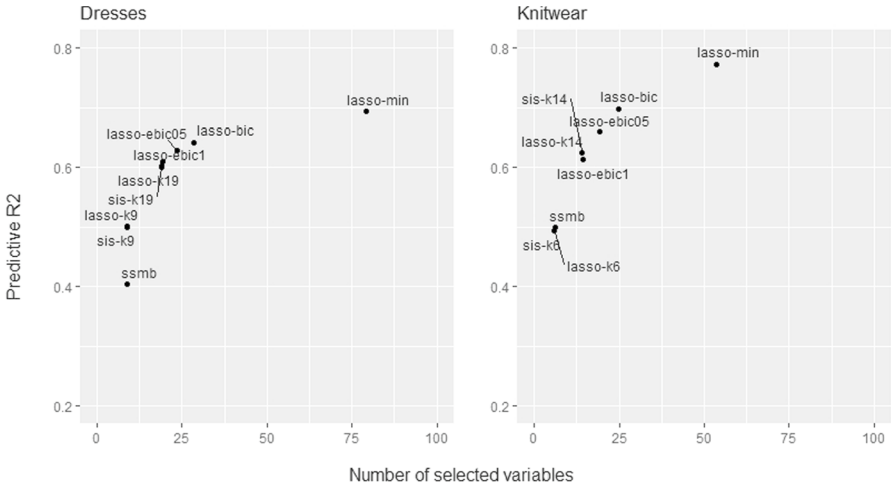

Fig. 1 Prices. Number of selected variables and predictive R^2. Averages over bootstrap replications

by explaining 0.773 of price variation for knitwear and 0.694 for dresses replicated datasets. Indeed, as it has been observed it tends to be generous in selecting noisy variables. The number of predictors selected by `lasso-min` displays the highest standard deviation.

When lasso is optimised by minimising the eBIC, it employs few parameters; the higher the value of the tuning parameter, the smaller the number of selected variables, on average. The most parsimonious eBIC selects, on average, 14.2 predictors for knitwear and 19.4 predictors for dresses, to explain about 61% of the price variability in new datasets. Stability selection is quite parsimonious as well: it selects 6.3 predictors for knitwear and 9.0 for dresses, explaining a share of 0.499 and 0.404 of prices' predictive variance within categories. Predictive R^2 are just lower than the ones of `lasso-ebic10`, though based on quite less predictors. Moreover, it has to be observed that lasso models based on the ebic optimisation tend to select a limited number of coefficients on average, but with a certain degree of heterogeneity over the replications.

For sake of brevity, having results for $k = 1, 2, \ldots, 25$, Table 5 presents results for SIS and lasso computed with the same (average) fixed number of predictors as `ssmb` and `lasso-ebic10`, i.e. $\bar{P}^* = 6$ (knitwear) and $\bar{P}^* = 9$ (dresses) for stability selection and $\bar{P}^* = 14$ (knitwear) and $\bar{P}^* = 19$ (dresses) for lasso optimised with eBIC with $\gamma = 1$.

Focusing on the SIS method, we note that it produces very similar results as compared to `ssmb` and `lasso-ebic10` when a comparable number of predictors are imposed. The same pattern may be observed when lasso is used by fixing the number of selected variables. Figure 1 sketches an overview of presented methods for

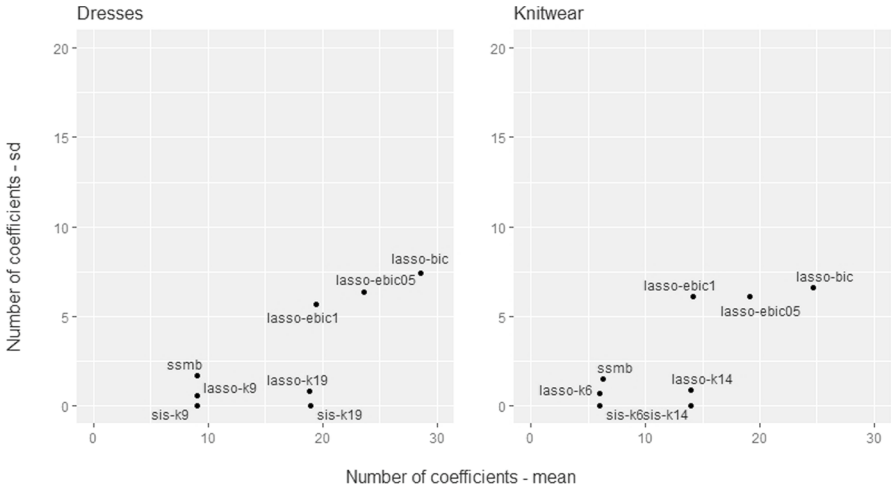

Fig. 2 Prices. Number of coefficients. Averages over bootstrap replication

the two categories. Note that, as expected, the performance in terms of predictive R^2 increases, with diminishing derivative, as long as the number of predictors grows.

