Statistics and Its Interface Volume 11 (2018) 41–50

Feature screening for ultrahigh dimensional binary data

Guoyu Guan, Na Shan, and Jianhua Guo*

With the rapid development of information technology, ultrahigh dimensional binary data have increased dramatically, for which feature screening has become a necessary step in real data analysis. In this article, we propose a L_0-regularization feature screening procedure for naive Bayes classifier, which is equivalent to the classical mutual information screening method. However, the turning parameter in L_0-regularization is hard to be selected and lack of theoretical support. To this end, a BIC-type criterion is applied to identify important features. Moreover, the asymptotic properties of the proposed method is theoretically investigated under some mild assumptions. Lastly, its outstanding performance is numerically confirmed on simulated data, and a real example of Chinese document classification is presented for illustration purpose.

AMS 2000 subject classifications: Primary 62F07; secondary 62H30.
Keywords and phrases: Feature screening, L_0-regularization, Naive Bayes, Screening consistency.

1. INTRODUCTION

With the rapid development of information technology, ultrahigh dimensional data have attracted considerable attentions in the literature. For example, in text documents, the total number of keywords (i.e., the feature dimension p) is ultrahigh, which grows much faster than the sample size n. As noted by [4] and [5], a large number of features (or variables) may be irrelevant for some specific research purpose. To this end, feature (or variable) selection becomes an important but challenging task, which appeared early in the area of machine learning. The most popular used feature selection criteria include but not limited to information gain, mutual information, chi-square test and many others [20, 13]. However, these methods are lack of theoretical support, i.e., how many features should be selected is not clear in real applications.

For the last decade, feature selection and feature screening for ultrahigh dimensional data have been discussed extensively in the statistical literature. Fan and Lv [3] proposed an independent screening framework by ranking the marginal correlations, which is a seminal work on ultrahigh dimensional feature screening. [17] proposed a forward regression method for feature screening in ultrahigh dimensional linear models. [22] proposed a feature screening procedure under a model-free framework, which covers a variety of commonly used parametric and semiparametric models. The aforementioned methods are all developed for dealing with ultrahigh dimensional continuous predictors. [8] proposed a Pearson chi-square based sure independence feature screening procedure for categorical response with ultrahigh dimensional categorical predictors. In this article, we propose a feature screening procedure especially for ultrahigh dimensional binary data, which is motivated by an empirical study of Chinese document classification.

In real study, we consider the classification of Chinese text documents. The documents were generated by the Mayor Public Hotline (MPH) of Changchun, the capital city of Jilin Province in Northeast China. The goal is to automatically classify the MPH telephone records (i.e., Chinese text documents) according to their corresponding functional departments in the local government. To be specific, a bag of frequently used Chinese words are collected. Note that, the documents are very short which causes the vast majority of words appear at most once per document. Hence each document is converted into a binary vector, which records whether these words appear in the document or not. Because the total number of words (i.e., p) collected is huge, the feature dimension p is ultrahigh. To this end, many existing classification methods can be applied, such as support vector machine, classification and regression trees, k-nearest neighbor, random forest, naive Bayes and many others [1, 19, 7, 21]. Nevertheless, experienced experts tell us that only a small number of words are relevant for classification and a large amount of them are irrelevant, which results that the classification accuracy is not optimal. Hence feature selection on this dataset is necessary. Among all aforementioned classification methods, we find that naive Bayes (NB) [9, 6, 11] is particularly attractive and always performs well in document classification, due to its simplicity and effectiveness. Furthermore, NB is fairly sensitive to feature selection, having its performance peak at a much lower number of features selected [4]. Meanwhile, among all existing feature (or variable) selection methods, L_0-regularization is shown to have the oracle property under exponentially large $p = e^{o(n)}$ [15, 12], which implies that it can handle ultrahigh dimensional data. But it is computationally infeasible to optimize such a discontinuous objective function (i.e., L_0-regularized

*Corresponding author.

likelihood). Motivated by these facts, we construct a feasible L_0-regularization feature selection framework based on a NB model, which is equivalent to the feature screening method by ranking mutual information. Then the resulting model can be only selected from finite nested candidate models, which avoids solving the discontinuous objective function. However, the turning parameter in L_0-regularization is hard to be selected in practice and lack of theoretical support. To this end, a BIC-type criterion is applied to identify important features, which is a natural extend of the classical Bayesian information criterion [14]. Moreover, the screening consistency of the BIC-type criterion is investigated under some mild assumptions. Although this method is especially proposed for classification of ultrahigh dimensional binary data, the idea can be naturally applied to other categorical or continuous cases.

The rest of this article is organized as follows. In Section 2, we introduce the double truncated estimators and the L_0-regularization method for feature selection. Then, a BIC-type criterion is applied to select important features. In Section 3, the outstanding performance of the proposed method is numerically confirmed on both simulated and empirical datasets. Lastly, some concluding remarks are given in Section 4. All the theoretical proofs are left to the Appendices.

2. METHODOLOGY

2.1 Model and notations

Let $(\mathbf{X}, \mathbf{Y}) = (X_i, Y_i)_{1\le i\le n}$ be a vector of n independent and identically distributed observations collected from the real world. For the i-th subject, $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \{0,1\}^p$ is the associated p-dimensional binary feature, and $Y_i \in \{1, \cdots, K\}$ is the corresponding class label. Note that the feature dimension p is assumed to be ultrahigh and may greatly exceed the sample size n. We simply assume that the features are conditionally independent given the class label. Thus, a so-called naive Bayes (NB) model assumes that

$$P(X_i = x|Y_i = k) = \prod_{j=1}^{p} \theta_{kj}^{x_j}(1-\theta_{kj})^{1-x_j}, \tag{1}$$

where $\theta_{kj} = P(X_{ij} = 1|Y_i = k)$, and $x = (x_1, \cdots, x_p)^\top \in \{0,1\}^p$ stands for one particular realization of X_i. Moreover, we define $P(Y_i = k) = \pi_k$.

To estimate the parameters, the maximum likelihood estimators (MLE) can be adopted, i.e., $\hat{\pi}_k^{ML} = n^{-1}\sum_i Z_{ik}$ and $\hat{\theta}_{kj}^{ML} = \{\sum_i Z_{ik}\}^{-1}\sum_i X_{ij}Z_{ik}$, where $Z_{ik} = \mathbb{1}(Y_i = k)$. Note that $\mathbb{1}(\cdot)$ represents the indicator function. The trouble of the MLE is that it may overfit the data, which means it can reach 1 or 0 with some positive probability. In practical applications, Laplacian smoothing [10] is commonly used to enforce the estimates bounded away from 0 and 1. But we only care about the classification of short Chinese documents in which the majority of word frequencies are fairly small. Thus, the Laplacian smoothing may lead to large bias of estimators, which are not our main concern in this article. To this end, we define the double truncated estimators (DTE) as follows. Without loss of generality, assume $\sum_i Z_{iK} = \max_k\{\sum_i Z_{ik}\}$. For $1 \le k \le K-1$, define $\hat{\pi}_k = \max\{n^{-1}, \min\{1-n^{-1}, n^{-1}\sum_i Z_{ik}\}\}$. Then it becomes apparent that $\hat{\pi}_K = 1 - \sum_{k=1}^{K-1}\hat{\pi}_k$ lies between n^{-1} and $1-n^{-1}$. Next, define $\hat{\theta}_{kj} = \hat{\pi}_k^{-1}\max\{n^{-1}, \min\{1-n^{-1}, n^{-1}\sum_i Z_{ik}X_{ij}\}\}$, for every k and j. Note that, this trick can enforce the estimators bounded away from 0 and 1.

