JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 542, 805–817: Applications and Case Studies
https://doi.org/10.1080/01621459.2022.2152342

Feature Screening for Interval-Valued Response with Application to Study Association between Posted Salary and Required Skills

Wei Zhong[a], Chen Qian[b], Wanjun Liu[c], Liping Zhu[d], and Runze Li[e]

[a]Xiamen University, Xiamen, China; [b]Virginia Tech, Blacksburg, VA; [c]LinkedIn, Mountain View, CA; [d]Renmin University of China, Beijing, China; [e]Pennsylvania State University, University Park, PA

ABSTRACT

It is important to quantify the differences in returns to skills using the online job advertisements data, which have attracted great interest in both labor economics and statistics fields. In this article, we study the relationship between the posted salary and the job requirements in online labor markets. There are two challenges to deal with. First, the posted salary is always presented in an interval-valued form, for example, 5k–10k yuan per month. Simply taking the mid-point or the lower bound as the alternative for salary may result in biased estimators. Second, the number of the potential skill words as predictors generated from the job advertisements by word segmentation is very large and many of them may not contribute to the salary. To this end, we propose a new feature screening method, Absolute Distribution Difference Sure Independence Screening (ADD-SIS), to select important skill words for the interval-valued response. The marginal utility for feature screening is based on the difference of estimated distribution functions via nonparametric maximum likelihood estimation, which sufficiently uses the interval information. It is model-free and robust to outliers. Numerical simulations show that the new method using the interval information is more efficient to select important predictors than the methods only based on the single points of the intervals. In the real data application, we study the text data of job advertisements for data scientists and data analysts in a major China's online job posting website, and explore the important skill words for the salary. We find that the skill words like optimization, long short-term memory (LSTM), convolutional neural networks (CNN), collaborative filtering, are positively correlated with the salary while the words like Excel, Office, data collection, may negatively contribute to the salary. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2021
Accepted November 2022

KEYWORDS
Feature screening; Interval-valued response; Job advertisements; Returns to skills; Ultrahigh dimensionality

1. Introduction

It is of fundamental importance to depict the differences in returns to skills in labor economics because it can thoroughly present the detailed structure of wages, which is beneficial to understand the systematical demand and supply change in labor markets. Card and DiNardo (2002) claimed that the effect of technology on the labor market associated with differences in returns to skills has long been a core concern for economists. Frank et al. (2019) stated that depicting the returns to granular skills can capture the dynamic nature of the job markets. Many empirical studies have been conducted to estimate the returns to skills, for example, cognitive skills and social skills (Beaudry, Green, and Sand 2016). This work was motivated by an empirical analysis of data collected from online job advertisements. Comparing with the occupational requirements survey (ORS) conducted by U.S. Bureau of Labor Statistics, online job advertisements timely reflects the skills needed in the current job markets. Thus, analysis of online job advertisements is of great importance, and may provide insights into the current needs of job market more timely than the conventional methods such as the ORS used by labor economists. Recently, the trending online job markets, led by LinkedIn, embark convenient ways for employment and also provide accessible ways to explore more details toward the labor markets. Kuhn and Shen (2013) found at least 11.2% Chinese companies used online job websites to conduct the employment in 2012. This percentage has grown substantially with the fast development of the internet industry. Online job postings cover adequate employment information including job title, job description, job requirements, the expected salary, benefits, etc. Marinescu and Wolthoff (2020) found that the job title plays a critical role in explaining the number and quality of applicants. Hershbein and Kahn (2018) investigated how the demand for skills changes over the business cycle. Online job postings provide an informative way for labor economists and statisticians to investigate the differences among returns to skills and establish a thorough understanding of the structure of the job markets. As detailed below, interval-valued response and high-dimensionality of predictor vectors are two major challenges in analyzing job advertisement data. Interval-valued response implies that the response is interval censored, and it is not uncommon in various job advertisement database including newspaper as well as online job listings. Furthermore, most job advertisements include many key words to describe the job precisely and attractively. This requires to create high-dimensional predictors for analysis of job advertisement

CONTACT Runze Li *ril4@psu.edu* Pennsylvania State University, University Park, PA
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

data. The statistical procedures developed in this article are applicable for various job databases including both online and newspaper job advertisements. The new procedures are also broadly applicable for other research areas where people collect interval-valued responses along with ultrahigh dimensional predictors.

In the analysis of job advertisement data, people aim to establish the relationship between the posted salary and the job requirements using the text data of online job advertisements. However, there are at least two challenges to deal with from a statistical perspective. The first challenge is how to deal with the interval-valued data. In the job postings, the posted salary is always presented in an interval-valued form, for example, 5k–10k Chinese yuan per month. In other words, we can not observe the real salary for a posted job but an interval which is believed to include the real salary. In many empirical analyses, the mid-point or the lower bound of the interval is simply taken as the alternative for the salary, for instance, Marinescu and Wolthoff (2020). However, the interval-valued data contain both the scale and position information as discussed in Sun et al. (2018). Simply taking the single point to represent the interval information may result in biased estimation. In this article, we will treat the interval-valued form of the salary in the online job markets as the case 2 interval censoring which includes interval censoring (e.g., 5k–10k yuan per month), right censoring (e.g., at least 20k yuan per month) and left censoring (e.g., up to 15k yuan per month). In the literature on interval censoring, some researchers have studied nonparametric maximum likelihood estimation of the distribution function of the interval-censored data and its asymptotic properties, such as Aragón and Eberly (1992), Wellner and Zhan (1997), Geskus and Groeneboom (1999) and among others. These existing works on interval censoring pave a way to construct a dependence measure between the interval-valued posted salary and a potential skill.