Figure 2 displays the standard deviation versus the average number of coefficients for a selection of models, highlighting that the `ssmb` constantly ensures parsimonious models.

A first finding is that the two methods producing the most parsimonious variables selection are `lasso-ebic10` and `ssmb`, where the last one seems to be preferable in terms of robustness. A similar performance is reached by simple lasso or SIS, with the same, fixed, number of predictors.

5.1.2 Inclusion frequency and model class reliance

We now come to the analysis of inclusion frequencies and the model reliance. The analysis covers a selection of models: lasso optimised with eBIC with $\gamma = 1$; stability selection; SIS and lasso computed with the same (average) fixed number of predictors as `ssmb` and `lasso-ebic10`, and LDA with the number of topics equal to the same (average) fixed number of predictors as `ssmb` and `lasso-ebic10`. Tables 6 and 7 display the most occurring variables picked in `lasso-ebic10`, `ssmb` and in SIS or lasso with a priori fixed coefficients over the knitwear and dresses dataset, respectively. For each specification, the tables present the average inclusion frequency and the highest model class reliance (MCR) by word and on average.

Table 6 Selected words and inclusion frequency for the Knitwear dataset

Knitwear											
ssmb	Inclusion	Highest	lasso-k6	Inclusion	High	sis-k6	Inclusion	Highest	lda-k6	Inclusion	Highest
$P^* = 6$	Frequency	MCR		Frequency	MCR		Frequency	MCR		Frequency	MCR
cashmere	0.996	1.262	cashmere	1.000	1.262	cashmere	1.000	1.262	sweater	1.000	1.007
wool	0.810	1.030	turtleneck	0.654	1.066	turtleneck	0.722	1.066	jumper	0.896	1.002
turtleneck	0.626	1.066	wool	0.546	1.030	silk	0.478	1.032	top	0.458	1.010
blend	0.494	1.022	beach	0.456	1.036	wool	0.446	1.030	cashmere	0.270	1.262
sweater	0.406	1.007	belted	0.342	1.005	beach	0.432	1.036	cotton	0.150	1.003
beach	0.388	1.036	blend	0.328	1.022	collar	0.314	1.004	cardigan	0.146	1.003
Mean	0.620	1.071		0.554	1.070		0.565	1.072		0.487	1.048
lasso-ebic1	Inclusion	Highest	lasso-k14	Inclusion	Highest	sis-k14	Inclusion	Highest	lda-k14	Inclusion	Highest
$P^* = 14$	Frequency	MCR		Frequency	MCR		Frequency	MCR		Frequency	MCR
cashmere	1.000	1.262	cashmere	1.000	1.262	cashmere	1.000	1.262	sweater	1.000	1.007
wool	0.868	1.030	wool	0.926	1.033	turtleneck	0.916	1.066	jumper	0.872	1.002
turtleneck	0.842	1.066	turtleneck	0.842	1.066	silk	0.878	1.032	top	0.842	1.010
blend	0.716	1.022	blend	0.786	1.022	wool	0.842	1.030	knit	0.812	1.000
beach	0.682	1.036	beach	0.734	1.036	beach	0.752	1.036	cardigan	0.804	1.007
sweater	0.544	1.007	sweater	0.624	1.007	blend	0.728	1.022	cashmere	0.802	1.262
reissued	0.524	1.013	reissued	0.554	1.013	sweater	0.538	1.007	crop	0.770	1.005
belted	0.474	1.005	belted	0.500	1.005	top	0.510	1.010	stripe	0.742	1.002
cricket	0.450	1.024	top	0.486	1.010	belted	0.506	1.005	cotton	0.644	1.003
squareneck	0.416	1.007	cricket	0.478	1.024	cricket	0.476	1.024	check	0.630	1.001
intarsia	0.404	1.005	intarsia	0.436	1.005	collar	0.434	1.004	silk	0.616	1.032
double	0.392	1.015	squareneck	0.424	1.007	squareneck	0.404	1.007	sleeve	0.576	1.002
moulin	0.346	1.004	double	0.420	1.015	reissued	0.372	1.013	wool	0.530	1.030

Table 6 (continued)

lasso-ebic1	Inclusion	Highest	lasso-k14	Inclusion	Highest	sis-k14	Inclusion	Highest	lda-k14	Inclusion	Highest
$P^* = 14$	Frequency	MCR		Frequency	MCR		Frequency	MCR		Frequency	MCR
saddle	0.316	1.000	moulin	0.388	1.004	double	0.364	1.015	detail	0.470	1.000
Mean	0.570	1.035		0.614	1.036		0.623	1.038		0.707	1.026