In classification problems, one main goal is to compute the posterior probability, which can be estimated as

$$\widehat{P}(Y_i = k|X_i = x) = \left\{\hat{\pi}_k \prod_{j=1}^{p}\hat{\theta}_{kj}^{x_j}(1-\hat{\theta}_{kj})^{1-x_j}\right\} \times\left\{\sum_{k'=1}^{K}\hat{\pi}_{k'}\prod_{j=1}^{p}\hat{\theta}_{k'j}^{x_j}(1-\hat{\theta}_{k'j})^{1-x_j}\right\}^{-1}.$$

Subsequently, the unknown class label of a new observation with $X_{new} = x$ can be predicted as $\hat{Y}_{new} = \text{argmax}_k\widehat{P}(Y_{new} = k|X_{new} = x)$.

2.2 Feature selection by L_0-regularization

In the above prediction rule, those features irrelevant to (or independent of) the class label should be filtered out. Thus, the task of feature selection becomes indispensable. Intuitively, if the j-th feature is irrelevant to the class label, its response probabilities across all classes should be the same. As a consequence, we should have $\theta_{kj} = \theta_j$ for every k, where $\theta_j = P(X_{ij} = 1) = \sum_k \pi_k\theta_{kj}$ (the corresponding estimator is $\hat{\theta}_j = \sum_k \hat{\pi}_k\hat{\theta}_{kj}$). In contrast, if the j-th feature is relevant to the class label, the response probabilities should be different for at least two classes. Then we define the true model as $\mathcal{M}_T = \{1 \le j \le p : \sum_k |\theta_{kj} - \theta_j| > 0\}$ with size $|\mathcal{M}_T| = d_0$. Define $\mathcal{M}_F = \{1, \cdots, p\}$ to be the full model. We further define a generic notation $\mathcal{M} = \{j_1, \cdots, j_d\}$ to be an arbitrary model with $X_{ij_1}, \cdots, X_{ij_d}$ as relevant features.

As we know, if the maximum likelihood estimation is applied, the result of feature selection will be overfit, i.e., $\widehat{\mathcal{M}} = \mathcal{M}_F$, which is not our purpose. A natural way, L_0-penalty should be used to the model size $|\mathcal{M}|$. To derive the L_0-penalized log-likelihood, we introduce the feature indicator $\delta = (\delta_1, \cdots, \delta_p)^\top \in \{0,1\}^p$, where $\delta_j = \mathbb{1}(j \in \mathcal{M})$. More specifically, $\delta_j = 1$ means that the j-th feature is relevant, and $\delta_j = 0$ otherwise. Next, we can treat δ as unknown parameters, and select relevant features by maximizing the following L_0-penalized log-likelihood,

$$\begin{aligned}\mathcal{L}_p &= \mathcal{L}_t(\mathcal{M}) - n\lambda|\mathcal{M}| \\ &= n\left[\sum_{k=1}^{K}\hat{\pi}_k\log\hat{\pi}_k + \sum_{j=1}^{p}\{\hat{\theta}_j\log\hat{\theta}_j + (1-\hat{\theta}_j)\log(1-\hat{\theta}_j)\}\right. \\ &\quad \left. + \sum_{j=1}^{p}\delta_j(\hat{I}_j - \lambda)\right],\end{aligned}$$

where $\mathcal{L}_t(\mathcal{M})$ is the truncated log-likelihood, $\lambda \geq 0$ is a regularization constant and

$$\hat{I}_j = \sum_{k=1}^{K} \hat{\pi}_k \left[\hat{\theta}_{kj} \log \frac{\hat{\theta}_{kj}}{\hat{\theta}_j} + (1 - \hat{\theta}_{kj}) \log \frac{1 - \hat{\theta}_{kj}}{1 - \hat{\theta}_j} \right].$$

Detailed derivations of $\mathcal{L}_t(\mathcal{M})$ and $\mathcal{L}_p$ are left to Appendix A. Then δ can be estimated by $\hat{\delta} = \text{argmax}_{\delta \in \{0,1\}^p} \sum_j \delta_j (\hat{I}_j - \lambda)$. Thus, we can easily get $\hat{\delta}_j = \mathbb{1}(\hat{I}_j > \lambda)$. This suggests that the true model $\mathcal{M}_T$ can be estimated by $\widehat{\mathcal{M}}_\lambda = \{1 \leq j \leq p : \hat{I}_j > \lambda\}$, for some appropriately regularization constant λ. Here $\hat{I}_j$ is an estimator of mutual information

$$\begin{aligned} I_j &= \sum_{k=1}^{K} \sum_{x_j=0}^{1} P(X_{ij} = x_j, Y_i = k) \log \frac{P(X_{ij} = x_j, Y_i = k)}{P(X_{ij} = x_j) P(Y_i = k)} \\ &= \sum_{k=1}^{K} \pi_k \left[\theta_{kj} \log \frac{\theta_{kj}}{\theta_j} + (1 - \theta_{kj}) \log \frac{1 - \theta_{kj}}{1 - \theta_j} \right]. \end{aligned}$$

As is known to all, the mutual information $I_j \geq 0$ and $\hat{I}_j \geq 0$ by Jensen's inequality. $I_j = 0$ if and only if feature X_{ij} and class label Y_i are mutually independent. Accordingly, we intend to select those features far away from independency. Intuitively, those features with larger $\hat{I}_j$ values are more likely to be relevant for classification. In contrast, those with smaller $\hat{I}_j$ values are less likely.

To gain some theoretical insight, some basic technical assumptions should be clear first. **Basic Setting:** The class number K is fixed, and the feature dimension p is ultrahigh but subjects to $\log p = o(n)$. **Boundedness Assumption:** There exists some positive constant $\nu < 1/3$, such that $\pi_k \geq \nu$ and $\nu \leq \theta_{kj} \leq 1 - \nu$ for all k and j, and $\max_k \{|\theta_{kj} - \theta_j|\} \geq \nu$ for any $j \in \mathcal{M}_T$. Under this assumption, all parameters are bounded away from 0 and 1, and all relevant features are apart from the irrelevant ones. By the next theorem, we know that $\widehat{\mathcal{M}}_\lambda$ is consistent to $\mathcal{M}_T$, as long as the regularization constant λ is appropriately selected.

Theorem 1. *Under the basic setting and boundedness assumption, we have* $\max_j |\hat{I}_j - I_j| = O_P(\sqrt{\log p / n})$. *Additionally, there exists some* $\lambda_{max} > 0$, *such that for any* $0 < \lambda < \lambda_{max}$, $P(\widehat{\mathcal{M}}_\lambda = \mathcal{M}_T) \to 1$, *as* $n \to \infty$.