The second challenge is how to deal with the ultrahigh dimensionality of the potential skill words as predictors generated from the job advertisements by word segmentation. In the literature such as Deming and Kahn (2018), Modestino, Shoag, and Ballance (2016), and Hershbein and Kahn (2018), labor economists always first defined a small pool of job skills, such as cognitive abilities and social skills, to investigate the returns to different skills. The prespecified small group of job skills may ignore some important information in the job postings and lead to the omitted-variable bias. Instead, to sufficiently use the textual information in the job postings, we generate all potential skill words by word segmentation and define the predictor $X_{ik} = 1$ if word k exists in job i and $X_{ik} = 0$ otherwise. As the result, the number of the generated predictors is very large and many of them may not contribute to the salary. It is necessary to select the important words for the salary before establishing the regression between them. In the literature, feature screening approaches have been proposed to screen out the unimportant variables and select the important ones for ultrahigh dimensional data. In the seminal work of Fan and Lv (2008), the Sure Independence Screening (SIS) was proposed to rank variables based on the magnitude of Pearson's correlation between each variable and the response and its sure screening property has been theoretically established. The essence of a feature screening is to construct a proper marginal utility to measure the importance of each covariate for the response. Most feature screening procedures can be classified into two types. The first type is model-based, including Fan and Song (2010) for generalized linear models, Fan, Feng, and Song (2011) for nonparametric additive models, etc. The second type is model-free based on various dependence measures, such as distance correlation (Li, Zhong, and Zhu 2012), Kolmogorov-Smirnov test statistic (Mai and Zou 2013), martingale difference correlation (Shao and Zhang 2014), mean-variance index (Cui, Li, and Zhong 2015), ball correlation (Pan et al. 2019) and among others. However, the existing methods can not directly be applied to the interval-valued response.

In this article, we are motivated by the real application in labor economics to explore the relationship between the posted interval-valued salary and the ultrahigh dimensional skill words in job advertisements of a certain job category. To deal with the first challenge, we treat the interval-valued posted salary as the case 2 interval censoring and define a new dependence measure, Absolute Distribution Difference (ADD), between an interval-valued response and each binary covariate. The estimation of the new ADD measure is based on the nonparametric maximum likelihood estimation of the distribution function of the interval-censored data. Geometrically, the ADD characterizes the area of the absolute difference between two distributions of the salary with/without the given word. It is able to sufficiently retain the distribution information of the interval-valued data. To deal with the second challenge, a new ADD-based sure independence screening (ADD-SIS) is proposed to rank the ultrahigh dimensional predictors generated by word segmentation. The sure screening property is established. Numerical simulations show that the new ADD-SIS using the interval information is more efficient to select important predictors than the methods only based on the single point of the interval. In the real data analysis, we study the text data of job advertisements for data scientists and data analysts in a major China's online job posting website, and explore the important skill words for the salary. It is interesting to find that the skill words like "optimization," "LSTM," "CNN," "collaborative filtering," are positively correlated with the salary while the words like "Excel," "data collection," "data management" may negatively contribute to the salary.

We conclude this section by summarizing the main contributions of this work to the field of labor economics. First, it provides researchers in labor economics studies an effective approach to identifying the important skills for better pay from the certain job advertisements data, which is more comprehensive than the predefined pool of job skills to quantify the differences among returns to skills. Second, we establish the association between the salary and the selected skill words using the accelerated failure time (AFT) model. A good return-to-skills association model can benefit both employees and employers in the job markets. For employees, especially for new graduates and people switching jobs, knowing the expected salary based on their experienced job skills may help them learn the current market value to find satisfactory employment. It can help employers set an appropriate and competitive salary in their new postings and intelligently optimize hiring plans. It may also benefit the job posting platforms (such as LinkedIn) which may use the proposed procedures for better matching between

employees and employers. Third, the proposed procedures may help undergraduate and graduate programs to adjust their curriculums so that their students are well prepared for their target jobs. For example, our empirical data analysis in Section 5 suggests that the graduate programs in applied statistics could add new courses on modern machine learning procedures like collaborative filtering and deep learning, and popular analytic softwares like Python, etc. These new skills can effectively help students in statistics to meet the demand of the high-wage jobs in the modern data science era.

The rest of this article is organized as follows. We first introduce the data and the associated statistical problem in Section 2. Then, we define the Absolute Distribution Difference (ADD) and the ADD-based feature screening procedure in Section 3. In Section 4, we evaluate the finite sample performance via Monte Carlo simulations. An application on empirical analysis of job advertisements on data scientists/data analysts is conducted in Section 5. Section 6 concludes the article.