Table 7 Selected words and inclusion frequency for the Dresses dataset

Dresses											
`ssmb`	Incl	Highest	`lasso-k9`	Incl	Highest	`sis-k9`	Incl	Highest	`lda-k9`	Incl	Highest
$P^* = 9$	freq	MCR		freq	MCR		freq	MCR		freq	MCR
Silk	0.996	1.093	silk	1.000	1.093	silk	0.970	1.093	Print	0.864	1.000
Cotton	0.988	1.036	Wool	0.856	1.101	Wool	0.822	1.101	floral	0.728	1.000
Wool	0.754	1.101	tulle	0.754	1.124	Tulle	0.714	1.124	lace	0.710	1.007
Crepe	0.714	1.023	cotton	0.728	1.036	Cotton	0.624	1.036	maxi	0.682	1.001
Archive	0.626	1.010	Evening	0.548	1.065	Evening	0.568	1.065	midi	0.680	1.001
Scribble	0.504	1.008	tilly	0.510	1.035	cady	0.566	1.022	glamorous	0.564	1.002
Lace	0.446	1.007	wildflower	0.496	1.026	Charmeuse	0.498	1.032	mini	0.560	1.003
Tulle	0.386	1.124	oreilly	0.494	1.022	Tilly	0.486	1.035	club	0.440	1.003
Crochet	0.364	1.042	charmeuse	0.460	1.032	oreilly	0.460	1.022	Petite	0.428	1.002
Mean	0.642	1.049		0.650	1.059		0.634	1.059		0.628	1.002
`lasso-ebic1`	Incl	Highest	`lasso-k19`	Incl	Highest	`sis-k19`	Incl	Highest	`lda-k19`	Incl	Highest
$P^* = 19$	freq	MCR		Freq	MCR		Freq	MCR		Freq	MCR
Silk	1.000	1.093	Silk	1.000	1.093	Silk	1.000	1.093	Lace	0.904	1.007
Cotton	0.998	1.036	Cotton	0.996	1.036	Cotton	0.978	1.036	Print	0.854	1.000
Wool	0.920	1.101	Wool	0.954	1.101	Wool	0.940	1.101	Floral	0.830	1.000
Crepe	0.810	1.023	Crepe	0.900	1.023	Archive	0.796	1.010	midi	0.808	1.001
Tulle	0.800	1.124	Tulle	0.792	1.124	Crepe	0.782	1.023	Club	0.808	1.003
Crochet	0.754	1.042	Scribble	0.750	1.008	Tulle	0.746	1.124	Maxi	0.804	1.001
Scribble	0.698	1.008	Archive	0.746	1.010	Crochet	0.738	1.042	Glamorous	0.764	1.002
Lace	0.694	1.007	Crochet	0.688	1.042	Cady	0.680	1.022	Mini	0.738	1.003
Archive	0.690	1.010	lace	0.624	1.007	Scribble	0.664	1.008	shirt	0.648	1.011
Cady	0.534	1.022	Evening	0.554	1.065	Shirt	0.600	1.011	Petite	0.628	1.002
Evening	0.534	1.065	Shirt	0.538	1.011	Evening	0.568	1.065	Bodycon	0.610	1.001

Table 7 (continued)

`lasso-ebic1`	Incl	Highest	`lasso-k19`	Incl	Highest	`sis-k19`	Incl	Highest	`lda-k19`	Incl	Highest
$P^* = 19$	freq	MCR		Freq	MCR		Freq	MCR		Freq	MCR
Oreilly	0.498	1.022	Cady	0.530	1.022	Jumper	0.512	1.007	Jersey	0.582	1.002
Charmeuse	0.490	1.032	Oreilly	0.504	1.022	Charmeuse	0.506	1.032	Slip	0.532	1.002
Cottoncash-mere	0.486	1.006	Tilly	0.492	1.035	Oreilly	0.498	1.022	Polka	0.496	1.000
Tilly	0.486	1.035	Cottoncash-mere	0.490	1.006	lace	0.492	1.007	Jumper	0.470	1.007
Wildflower	0.484	1.026	Wildflower	0.472	1.026	logo	0.492	1.003	Front	0.386	1.001
Shirt	0.470	1.011	Jumper	0.462	1.007	Tilly	0.492	1.035	Button	0.348	1.001
Foil	0.436	1.009	Foil	0.456	1.009	Wildflower	0.492	1.026	wrap	0.328	1.000
Georgette	0.430	1.004	Charmeuse	0.454	1.032	Cottoncash-mere	0.316	1.006	sleeve	0.290	1.000
Mean	0.643	1.036		0.653	1.036		0.653	1.035		0.624	1.002