2.3 Feature screening by a BIC-type criterion

Theorem 1 shows that, $\widehat{\mathcal{M}}_\lambda$ is a consistent estimator of $\mathcal{M}_T$, provided by the regularization constant λ is appropriately selected. Although there exist infinitely many different choices for λ on $[0, \infty)$, the resulting model $\widehat{\mathcal{M}}_\lambda$ can be only selected from finite nested candidate models. Without loss of generality, assume that the features' indices have been appropriately re-labeled such that $\hat{I}_1 > \hat{I}_2 > \cdots > \hat{I}_p$. Thus, the solution path can be given by $\mathbb{M} = \{\mathcal{M}_{(d)} : 0 \leq d \leq p\}$ with $\mathcal{M}_{(0)} = \emptyset$ and $\mathcal{M}_{(d)} = \{1, \cdots, d\}$ for $1 \leq d \leq p$, which is a finite set with a total of $p + 1$ nested candidate models. The conclusion of Theorem 1 also implies that $P(\mathcal{M}_T \in \mathbb{M}) \to 1$, as $n \to \infty$. Subsequently, the original problem of determination for regularization constant λ is converted into a model selection problem with respect to the solution path $\mathbb{M}$. However, the true model size is unknown and λ is hard to be selected practically. Then we follow [18] and consider the following BIC-type criterion

$$\text{BIC}_{\mathcal{M}} = -\frac{2}{n} \mathcal{L}_t(\mathcal{M}) + df(\mathcal{M}) \times \frac{\log(n)}{n}, \tag{2}$$

where $df(\mathcal{M}) = (K-1) + K|\mathcal{M}| + (p - |\mathcal{M}|)$ is the number of free parameters estimated in the truncated log-likelihood $\mathcal{L}_t(\mathcal{M})$. Thus, the best model is selected as $\widehat{\mathcal{M}} = \mathcal{M}_{(\hat{d})}$, where $\hat{d} = \text{argmin}_{1 \leq d \leq p} \text{BIC}_{\mathcal{M}_{(d)}}$. It can filter out the features that have weak dependency with the class label. Then the following theorem states the screening consistency of this BIC-type criterion.

Theorem 2. *Under the basic setting and boundedness assumption, we have* $P(\mathcal{M}_T \subset \widehat{\mathcal{M}}) \to 1$ *as* $n \to \infty$.

In conclusion, the proposed feature screening method could be called as L_0-regularized naive Bayes (L0NB) method. After the feature screening step, the prediction rule of NB model can be simplified as $\hat{Y}_{new} = \text{argmax}_k \widehat{P}(Y_{new} = k | X_{new}(\widehat{\mathcal{M}}) = x(\widehat{\mathcal{M}}))$, where $X_{new}(\widehat{\mathcal{M}})$ and $x(\widehat{\mathcal{M}})$ are the subvectors of X_{new} and x corresponding to the estimated model $\widehat{\mathcal{M}}$. The performance of numerical studies in the next section suggests that L0NB works quite well both on simulated and real data.

3. NUMERICAL STUDIES

3.1 Competitive methods

To evaluate the performance of the proposed feature screening method, we consider two alternative methods as competitors, i.e., (1) the Pearson chi-square based sure independence feature screening (PC-SIS) proposed by [8], and (2) the L_1-penalized naive Bayes (L1NB) method, which is motivated by the lasso [16] method of linear regression model. Note that, the penalized method relative to naive Bayes has not been well defined before. The L_1-penalized log-likelihood function for NB can be represented as

$$\begin{aligned} \mathcal{L}_1 = &\sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{K} Z_{ik} \{ X_{ij} \log \theta_{kj} + (1 - X_{ij}) \log(1 - \theta_{kj}) \} \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k - \lambda_1 \sum_{j=1}^{p} \sum_{k=1}^{K} \left| \theta_{kj} - n^{-1} \sum_{i=1}^{n} X_{ij} \right|. \end{aligned} \tag{3}$$

Thus, the L_1-penalty can make some θ_{kj} shrinking to the common mean of j-th feature. By maximizing (3), we get the corresponding parameter estimators. After some mathematical derivations, the true model can be estimated by $\widehat{\mathcal{M}}_{\lambda_1} =$

$\{1 \le j \le p : H_j > \lambda_1\}$, after giving a regularization parameter λ_1, where $H_j = \max_{1\le k\le K}\{n\hat{\pi}_k\hat{\theta}_j^{-1}(1-\hat{\theta}_j)^{-1}|\hat{\theta}_{kj}-\hat{\theta}_j|\}$. Note that, $\hat{\pi}_k$, $\hat{\theta}_{kj}$ and $\hat{\theta}_j = \sum_k \hat{\pi}_k\hat{\theta}_{kj}$ are the DTEs which have been defined in section 2.1. Subsequently, the regularization parameter λ_1 can be selected by BIC-type criterion (2) or cross-validation. Detailed derivations of L1NB are left to Appendix B.

To evaluate the classification accuracy of the proposed method, we also consider five alternative classification algorithms as competitors. They are k-nearest neighbor (KNN), adaptive boosting (AdaB), random forest (RF), support vector machine (SVM) and NB. The detailed settings of these algorithms should be declared first. (1) For KNN method, the Hamming distance is used as the distance metric, because it behaves quite well on binary features. The smoothing parameter k is selected by handout cross-validation from 1 to 10. (2) For AdaB method, tree classifiers are used as base learners, because they could be usually viewed as weak learners and especially suitable for categorical predictors. More specifically, there are a total of 200 trees used for simulated data and 500 trees used for real high dimensional data. (3) There are a total of 200 tree classifiers used in RF, and $n_f = \min\{\lfloor\sqrt{p}\rfloor, 30\}$ features are randomly selected for each tree, where $\lfloor p\rfloor$ is the maximum integer not grater than p. These trees are built using a greedy, top-down recursive partitioning strategy, and the minimum number of observations in per tree leaf is restricted to 1. Hence the depth of each tree is not grater than n_f attributing to binary features. (4) The linear kernel is used in SVM. Because a binary feature will still be a binary feature after any nonlinear transformation. (5) NB is the traditional naive Bayes classifier as described in section 2.1.

3.2 Simulation studies

To evaluate the finite sample performance of the newly proposed method, our simulation studies consider naive Bayes models with different sample sizes (i.e., n =500, 1,000, 2,000), feature dimensions (i.e., p =500, 1,000), and true model sizes (i.e., d_0 =20, 30, 50). For each fixed parameter setting, a total of 500 simulation replications are conducted. For each simulated dataset, the three feature screening methods are adopted, i.e., PC-SIS, L1NB (with BIC-type criterion for model selection) and L0NB. Subsequently, the percentage of incorrect zeros [2], that is PIZ $= 100\% \times \{|(\mathcal{M}_F\backslash\widehat{\mathcal{M}})\cap\mathcal{M}_T|\}|\mathcal{M}_T|^{-1}$ of three methods, are computed and averaged. The same is also done to the percentage of correct zeros, that is PCZ $= 100\% \times \{|(\mathcal{M}_F\backslash\widehat{\mathcal{M}})\cap(\mathcal{M}_F\backslash\mathcal{M}_T)|\}\{|(\mathcal{M}_F\backslash\mathcal{M}_T)|\}^{-1}$. Average PCZ and PIZ values of three feature screening methods, and the corresponding standard errors (SE) over 500 replications are reported in percentage. Lastly, in order to evaluate the performance of classification, another 1,000 independent testing samples are generated. Then, the classification accuracies of NB on three estimated models (separately selected by PC-SIS, L1NB and L0NB) are evaluated on the testing sample. The corresponding average mis-classification rates (AMR) and their SE values over 500 replications are computed and reported in percentage. For comparison's sake, the AMR and SE values of KNN, AdaB, RF, SVM and NB but based on the true model $\mathcal{M}_T$ are also included.

We consider here a standard NB model with irrelevant features. As a result, the standard NB method on the true model is expected to perform best. In contrast, due to the unnecessary noise introduced by the irrelevant features, the $\widehat{\mathcal{M}}$-based NB method should perform worse. We then use this example to numerically investigate the feature screening ability and the classification accuracy of the proposed method. More specifically, we generate the class label $Y_i \in \{1, \cdots, K\}$ with probability $P(Y_i = k) = 1/K$ and $K = 3$. Next, given Y_i, the j-th binary feature X_{ij} is generated from a Bernoulli distribution with probability $P(X_{ij} = 1|Y_i = k) = \theta_{kj}$ for $j \in \mathcal{M}_T$, and $P(X_{ij} = 1|Y_i = k) = \theta_j$ for $j \notin \mathcal{M}_T$, where $\mathcal{M}_T = \{1, \cdots, d_0\}$ is the true model with size d_0. In addition to that, $\{\theta_{kj}\}_{1\le k\le m, j\in\mathcal{M}_T}$ and $\{\theta_j\}_{j\notin\mathcal{M}_T}$ are simulated from a uniform distribution on $[0.1, 0.9]$.

The detailed simulation results are given in Tables 1 and 2. From Table 1, we find that L0NB performs the best, L1NB performs a little worse than L0NB, and PC-SIS performs the worst, in term of PCZ. Then PC-SIS performs better than L1NB and L0NB in term of PIZ, but PIZ of PC-SIS cannot tend to 0% as n gets larger. Both PCZ and PIZ values of PC-SIS are smaller than other methods, which shows that PC-SIS is a relatively conservative feature screening method. Moreover, the SE values of PCZ and PIZ for PC-SIS are much larger than that of L1NB and L0NB, which implies that PC-SIS is less robust than L1NB and L0NB. We also find that the PCZ values of L1NB and L0NB approach 100% and the PIZ values of L1NB and L0NB approach 0% quickly as n gets larger. This confirms that the proposed BIC-type criterion is indeed consistent for feature selection. From Table 2, we find that the $\mathcal{M}_T$-based NB method performs the best in term of AMR. This is as expected because the true model is indeed NB. In the meanwhile, the performance of NB based on $\widehat{\mathcal{M}}$ selected by L0NB is extremely comparable, and NB based on $\widehat{\mathcal{M}}$ selected by L1NB also works quite well. This suggests that even if the true model is NB, the efficiency loss suffered by irrelevant features is very limited. Furthermore, the SE values of NB based on $\widehat{\mathcal{M}}$ selected by PC-SIS are lager than that of L1NB and L0NB, which also implies that PC-SIS is less robust than L1NB and L0NB. We also find that from both $\mathcal{M}_T$-based and $\widehat{\mathcal{M}}$-based results, larger sample size n leads to smaller AMR and SE values, if both p and d_0 are fixed. This is expected because larger sample size leads to more accurate estimation. In the meanwhile, for a fixed p and n, we find that larger true model size d_0 leads to smaller AMR, because the more relevant features involved the better we can predict. Lastly, with a fixed d_0 and n, we find that larger feature dimension p leads to worse performance in terms of AMR. This is also reasonable, because larger feature dimension leads to more challenge for feature selection and then worse prediction.

Table 1. Simulation results for evaluating the performance of three feature screening methods (i.e., PC-SIS, L1NB, L0NB). The average PCZ and PIZ values and their standard errors (SE) are reported in percentage over 500 replications

			PCZ×10²(SE×10²)			PIZ×10²(SE×10²)		
p	d_0	n	PC-SIS	L1NB	L0NB	PC-SIS	L1NB	L0NB
500	20	500	27.3(44.3)	99.7(0.3)	99.8(0.2)	3.5(6.2)	9.5(3.1)	8.3(2.8)
		1000	47.6(49.8)	99.9(0.1)	99.9(0.2)	4.4(5.1)	6.7(2.5)	5.8(2.4)
		2000	62.1(48.4)	100.0(0.1)	100.0(0.1)	4.5(4.0)	4.9(1.2)	4.4(1.7)
	30	500	10.0(29.4)	99.6(0.4)	99.8(0.2)	2.0(7.1)	12.1(3.8)	9.8(3.3)
		1000	25.3(43.2)	99.9(0.2)	99.9(0.2)	2.5(4.9)	7.1(2.6)	5.9(2.5)
		2000	39.2(48.7)	100.0(0.1)	100.0(0.1)	2.6(3.9)	4.0(1.6)	3.3(1.7)
	50	500	4.9(20.4)	99.4(0.5)	99.8(0.2)	0.8(4.1)	15.0(3.3)	12.4(2.4)
		1000	17.0(37.1)	99.7(0.3)	99.9(0.1)	1.9(4.7)	9.3(2.3)	8.4(1.9)
		2000	25.1(43.0)	99.9(0.1)	100.0(0.1)	2.5(4.5)	6.0(1.9)	5.2(1.9)
1000	20	500	24.7(43.0)	99.7(0.2)	99.8(0.2)	3.6(7.6)	9.9(3.3)	8.1(2.8)
		1000	47.5(49.9)	99.9(0.1)	99.9(0.1)	4.6(5.3)	7.1(2.5)	5.7(2.4)
		2000	57.5(49.4)	100.0(0.1)	100.0(0.1)	4.3(4.2)	5.1(1.0)	4.5(1.7)
	30	500	9.6(29.1)	99.7(0.3)	99.8(0.1)	2.4(9.8)	13.3(3.7)	10.0(3.1)
		1000	22.5(41.7)	99.9(0.1)	99.9(0.1)	2.5(5.7)	7.6(2.7)	5.8(2.4)
		2000	36.9(48.2)	100.0(0.1)	100.0(0.1)	2.6(3.9)	4.4(1.7)	3.5(1.6)
	50	500	3.8(18.6)	99.5(0.3)	99.8(0.2)	1.0(6.3)	16.5(3.5)	12.4(2.5)
		1000	12.0(32.2)	99.8(0.2)	99.9(0.1)	1.4(4.0)	10.1(2.4)	8.5(2.0)
		2000	22.4(41.5)	99.9(0.1)	100.0(0.1)	2.2(4.3)	6.4(1.9)	5.2(1.8)

Table 2. Simulation results for evaluating the classification accuracy. The AMR values of NB based on $\widehat{\mathcal{M}}$ (separately estimated by PC-SIS, L1NB, L0NB) and their SE values are reported in percentage over 500 replications. For comparison's sake, the AMR and SE values of KNN, AdaB, RF, SVM and NB but based on the true model $\mathcal{M}_T$ are also are reported in percentage over 500 replications