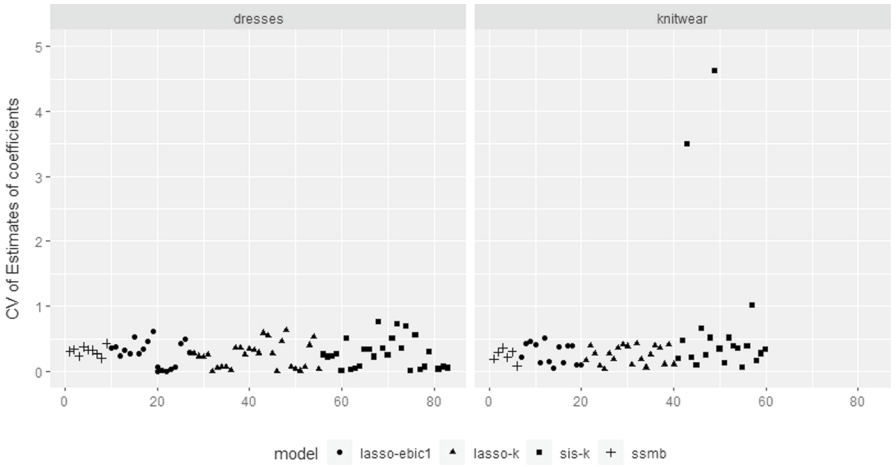

Fig. 3 Prices. CV of estimates of coefficients

The results display levels of mean inclusion frequency comparable across the variable selection methods with some heterogeneity. A similar picture is found in terms of the highest MCR, i.e. comparable levels across lasso-based methods, always higher as compared to LDA results.

In the knitwear dataset (Table 6), MCR shows that, when 6 variables are selected, the gain in the residual standard error amounts to 7.1 percentage points in mean, by 3.5 to 3.8 percentage points when 14 variables are selected.

Table 7 displays selected words and bootstrap inclusion frequencies over the dresses dataset. Text regression methods provide similar levels of average inclusion frequencies. The highest MCR values confirm that lasso-based variable selection methods, more often than LDA, identify words indicating important variables. When 9 words are selected, the MCR increase the gain of about 5 percentage points in mean, in case of selection of 19 words by about 3.5 percentage points in mean.

Eventually, by the analysis of the estimated coefficients and the coefficients of variations, see Fig. 3, it clearly emerges that few of them are detected out of the bulk of data, thus indicating a negative note in terms of model robustness for `sis`, which have selected features whose coefficients result to exhibit a disproportionate variability.

5.1.3 Summary

We may conclude that the results attained through text regression methods always significantly overcome the performance of `lda`, in terms of predictive R^2. Mean inclusion frequencies of `lda` models, on the other hand, are comparable to the ones attained by other methods. At the same time, the ability to select important variables favours text regressions. Overall, for prices datasets, our findings recommend the use of text regression methods. Among the latter, the best performance in terms of the three measures considered throughout the analysis is provided by `ssmb`, i.e.

Table 8 Ratings. Number of selected variables and predictive R^2

	Case study 2			
	Ratings			
	N. vars.		Pred. R^2	
	Mean	SD	Mean	SD
`lasso-min`	171.0	37.4	0.851	0.067
`lasso-bic`	213.5	159.1	0.784	0.223
`lasso-ebic05`	16.8	32.9	0.390	0.087
`lasso-ebic10`	7.0	17.7	0.302	0.067
`lasso-k k=lasso-ebic10`	7.0	0.6	0.319	0.042
`sis-k k=lasso-ebic10`	7.0	0.0	0.305	0.045
`ssmb`	8.8	1.6	0.338	0.049
`lasso-k k=ssmb`	9.0	0.6	0.345	0.043
`sis-k k=ssmb`	9.0	0.0	0.332	0.045
`lda-k k=ssmb`	9.0	0.0	0.390	0.051
`lda-k k=lasso-ebic10`	7.0	0.0	0.383	0.055

Averages over bootstrap replications

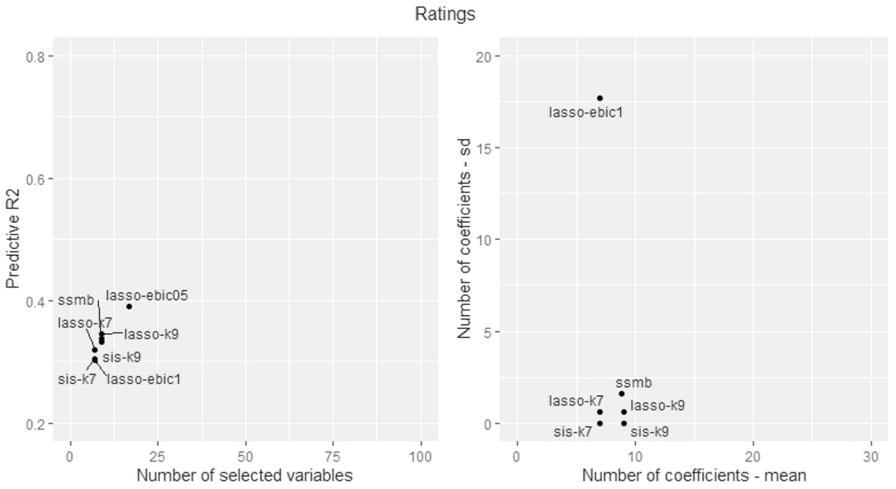

Fig. 4 Ratings. Number of selected variables and predictive R^2 (left). Number of coefficients (right). Averages over bootstrap replications