			$\mathcal{M}_T$-based results for five competitors AMR×10²(SE×10²)					$\widehat{\mathcal{M}}$-based results for NB AMR×10²(SE×10²)		
p	d_0	n	KNN	AdaB	RF	SVM	NB	PC-SIS	L1NB	L0NB
500	20	500	11.3(2.2)	10.5(1.4)	8.9(1.0)	10.2(1.3)	6.7(0.8)	10.7(2.6)	6.9(0.8)	6.9(0.8)
		1000	10.7(2.0)	9.9(1.2)	8.4(0.9)	9.7(1.2)	6.6(0.8)	8.0(1.6)	6.6(0.8)	6.6(0.7)
		2000	10.2(1.9)	9.5(1.1)	8.0(0.9)	9.5(1.2)	6.4(0.8)	7.0(1.0)	6.5(0.8)	6.5(0.8)
	30	500	8.0(2.0)	7.7(1.2)	5.8(0.8)	6.7(0.9)	3.7(0.6)	6.6(1.1)	3.9(0.6)	3.8(0.6)
		1000	7.3(1.8)	7.1(1.0)	5.4(0.8)	6.2(0.9)	3.6(0.6)	4.8(0.9)	3.7(0.6)	3.7(0.6)
		2000	6.8(1.6)	6.8(0.9)	5.2(0.7)	6.0(0.8)	3.6(0.6)	4.1(0.7)	3.6(0.6)	3.6(0.6)
	50	500	2.7(1.1)	3.2(0.7)	1.9(0.4)	2.4(0.6)	0.8(0.3)	1.5(0.4)	0.9(0.3)	0.8(0.3)
		1000	2.3(1.0)	2.7(0.6)	1.7(0.4)	2.0(0.5)	0.8(0.3)	1.0(0.3)	0.8(0.3)	0.8(0.3)
		2000	2.1(0.9)	2.5(0.5)	1.7(0.4)	1.8(0.5)	0.7(0.3)	0.8(0.3)	0.7(0.3)	0.7(0.3)
1000	20	500	11.4(2.2)	10.5(1.4)	8.9(1.0)	10.2(1.3)	6.8(0.8)	14.4(4.4)	7.0(0.8)	7.0(0.8)
		1000	10.6(1.9)	9.8(1.1)	8.4(0.9)	9.8(1.2)	6.5(0.8)	9.4(2.8)	6.6(0.8)	6.6(0.8)
		2000	10.3(1.9)	9.7(1.0)	8.2(0.9)	9.7(1.2)	6.5(0.8)	7.7(1.6)	6.5(0.8)	6.5(0.8)
	30	500	7.9(2.0)	7.7(1.2)	5.8(0.8)	6.6(0.9)	3.7(0.6)	9.8(4.0)	4.0(0.7)	3.9(0.7)
		1000	7.2(1.8)	7.1(1.0)	5.3(0.8)	6.2(0.8)	3.6(0.6)	6.1(1.5)	3.7(0.6)	3.7(0.6)
		2000	6.9(1.6)	6.8(0.9)	5.1(0.7)	5.9(0.8)	3.6(0.6)	4.6(0.9)	3.6(0.6)	3.6(0.6)
	50	500	2.7(1.1)	3.2(0.7)	2.0(0.5)	2.4(0.6)	0.8(0.3)	2.5(2.9)	0.9(0.3)	0.9(0.3)
		1000	2.4(1.0)	2.8(0.6)	1.8(0.4)	2.0(0.5)	0.8(0.3)	1.4(0.4)	0.8(0.3)	0.8(0.3)
		2000	2.1(0.9)	2.5(0.5)	1.7(0.4)	1.9(0.5)	0.7(0.3)	1.0(0.3)	0.7(0.3)	0.7(0.3)

3.3 A real example

To demonstrate the practical usefulness of the newly proposed feature screening method, we consider here an example of Chinese document classification, i.e., MPH dataset, which has been introduced in the Introduction. More specifically, the dataset used here consists $n = 13{,}613$ documents from $K = 8$ different functional departments with $p = 4{,}788$ words (or binary features). It is noted that stop words and low-frequency (frequency less than 5) words have been filtered out. Nevertheless, experienced experts tell us only a small number of words are relevant for classification, a large amount of them are irrelevant. Hence, the proposed feature screening method is applied on this dataset for selecting important words and getting higher classification accuracy. To get a reliable conclusion, a total of 100 replications are conducted here. For each replication, half of these documents (i.e., $n = 6{,}806$) are randomly selected for training while the rest are used for testing. We apply three feature screening methods, i.e., PI-SIS, L1NB and L0NB, to the training set separately, then the corresponding models are estimated. The average model size (AMS) of these three estimated models over 100 replications, and the corresponding SE values of them are reported in Table 3. It is noted that, the AMS value is equal to 4.36 of L1NB method with BIC-type criterion for model selection. According to our experience, so few feature words cannot classify the documents to their correct classes for this multi-class classification. Hence, 5-fold cross-validation is adopted for model selection in L1NB, instead of BIC-type criterion in Table 3. For comparison's sake, all competitors (i.e., KNN, AdaB, RF, SVM and NB) are considered on full model and three estimated models. The AMR and their SE values of all cases over 100 replications are also reported in percentage. From Table 3, we find that all classification methods except NB, perform the best on the model selected by L0NB in term of AMR (and SE). Although NB preforms a little worse on L0NB, it is also practically useful in real applications, because only about $289/4{,}778 \approx 6\%$ of features used for classification. This further verifies that the excellent performance of L0NB feature screening method in real applications. We also find that, KNN and SVM perform relatively worse on full model and the models selected by PC-SIS and L1NB, but they preform well on the model selected by L0NB. It indicates that KNN and SVM are fairly sensitive to feature selection. Furthermore, AdaB performs the worst on all models, because only 500 base learners have been used in such high dimensional data. Therefore, in order to improve the classification accuracy of AdaB, more base learners should be used, but it will lead to higher computational cost.

4. CONCLUDING REMARKS

We propose here a L_0-regularization feature selection method based on NB model. But the turning parameter is hard to select in real applications, then a BIC-type screening criterion is proposed to select important features. The screening consistency is investigated under some mild assumptions, which provides theoretical support of mutual information screening method under the naive Bayes assumption. Although the L0NB method is especially proposed for classification of ultrahigh dimensional binary data, the idea can be naturally applied to other categorical or continuous cases. To conclude this article, we discuss two interesting topics for future research. Firstly, interaction screening is a direct future direction that some features are usually used together for improving the prediction accuracy. Secondly, if the conditional independence (1) does not hold, i.e., more complex structures (e.g., tree-structure and clique-structure) can also be learned from the high dimensional binary data. Feature screening for these cases are need to be theoretically investigated in further studies.

APPENDIX A. DERIVATION OF L_0-REGULARIZATION

Because we assume that the model $\mathcal{M}$ contains all potential relevant features, the conditional probability mass function can be represented as

$$P(X_i|Y_i,\mathcal{M}) = \prod_{j=1}^{p}\prod_{k=1}^{K}\left[\left\{\theta_{kj}^{X_{ij}}(1-\theta_{kj})^{1-X_{ij}}\right\}^{\delta_j} \times\left\{\theta_j^{X_{ij}}(1-\theta_j)^{1-X_{ij}}\right\}^{1-\delta_j}\right]^{Z_{ik}}.$$

Thus, the log-likelihood $\mathcal{L}(\mathcal{M}) = \log\prod_{i=1}^{n} P(X_i,Y_i|\mathcal{M})$ is equal to

$$\begin{aligned}
&\sum_{i=1}^{n}\log P(Y_i)+\sum_{i=1}^{n}\log P(X_i|Y_i,\mathcal{M})\\
&=\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{K} Z_{ik}\Big[\delta_j\Big\{X_{ij}\log\theta_{kj}+(1-X_{ij})\log(1-\theta_{kj})\Big\}\\
&\quad+(1-\delta_j)\Big\{X_{ij}\log\theta_j+(1-X_{ij})\log(1-\theta_j)\Big\}\Big]\\
&\quad+\sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\log\pi_k\\
&=n\sum_{k=1}^{K}\hat{\pi}_k^{ML}\log\pi_k\\
&\quad+n\sum_{j=1}^{p}\{\hat{\theta}_j^{ML}\log\theta_j+(1-\hat{\theta}_j^{ML})\log(1-\theta_j)\}\\
&\quad+n\sum_{j=1}^{p}\delta_j\left[\sum_{k=1}^{K}\hat{\pi}_k^{ML}\left\{\hat{\theta}_{kj}^{ML}\log\frac{\theta_{kj}}{\theta_j}\right.\right.\\
&\qquad\left.\left.+(1-\hat{\theta}_{kj}^{ML})\log\frac{1-\theta_{kj}}{1-\theta_j}\right\}\right].
\end{aligned}$$