stability selection, that is both parsimonious and stable, in the sense that it is not affected by high variability in the estimated coefficients.

5.2 Case study 2: ratings

5.2.1 Predictive R^2

Table 8 displays the results related to the second case study, conducted with the aim of explaining ratings with tokens drawn by open questions. In this example, `lasso-bic` tends to select the largest number of variables, capable of explaining around 78.4% of predictive ratings variation. As in the previous example, also in this case, `lasso-min` produces, on average, a high number of predictors, to reach the highest predictive R^2. Figure 4 (left panel) confirms that high predictive power may be reached only at the cost of selecting a very large number of predictors. Note that `lasso-min` and `lasso-bic` are not included in the figure as we have maintained the same scale of Fig. 1, for sake of comparison. More parsimonious models are selected by `ssmb` and `lasso-ebic10`, based on less nine variables or less, even if also lasso optimising the ebic yet displays higher variability for the number of coefficients (see Fig. 4, right panel, in the same scale of Fig. 2). Both the lasso model which minimises eBIC and stability selection guarantee an acceptable trade-off between explanatory power and parsimony. As in case study 1, lasso and SIS perform comparably well as `ssmb` and `lasso-ebic10`, when the number of predictors is a priori fixed, with an out-of-sample R^2 oscillating around 30 to 35%.

Differently than in the case of prices, with ratings, our findings are slightly in favour of topic modelling, as the `lda` predictive R^2, corresponding, respectively, to solutions with 7 and 9 topics, to preserve comparability, are slightly higher than the ones reached through text regression methods.

5.2.2 Inclusion frequency and model class reliance

Table 9 compares the most frequently selected features over the five hundred bootstrap replications by the methods, as well as the highest MCR of selected variables. Words have been drawn by the two open questions asking to indicate, respectively, pros and cons. Words deriving by the positive field, pros, are presented preceded by the prefix p while words deriving by the negative significant, cons, are preceded by the prefix c. The more parsimonious the models, the higher the inclusion frequency of selected variables. For both `ssmb` and `lasso-ebic10`, the most relevant variables are selected by more than 65% of replications. It is evident that in terms of both persistence and relevant variables, `ssmb` and `lasso-ebic10` outperform all the other methods. As far as `lda` is concerned, top selected words display slightly lower mean inclusion frequency rates over the considered methods.

Concerning the ability to detect important variables, note that, in all cases, the highest MCR values only negligibly overcome the unitary value, meaning that the selected variables, on average, are able to decrease the loss in predictive power only of about 1 percentage point.

Table 9 Selected words and inclusion frequency for the ratings dataset

Ratings											
ssmb	Incl	Highest	lasso-k9	Incl	Highest	sis-k9	Incl	Highest	lda-k9	Incl	Highest
$P^* = 9$	freq	MCR		freq	MCR		freq	MCR		freq	MCR
p.culture	0.996	1.032	p.culture	0.998	1.032	p.culture	0.986	1.032	p.free	0.808	1.004
c.management	0.980	1.036	c.management	0.988	1.036	c.management	0.956	1.036	c.people	0.804	1.006
p.freedom	0.882	1.012	p.freedom	0.850	1.012	c.fired	0.742	1.015	p.company	0.710	1.007
c.fear	0.720	1.013	c.fired	0.840	1.015	c.job	0.688	1.007	p.companyname	0.594	1.003
c.fired	0.668	1.015	c.job	0.740	1.007	p.freedom	0.654	1.012	c.companyname	0.526	1.003
p.company	0.508	1.007	c.fear	0.738	1.013	c.fear	0.576	1.013	p.people	0.460	1.004
p.pay	0.442	1.004	p.pay	0.566	1.004	p.pay	0.498	1.004	c.culture	0.460	1.002
c.job	0.430	1.007	c.worst	0.422	1.005	c.worst	0.304	1.005	c.management	0.438	1.036
p.free	0.408	1.004	p.free	0.374	1.004	p.company	0.264	1.007	p.culture	0.408	1.032
Mean	0.670	1.014		0.724	1.014		0.630	1.015		0.588	1.011
lasso-ebic1	Incl	Highest	lasso-k7	Incl	Highest	sis-k7	Incl	Highest	lda-k7	Incl	Highest
$P^* = 7$	freq	MCR		freq	MCR		freq	MCR		freq	MCR
p.culture	0.978	1.032	p.culture	0.990	1.032	p.culture	0.982	1.032	p.free	0.832	1.004
c.management	0.942	1.036	c.management	0.976	1.036	c.management	0.942	1.036	c.people	0.754	1.006
c.fired	0.694	1.015	c.fired	0.760	1.015	c.fired	0.652	1.015	p.company	0.652	1.007
p.freedom	0.614	1.012	p.freedom	0.758	1.012	p.freedom	0.542	1.012	p.company-name	0.532	1.003
c.fear	0.524	1.013	c.job	0.672	1.007	c.job	0.484	1.007	c.company-name	0.436	1.003
c.job	0.514	1.007	c.fear	0.622	1.013	c.fear	0.482	1.013	p.people	0.434	1.004
p.pay	0.290	1.004	p.pay	0.434	1.004	p.pay	0.304	1.004	p.culture	0.382	1.032