If we maximize $\mathcal{L}(\mathcal{M})$ directly, all features will be selected. That is not we expect. Next, the original L_0-regularization

Table 3. *Results of Chinese document classification. The AMR and the corresponding SE values of five classification methods (i.e., KNN, AdaB, RF, SVM and NB) based on full model (FULL) and three estimated models (by PC-SIS, L1NB and L0NB) over 100 replications are reported in percentage. The AMS and the corresponding SE values of these models over 100 replications are also reported*

	AMR×10²(SE×10²)					
	KNN	AdaB	RF	SVM	NB	AMS(SE)
FULL	22.4(2.5)	30.7(1.3)	5.8(0.3)	28.8(0.9)	5.6(0.3)	4778(0)
PC-SIS	32.5(14.3)	35.0(6.2)	19.2(17.9)	40.1(15.0)	19.1(18.1)	3058.6(2301.7)
L1NB	20.1(3.6)	30.7(1.3)	5.68(0.3)	27.65(1.8)	5.4(0.4)	4590.8(238.1)
L0NB	11.7(1.6)	30.7(1.3)	4.95(0.3)	9.61(0.7)	5.9(0.2)	289.0(14.6)

log-likelihood should be considered as $\mathcal{L}_p^o = \mathcal{L}(\mathcal{M}) - n\lambda|\mathcal{M}|$. Subsequently, we can get the penalized maximum likelihood estimators (PMLE) of all parameters by maximizing $\mathcal{L}_p^o$. One can easily check that the PMLE of π_k, θ_{kj} and θ_j are same as the MLE of them, because they are not dependent on the regularization constant λ. Then estimate δ based on these estimators. However, MLE is overfit, which leads to serious trouble that $\mathcal{L}(\mathcal{M})$ maybe reach infinity. Thus, we use the DTE instead and obtain the truncated log-likelihood $\mathcal{L}_t(\mathcal{M})$ as

$$n\sum_{k=1}^{K}\hat{\pi}_k\log\hat{\pi}_k + n\sum_{j=1}^{p}\{\hat{\theta}_j\log\hat{\theta}_j + (1-\hat{\theta}_j)\log(1-\hat{\theta}_j)\}$$
$$+n\sum_{j=1}^{p}\delta_j\left[\sum_{k=1}^{K}\hat{\pi}_k\left\{\hat{\theta}_{kj}\log\frac{\hat{\theta}_{kj}}{\hat{\theta}_j} + (1-\hat{\theta}_{kj})\log\frac{1-\hat{\theta}_{kj}}{1-\hat{\theta}_j}\right\}\right].$$

Hence we could work on the modified L_0-regularization log-likelihood $\mathcal{L}_p = \mathcal{L}_t(\mathcal{M}) - n\lambda|\mathcal{M}|$ to select relevant features.

APPENDIX B. DERIVATION OF L_1-REGULARIZATION

Now, we derive the method of L1NB. As expected, the true model can be estimated by $\widehat{\mathcal{M}}_{\lambda_1} = \{1 \le j \le p : \sum_k |\tilde{\theta}_{kj} - \sum_{k'}\tilde{\pi}_{k'}\tilde{\theta}_{k'j}| > 0\}$, after giving a regularization parameter λ_1, where $\tilde{\pi}_k$ and $\tilde{\theta}_{kj}$ are estimators by maximizing the L_1-penalized log-likelihood function (3), Denote $S_{zk} = \sum_i Z_{ik}$, $S_{xj} = \sum_i X_{ij}$ and $S_{zxkj} = \sum_i Z_{ik}X_{ij}$. After some simple mathematical derivations, we have $\tilde{\pi}_k = n^{-1}S_{zk}$ and

$$\tilde{\theta}_{kj} = \frac{1}{2} + \frac{U_{kj}}{2\lambda_1}\Big[S_{zk} - \{(\lambda_1 + U_{kj}S_{zk})^2 - 4\lambda_1 U_{kj}S_{zxkj}\}^{1/2}\Big] + (1-|U_{kj}|)(n^{-1}S_{xj} - 0.5),$$

where $U_{kj} = sign(V_{kj}\lfloor|V_{kj}|\rfloor)$ and $V_{kj} = n\lambda_1^{-1}S_{xj}^{-1}(n - S_{xj})^{-1}(nS_{zxkj} - S_{zk}S_{xj})$. Then, we can immediately have $\tilde{\theta}_{kj} = n^{-1}S_{xj}$ if and only if $nS_{xj}^{-1}(n-S_{xj})^{-1}|nS_{zxkj} - S_{zk}S_{xj}| \le \lambda_1$. Thus the true model can be estimated by

$$\widehat{\mathcal{M}}_{\lambda_1} = \Big\{1 \le j \le p : \max_{1\le k\le K}\{|\tilde{\theta}_{kj} - n^{-1}S_{xj}|\} > 0\Big\}$$
$$= \left\{1 \le j \le p : \max_{1\le k\le K}\left\{\frac{n|nS_{zxkj} - S_{zk}S_{xj}|}{S_{xj}(n-S_{xj})}\right\} > \lambda_1\right\}.$$

Nevertheless, S_{xj} can reach 0 with some positive probability. To avoid zero probability estimates, the DTEs defined in section 2.1 could be adopted here. Then we define $H_j = \max_{1\le k\le K}\{n\hat{\pi}_k\hat{\theta}_j^{-1}(1-\hat{\theta}_j)^{-1}|\hat{\theta}_{kj} - \hat{\theta}_j|\}$ as an approximation of $\max_{1\le k\le K}\{nS_{xj}^{-1}(n-S_{xj})^{-1}|nS_{zxkj} - S_{zk}S_{xj}|\}$, where $\hat{\pi}_k$, $\hat{\theta}_{kj}$ and $\hat{\theta}_j = \sum_k \hat{\pi}_k\hat{\theta}_{kj}$ are the DTEs. Thus the true model can be estimated by $\widehat{\mathcal{M}}_{\lambda_1} = \{1 \le j \le p : H_j > \lambda_1\}$. It's not hard to see that, although there exist infinitely many different choices for λ_1 on $[0,\infty)$, the resulting model $\widehat{\mathcal{M}}_{\lambda_1}$ can be only selected from finite nested candidate models. Consequently, we adopt the same technique in section 2.3 and select the resulting model by BIC-type criterion (2). Alternatively, one can also adopt the cross-validation method for model selection. Some details are omitted here.

APPENDIX C. TECHNICAL LEMMAS

Lemma 1. *Let $(X_i)_{1\le i\le n} \in \{0,1\}^n$ be independent and identically distributed with $P(X_i = 1) = \beta$, where $\nu \le \beta \le 1-\nu$ and $0 < \nu < 0.5$. $\hat{\beta} = \max\{n^{-1}, \min\{1 - n^{-1}, n^{-1}\sum_i X_i\}\}$ is an estimate of β. Then for any $\varepsilon > 0$ and sufficiently large n ($\ge 0.5\nu^{-1}$), we have $P(|\hat{\beta} - \beta| > \varepsilon) \le 2\exp(-2n\varepsilon^2)$.*

PROOF: By Hoeffding's inequality, for any $\varepsilon > 0$, $P(|n^{-1}\sum_i X_i - \beta| > \varepsilon) \le 2\exp(-2n\varepsilon^2)$. Next, under the assumption $\nu \le \beta \le 1-\nu$ and $n^{-1} \le \hat{\beta} \le 1 - n^{-1}$, then $|\hat{\beta} - \beta| \le |n^{-1}\sum_i X_i - \beta|$ for $n \ge 0.5\nu^{-1}$. Hence, for any $\varepsilon > 0$ and sufficiently large n ($\ge 0.5\nu^{-1}$), we have $P(|\hat{\beta} - \beta| > \varepsilon) \le P(|n^{-1}\sum_i X_i - \beta| > \varepsilon) \le 2\exp(-2n\varepsilon^2)$.