Table 9 (continued)

`lasso-ebic1`	Incl	Highest	`lasso-k7`	Incl	Highest	`sis-k7`	Incl	Highest	`lda-k7`	Incl	Highest
$P^* = 7$	freq	MCR		freq	MCR		freq	MCR		freq	MCR
Mean	0.651	1.017		0.745	1.017		0.627	1.017		0.575	1.008

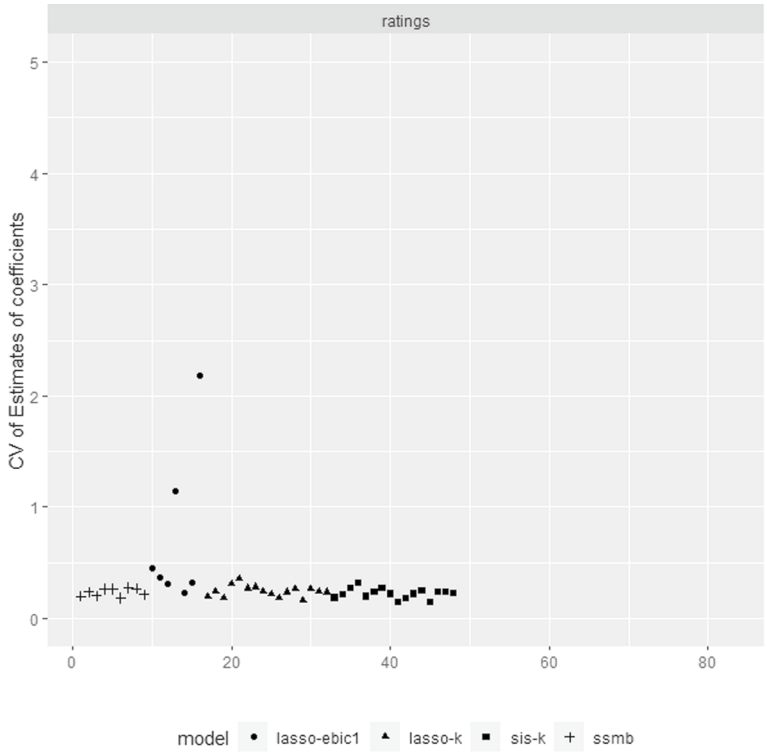

Fig. 5 Ratings. CV of estimates of coefficients

5.2.3 Summary

These results discussed above are corroborated by Fig. 5, confirming that, except for the highest coefficient of variation of `lasso-ebic10`, all the methods behave in a much more homogeneous way, with respect to the previous case study, in terms of the inclusion frequency and predictive R^2.

All in all, in this second case study, among text regression methods, stability selection is the preferred method, as it ensures an acceptable trade-off between explanatory power and parsimony. In addition, it outperforms the other methods in terms of persistence and relevant variables. However, differently than in the case of prices, with ratings, our findings are slightly in favour of topic modelling. The reasons can be related to the fact that in this example (a) sentences are not as short as in the case of prices and (b) the words have an emotional content, in this case, stronger than in the case of attributes, such as *freedom* as opposed to *wool*.

On this, and with the focus on interpretability, it is worth remarking that selected words have to be read jointly with most co-occurrent words. The task of understanding the meaning to which each word refers is usually performed in `lda` by looking at the top words with most probability to occur within each topic. To follow a similar

Table 10 Reconstructed topics (co-occurrence greater than 20%) from pivotal words

Pivotal word	Co-occurrent words
p.culture	p.deck; p.culture.deck; p.company.culture
c.management	c.upper.management; c.upper; c.exist
p.freedom	p.responsability
c.fear	c.culture; c.firings; c.performers; c.process; c.managing; c.simply c.fear.based; c.mistake
c.fired	c.people
p.company	p.direction
p.pay	p.benefits
c.job	c.security; c.job.security; c.comapny.name; c.pay; c.time; c.training
c.worst	c.coaching
p.free	p.free.company.name; p.free.food; p.coffee; p.free.movies

path in text regression, we consider the tokens that show the greatest co-occurrence, in a sort of topic reconstruction, displayed in Table 10. In this way, each selected word is accompanied by some further ones, among which it plays the pivotal role. As well as for standard topic models, such as `lda`, each group of topics has to be interpreted. The researchers focused on the field of study, jointly looking at the top hth token, find the best meaning for the topic.

In summary, in this case study, our results do not strongly support to use text regression methods but favour `lda`. The latter performs slightly better than text regression methods in explaining ratings data. We argue that this can be related to the length of the data text, in this second case study longer than in the first one, and on the very nature of the case study itself, where joint co-occurrence of words within the pros and cons fields—rather than single words alone—should explain the topic, as motivations underlying the ratings.

6 Concluding remarks

The paper has investigated the analytics that allow one to exploit the informative content and the explanatory power of unstructured, short texts on a response variable.

Interpretability of the results was a key issue for the scope of the present study; hence, we have restricted our focus to shrinkage methods, and within this class, we have favoured models that provide results of easier interpretation.

In this perspective, we have compared the explanatory power of variables selected through several variants of lasso, screening-based methods and randomisation-based models, such as sure independence screening and stability selection. A subsequent comparison has been also run with the widely applied topic model, i.e. LDA, used as a benchmark. The relative performance of the methods has been assessed based on the number and the importance of the selected variables.

We have considered two applications. The first application focused on explaining prices of goods within a product category, based on the captions provided by manufacturers on e-commerce platforms. In this case study, the nature of the texts is descriptive, as they characterise the goods for sale; after the text pre-processing phase, texts are very short, reduced in the median to only three words. The second application aimed to understand how to use open questions to obtain information on overall satisfaction within surveys. After the text pre-processing phase, texts are short, with 15 words in the median, but longer than in the previous case. Furthermore, here texts express opinions, and in particular satisfaction or dissatisfaction; thus, compared to the first case study, they are much more related to the emotional sphere.

The results of the study attain insights into two main directions concerning, on the one side, the performance of analysed models within the class of text regression and, on the other side, the different ability of text regression versus topic modelling methods in extracting information from short texts.

Along the first direction, our findings show that, in terms of explanatory power, both stability selection and lasso which optimises the eBIC criterion are able to improve the lasso, when the latter is optimised through standard criteria finalised to prediction purposes. Nevertheless, by limiting the number of selected variables, both lasso and sure independence screening are capable of attaining comparable results. As far as the ability to select relevant explanatory predictors is concerned, stability selection slightly outperformed the other methods, which, notwithstanding, exhibited good performance. In our opinion, a relevant finding is that lasso behaves as well as alternative computationally more intensive methods when the number of selected variables is limited.

Concerning the comparison of text regression with LDA, the former outperforms LDA in terms of explanatory power in the prices case study while LDA outperforms text regressions in the ratings case study. This is likely to happen both because texts are longer in the ratings case study and because of their contents, which are naturally more connected to latent topics.

In terms of quality of the selected words, text regression overcomes LDA in the case of prices but not entirely in the case of ratings. However, words selected based on text regressions are always more robust than LDA ones, so that, in both cases, text regressions appear highly suitable to pick up relevant words within a bag of words.

To conclude, we remark that the results of the paper describe how text regression and variable selection methods work over two specific applications and cannot be extensively generalised if not with further extensive analyses. However, our findings favour variable selection in text regressions as a method that may provide valuable solutions when texts are short and open the way to further investigations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10182-023-00472-0.

Acknowledgements We would like to thank the Editor, Prof. Yarema Okhrin, the Associate Editor and two referees for the careful reading of the manuscript and for the insightful comments that have led to an improved version of the original paper. The scientific output expressed in the paper does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the commission is responsible for the use which might be made of this publication.