Lemma 2. *Let $\beta = (\beta_1, \cdots, \beta_m)^\top \in \Theta = \Theta_1 \times \cdots \times \Theta_m$, where each Θ_j is a closed interval on $\mathbb{R}^1$. $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_m)^\top \in \Theta$ is an estimate of β. $f(\cdot)$ is a multivariate function, with bounded first-order partial derivatives on Θ. If for any $\varepsilon > 0$, $P(|\hat{\beta}_j - \beta_j| > \varepsilon) \le C_{j1}\exp(-C_{j2}n\varepsilon^2)$, where $\{C_{j1}, C_{j2}\}_{1\le j\le m}$ are some positive constants. Then for any $\varepsilon > 0$, we have*

$$P(|f(\hat{\beta}) - f(\beta)| > \varepsilon) \le C_1\exp(-C_2n\varepsilon^2),$$

where C_1 and C_2 are some positive constants.

PROOF: By Lagrange's mean value theorem, $f(\hat{\beta}) - f(\beta) = (\hat{\beta} - \beta)^\top\dot{f}(\tilde{\beta})$, where $\tilde{\beta}$ lies between $\hat{\beta}$ and β, and $\dot{f}(\tilde{\beta}) =$

$(\dot{f}_1(\tilde{\beta}), \cdots, \dot{f}_m(\tilde{\beta}))^\top$ is the gradient vector at $\tilde{\beta}$. Define random events $\mathcal{A}_j = \{|\hat{\beta}_j - \beta_j| \le m^{-1}M^{-1}\varepsilon\}$ for $1 \le j \le m$, and $\mathcal{A} = \{|f(\hat{\beta}) - f(\beta)| \le \varepsilon\}$, where $M = \max_j \sup_{\beta\in\Theta}\{|\dot{f}_j(\beta)|\}$. Then in event $\cap\mathcal{A}_j$, we have

$$|f(\hat{\beta}) - f(\beta)| \le m \max_j \{|\dot{f}_j(\tilde{\beta})|\} \max_j \{|\hat{\beta}_j - \beta_j|\} \le \varepsilon.$$

Hence we have $\cap\mathcal{A}_j \subset \mathcal{A}$. Subsequently, for any $\varepsilon > 0$,

$$P(\mathcal{A}^c) \le P(\cup\mathcal{A}_j^c) \le \sum_{j=1}^m P(\mathcal{A}_j^c)$$

$$\le \sum_{j=1}^m C_{j1}\exp\big(-C_{j2}nm^{-2}M^{-2}\varepsilon^2\big) \le C_1\exp(-C_2n\varepsilon^2).$$

where $C_1 = m\max_j\{C_{j1}\}$, $C_2 = m^{-2}M^{-2}\min_j\{C_{j2}\}$.

Lemma 3. *Let $0 < \nu < 1/3$, for any $(a,b) \in \Omega = \{(x,y) : \nu \le x \le 1-\nu, \nu \le y \le 1-\nu, |x-y| \ge \nu\}$, then we have*

$$\nu\log\frac{1+\nu}{1-\nu} \le a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b} \le (1-2\nu)\log\frac{1-\nu}{\nu}.$$

Proof: Let $g(x,y) = x\log(y^{-1}x) + (1-x)\log\{(1-y)^{-1}(1-x)\}$. Thus, it is equivalent to compute the minimum and maximum of $g(x,y)$ on Ω. We then define two univariate functions $g_0(x) = \min_{\nu\le y\le 1-\nu, |x-y|\ge\nu}\{g(x,y)\}$ and $g_1(x) = \max_{\nu\le y\le 1-\nu,|x-y|\ge\nu}\{g(x,y)\}$ on the domain $[\nu, 1-\nu]$. After some simple mathematical derivations, the analytical expressions of $g_0(x)$ and $g_1(x)$ can be obtained. Hence, it is not hard to show that $\min_{\nu\le x\le 1-\nu}\{g_0(x)\} = \nu\log\{(1-\nu)^{-1}(1+\nu)\}$ and $\max_{\nu\le x\le 1-\nu}\{g_1(x)\} = (1-2\nu)\log\{\nu^{-1}(1-\nu)\}$. Details are omitted here for brevity.

APPENDIX D. PROOF OF THEOREM 1

In order to prove Theorem 1, firstly denote $P(X_{ij}Z_{ik} = 1) = \pi_k\theta_{kj} = \mu_{kj}$. Subsequently, $\theta_{kj} = \pi_k^{-1}\mu_{kj}$, $\theta_j = \sum_k \mu_{kj}$ and $\hat{\mu}_{kj} = \max\{n^{-1}, \min\{1-n^{-1}, n^{-1}\sum_i Z_{ik}X_{ij}\}\}$. Together with $\hat{\pi}_k = \max\{n^{-1}, \min\{1-n^{-1}, n^{-1}\sum_i Z_{ik}\}\}$ for $1 \le k \le K-1$ and $\hat{\pi}_K = 1 - \sum_{k=1}^{K-1}\hat{\pi}_k$, we can rewrite I_j and $\hat{I}_j$ as follows,

$$\begin{aligned} I_j &= \sum_{k=1}^K \Big[\mu_{kj}\log\frac{\mu_{kj}}{\pi_k\sum_{k'}\mu_{k'j}} \\ &\quad + (\pi_k - \mu_{kj})\log\frac{\pi_k - \mu_{kj}}{\pi_k - \pi_k\sum_{k'}\mu_{k'j}}\Big], \\ \hat{I}_j &= \sum_{k=1}^K \Big[\hat{\mu}_{kj}\log\frac{\hat{\mu}_{kj}}{\hat{\pi}_k\sum_{k'}\hat{\mu}_{k'j}} \\ &\quad + (\hat{\pi}_k - \hat{\mu}_{kj})\log\frac{\hat{\pi}_k - \hat{\mu}_{kj}}{\hat{\pi}_k - \hat{\pi}_k\sum_{k'}\hat{\mu}_{k'j}}\Big]. \end{aligned}$$

By the conclusions of Lemmas 1 and 2, for any $\varepsilon > 0$ and sufficiently large n ($\ge 0.5\nu^{-1}$), we have

$$P(|\hat{I}_j - I_j| > \varepsilon) \le C_1\exp(-C_2n\varepsilon^2),$$

where C_1 and C_2 are some positive constants. Next, let $\gamma = \sqrt{2/C_2}$, by Bonferroni's inequality,

$$\begin{aligned} &P\big(\max_j|\hat{I}_j - I_j| > \gamma\sqrt{\log p/n}\big) \\ &\le \sum_{j=1}^p P\big(|\hat{I}_j - I_j| > \gamma\sqrt{\log p/n}\big) \\ &\le pC_1\exp(-C_2\gamma^2\log p) = C_1\exp(-\log p) \to 0, \end{aligned}$$

Consequently, we know that $\max_j|\hat{I}_j - I_j| = O_P(\sqrt{\log p/n})$. By the boundedness assumption and Lemma 3, it is clear that $\min_{j\in\mathcal{M}_T} I_j \ge K\tau\nu$ for some $\tau > 0$ and $I_j = 0$ for $j \notin \mathcal{M}_T$. We then set $\lambda_{max} = K\tau\nu$. For $0 < \lambda < \lambda_{max}$ and $\log p = o(n)$, we have

$$\begin{aligned} P(\widehat{\mathcal{M}}_\lambda = \mathcal{M}_T) &= P\Big(\min_{j\in\mathcal{M}_T}\hat{I}_j > \lambda, \max_{j\notin\mathcal{M}_T}\hat{I}_j < \lambda\Big) \\ &\ge P\Big(\min_{j\in\mathcal{M}_T}\hat{I}_j > \lambda\Big) + P\Big(\max_{j\notin\mathcal{M}_T}\hat{I}_j < \lambda\Big) - 1 \\ &\ge P\Big(\max_{j\in\mathcal{M}_T}|\hat{I}_j - I_j| < \lambda_{max} - \lambda\Big) \\ &\quad + P\Big(\max_{j\notin\mathcal{M}_T}|\hat{I}_j - I_j| < \lambda\Big) - 1 \\ &\to 1. \end{aligned}$$

This completes the proof.