References

Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)

Anderlucci, L., Viroli, C.: Mixtures of Dirichlet-multinomial distributions for supervised and unsupervised classification of short text data. Adv. Data Anal. Classif. **14**, 759–770 (2020)

Bach, F.R.: Consistency of the group lasso and multiple kernel learning. J. Mach. Learn. Res. **9**, 1179–1225 (2008)

Belloni, A., Chernozhukov, V.: Least squares after model selection in high-dimensional sparse models. Bernoulli **19**(2), 521–547 (2013)

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

Bogdan, M., Ghosh, J.K., Doerge, R.W.: Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics **167**(2), 989–999 (2004)

Breiman, L.: Better subset regression using the nonnegative garrote. Technometrics **37**(4), 373–384 (1995)

Broman, K.W., Speed, T.P.: A model selection approach for the identification of quantitative trait loci in experimental crosses. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **64**(4), 641–656 (2002)

Cachon, G.P., Swinney, R.: The value of fast fashion: quick response, enhanced design, and strategic consumer behavior. Manag. Sci. **57**(4), 778–795 (2011)

Candes, E., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**(6), 2313–2351 (2007)

Chen, J., Chen, Z.: Extended Bayesian information criteria for model selection with large model spaces. Biometrika **95**, 759–771 (2008)

Cheng, X., Yan, X., Lan, Y., Guo, J.: Btm: topic modeling over short texts. IEEE Trans. Knowl. Data Eng. **26**(12), 2928–2941 (2014)

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)

Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning, arxiv (2017)

Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis (1998)

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)

Fan, J., Lv, J.: Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **70**(5), 849–911 (2008)

Ferreira-Mello, R., Andre, M., Pinheiro, A., Costa, E., Romero, C.: Text mining in education. WIREs Data Min. Knowl. Discov. **9**(6), e1332 (2019)

Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. arXiv (2018)

Foster, D., Liberman, M., Stine, R.: Featurizing text: converting text into predictors for regression analysis (2013)
Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)
Friedman, J.H., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 033(i01) (2010)
Gentzkow, M., Kelly, B., Taddy, M.: Text as data. J. Econ. Lit. **57**(3), 535–74 (2019)
Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer, Berlin (2009)
Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
Hofner, B., Boccuto, L., Göker, M.: Controlling false discoveries in high-dimensional situations: boosting with stability selection. BMC Bioinform. **16**, 144 (2015)
Hofner, B., Hothorn, T.: Stabs: Stability Selection with Error Control. R package version 0.6-3 (2017)
Hollibaugh, G.E.: The use of text as data methods in public administration: a review and an application to agency priorities. J. Public Admin. Res. Theory **29**(3), 474–490 (2019)
Jentsch, C., Lee, E.R., Mammen, E.: Time-dependent Poisson reduced rank models for political text data analysis. Comput. Stat. Data Anal. **142**, 106813 (2020)
Jipeng, Q., Zhenyu, Q., Yun, L., Yunhao, Y., Xindong, W.: Short text topic modeling techniques, applications, and performance: a survey (2019)
Jones, T.: textmineR: functions for text mining and topic modeling. R Package Vers **3**, 4 (2019)
Lange, K.-R., Rieger, J., Jentsch, C.: Lex2sent: a bagging approach to unsupervised sentiment analysis, arxiv (2022)
Luque, C., Luna, J.M., Luque, M., Ventura, S.: An advanced review on text mining in medicine. WIREs Data Min. Knowl. Discov. **9**(3), e1302 (2019)
Margot, V., Luta, G.: A new method to compare the interpretability of rule-based algorithms. AI **2**(4), 621–635 (2021)
Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. Ann. Stat. **34**(3), 1436–1462 (2006)
Meinshausen, N., Bühlmann, P.: Stability selection. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **72**(4), 417–473 (2010)
Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning based text classification: a comprehensive review (2020)
Nowak, A., Smith, P.: Textual analysis in real estate. J. Appl. Econom. **32**(4), 896–918 (2017)
Reisenbichler, M., Reutterer, T.: Topic modeling in marketing: recent advances and research opportunities. J. Bus. Econ. **89**(3), 327–356 (2018)
Saldana, D.F., Feng, Y.: SIS: an R package for sure independence screening in ultrahigh-dimensional statistical models. J. Stat. Softw. **83**(2), 1–25 (2018)
Schwarz, C.: Ldagibbs: a command for topic modeling in stata using latent Dirichlet allocation. Stand. Genomic Sci. **18**(1), 101–117 (2018)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
Shah, R.D., Samworth, R.J.: Variable selection with error control: another look at stability selection. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **75**(1), 55–80 (2013)
Soysal, G.P., Krishnamurthi, L.: Demand dynamics in the seasonal goods industry: an empirical analysis. Mark. Sci. **31**(2), 293–316 (2012)
Stone, M.: Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc.: Ser. B (Methodol.) **36**(2), 111–133 (1974)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) **58**(1), 267–288 (1996)
Tuan, A.P., Tran, B., Nguyen, T.H., Van, L.N., Than, K.: Bag of biterms modeling for short texts. Knowl. Inf. Syst. **62**(10), 4055–4090 (2020)
Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc.: Series B (Stat. Methodol.) **67**(2), 301–320 (2005)
Zou, H., Hastie, T., Tibshirani, R.: On the degrees of freedom of the lasso. Ann. Stat. **35**(5), 2173–2192 (2007)