APPENDIX E. PROOF OF THEOREM 2

By the conclusion of Theorem 1, after given some appropriate regularization constant λ, we have $P(\mathcal{M}_T \in \mathbb{M}) \to 1$. It implies that $P(\mathcal{M}_{(d_0)} = \mathcal{M}_T) \to 1$, where $d_0 = |\mathcal{M}_T|$ is the true model size. To get the conclusion of Theorem 2, the following inequality should be needed, $P\big(\mathcal{M}_T \subset \widehat{\mathcal{M}}\big) \ge P\big(\mathcal{M}_T \subset \widehat{\mathcal{M}}\big|\mathcal{M}_{(d_0)} = \mathcal{M}_T\big)P\big(\mathcal{M}_{(d_0)} = \mathcal{M}_T\big)$. Now, the only thing left is $P(BIC_{\mathcal{M}_{(d)}} > BIC_{\mathcal{M}_{(d_0)}}) \to 1$ for $\mathcal{M}_{(d)} \subset \mathcal{M}_{(d_0)} = \mathcal{M}_T$.

By the BIC-type criterion (2) and the truncated log-likehood $\mathcal{L}_t(\mathcal{M})$, we can obtain the difference of BIC values between two models as

$$BIC_{\mathcal{M}_{(d)}} - BIC_{\mathcal{M}_{(d_0)}} = 2\sum_{j\in\mathcal{S}}\hat{I}_j - (d_0 - d)(K-1)\frac{\log n}{n},$$

where $\mathcal{S} = \mathcal{M}_{(d_0)} \setminus \mathcal{M}_{(d)}$. Subsequently, we know that

$$\begin{aligned} &P\Big(BIC_{\mathcal{M}_{(d)}} \le BIC_{\mathcal{M}_{(d_0)}}\Big) \\ &= P\Big(\sum_{j\in\mathcal{S}}\hat{I}_j \le (d_0 - d)(K-1)\frac{\log n}{2n}\Big). \end{aligned} \tag{4}$$

By Bonferroni's inequality, the right hand side of (4) can be further bounded by

$$P\Big(\min_{j\in\mathcal{S}}\hat{I}_j \le (K-1)\frac{\log n}{2n}\Big) \tag{5}$$

$$\leq \sum_{j\in\mathcal{S}} P\left(\hat{I}_j \leq (K-1)\frac{\log n}{2n}\right).$$

Because $\mathcal{M}_{(d)} \subset \mathcal{M}_{(d_0)} = \mathcal{M}_T$, then $\min_{j\in\mathcal{S}} I_j \geq K\tau\nu$. For sufficiently large n, $(K-1)\log n/(2n) \leq K\tau\nu/2$, the right hand side of (5) can be bounded by

$$\begin{aligned}(6)\qquad &\sum_{j\in\mathcal{S}} P\big(\hat{I}_j \leq I_j - K\tau\nu/2\big)\\ &\leq \sum_{j\in\mathcal{S}} P\big(\big|\hat{I}_j - I_j\big| \geq K\tau\nu/2\big)\\ &\leq \exp\big\{\log(d_0-d) + \log C_1 - C_2 n(K\tau\nu/2)^2\big\}.\end{aligned}$$

Under the assumption $\log p = o(n)$, the right hand side of (6) converges towards 0, as $n \to \infty$. Therefore, we have $P(BIC_{\mathcal{M}(d)} > BIC_{\mathcal{M}_{(d_0)}}) \to 1$ for $\mathcal{M}_{(d)} \subset \mathcal{M}_{(d_0)} = \mathcal{M}_T$, which is equivalent to the screening consistency result $P(\mathcal{M}_T \subset \widehat{\mathcal{M}}) \to 1$. The proof is completed.

ACKNOWLEDGEMENTS

We would like to thank the editor, an associate editor, and two anonymous referees for their constructive comments and suggestions that helped us to improve the manuscript. The research of Guoyu Guan is supported in part by National Natural Science Foundation of China (No.11501093), China Postdoctoral Science Foundation Funded Project (No.2015M581378), and the Fundamental Research Funds for the Central Universities (No.2412015KJ028,130028613). The research of Na Shan is supported in part by National Natural Science Foundation of China (No. 11401047, 11571050) and the Project of the Educational Department of Jilin Province of China (2016315). The research of all the authors is supported by National Natural Science Foundation of China (No.11631003, 11690012).

Received 14 September 2016

REFERENCES

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. MR2247587

[2] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

[3] 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. MR2530322

[4] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research* **3** 1289–1306.

[5] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* **3** 1157–1182.

[6] Hand, D. J. and Yu, K. (2001). Idiot's Bayes: not so stupid after all? *International Statistical Review* **69** 385–398.

[7] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd ed. Springer, New York. MR2722294

[8] Huang, D., Li. R. and Wang, H. (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business and Economic Statistics* **32** 237–244. MR3207836

[9] Lewis, D. D. (1998). Naive Bayes at forty: the independence assumption in information retrieval. *Proceedings of ECML-98, 10th European Conference on Machine Learning* 4–15.

[10] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press. MR1722790

[11] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

[12] Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics* **42** 789–817. MR3210987

[13] Peng, H., Long, F. and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27** 1226–1238.

[14] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6** 461–464. MR0468014

[15] Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223–232. MR2949354

[16] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288. MR1379242

[17] Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104** 1512–1524. MR2750576

[18] Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* **104** 747–757. MR2541592

[19] Wu, X. and Kumar, V. (2008). The top ten algorithms in data mining. *Knowledge and Information Systems* **14** 1–37. MR2779331

[20] Yang, Y. and Pedersen, J. (1997). A comparative study of feature selection in text categorization. *Proceedings of ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 412–420.

[21] Zhang, H. and Singer, B. H. (2010). *Recursive Partitioning and Applications*. Springer, New York. MR2674991

[22] Zhu, L., Li, L., Li, R. and Zhu, L. (2011). Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association* **106** 1464–1475. MR2896849

Guoyu Guan
Key Laboratory for Applied Statistics of MOE
School of Economics
Northeast Normal University
Changchun 130024, Jilin Province
China
E-mail address: guangy599@nenu.edu.cn

Na Shan
Key Laboratory for Applied Statistics of MOE
School of Psychology
Northeast Normal University
Changchun 130024, Jilin Province
China
E-mail address: dann285@nenu.edu.cn

Jianhua Guo
Key Laboratory for Applied Statistics of MOE
School of Mathematics and Statistics
Northeast Normal University
Changchun 130024, Jilin Province
China
E-mail address: jhguo@nenu.edu.cn