2. Data and Setups

2.1. Data

We take the job category of data scientists and data analysts, which are the major target job positions of students majoring in statistics, as an example to explore the critical skills from the job advertisements. The data for our analysis were scraped from the 51Job website (*https://www.51job.com*). This website is one of the largest online job platforms offering comprehensive recruitment solutions, targeting at serving the skilled workers in most Chinese cities. Employers can post job advertisements, contact applicants and arrange interviews through the website. Each job advertisement is comprised of several essential parts: the job title such as data scientist, the company information including its name and address, the expected salary which is the response of interest in our study and typically presented in the interval-valued form, the job requirements which illustrate the necessary job skills for applicants, such as deep learning, Python, Ph.D. degree, etc.

In our scraped data, we constrain the job titles to be either data analysts or data scientists posted only in September, 2019, which eliminates the potential time-varying effect. We also set our samples only from the tier-one cities (Beijing, Shanghai, Shenzhen, Guangzhou) in China to control the geographical heterogeneity issue. This dataset consists of 497 unique job advertisements. Figure 1 presents two typical samples of job advertisements. The left sample for data scientist is a high-wage job with the expected salary at 50k–70k yuan/month while the right one for data analyst only offers the expected salary at 4k–8k yuan/month. According to a simple comparison between two sample jobs, we can observe that the high-wage job requires the high-level education degree, modern machine learning techniques such as RNN, LSTM, and popular softwares for data science such as Python, TensorFlow. On the other hand, the low-wage job only needs applicants to be familiar with Excel, SAS, SPSS or other traditional statistical data analysis tools. Clearly, the expected salary is related to the job requirements. Later on, we will generate the potential skill words as predictors from the job advertisements by word segmentation, explore the important skill words for the expected salary and establish the regression relationship between the expected salary and the important skill words based on the survival models.

2.2. Setups

Let n denote the total number of job advertisements in our data. The response of interest is the posted salary for the ith job posting, denoted by Y_i, for $i = 1, 2, \ldots, n$. In general, the exact real value of Y_i can not be observed in the job postings. Instead, the employers present the expected salary in the interval-valued form which includes the real value of Y_i for the recruited employee, such as 50k–70k yuan/month. Besides, the minimum or the maximum of the expected salary may be presented in some special cases, for example at least 20k yuan per month or up to 25k yuan per month, respectively. In this work, we treat the interval-valued form of the salary data as the case 2 interval censoring (Aragón and Eberly 1992; Wellner and Zhan 1997; Yu et al. 1998; Geskus and Groeneboom 1999) which includes interval censoring, right censoring and left censoring. The observed interval-valued data can be defined as,

$$\left\{(y_i^L, y_i^U, \mathrm{I}(y_i \le y_i^L), \mathrm{I}(y_i^L < y_i \le y_i^U), \mathrm{I}(y_i > y_i^U)), i = 1, \ldots, n\right\}, \tag{2.1}$$

where y_i is the real value of the salary for the ith job which is generally unknown, (y_i^L, y_i^U) is the lower bound and the upper bound of the salary for the ith job, $\mathrm{I}(y_i \le y_i^L)$, $\mathrm{I}(y_i^L < y_i \le y_i^U)$ and $\mathrm{I}(y_i > y_i^U)$ are indicator functions for left censoring, interval censoring and right censoring, respectively. Clearly, $\mathrm{I}(y_i \le y_i^L) + \mathrm{I}(y_i^L \le y_i \le y_i^U) + \mathrm{I}(y_i > y_i^U) = 1$.

Next, we generate the potential skill words as predictors by word segmentation via Tsinghua University Lexical Analyzer for Chinese (THULAC)[1] from the job advertisements. We drop out all the stop words and collect $p = 1043$ words in the whole text corpora. We transform these words into binary predictors in the following way: for any specific word k ($1 \le k \le p$), the predictor $X_{ik} = 1$ if word k existed in job i and $X_{ik} = 0$ vice versa. Because the total number of Chinese characters in a given corpora is very large, the dimension p of predictors is typically very high. Obviously, not all generated predictors can contribute to the job salary. Thus, we need to introduce a new feature screening procedure to filter out trivial information and select the important predictors for the interval-valued salary.

3. Statistical Methodology

3.1. A New Dependence Statistic

To quantify the importance of each predictor for the response variable Y, we need to measure the dependence between Y and the kth binary predictor X_k for any $k = 1, 2, \ldots, p$. In a natural way, we can compare the conditional distribution functions given X_k. Let $F_{+k}(y) = \Pr(Y \le y | X_k = 1)$ and $F_{-k}(y) = \Pr(Y \le y | X_k = 0)$ be the conditional distribution functions given $X_k = 1$ and $X_k = 0$, respectively. If $F_{+k}(y) = F_{-k}(y)$ for all $y \in \mathbb{R}$, then Y and X_k are independent. This motivates us to

[1] *https://github.com/thunlp/THULAC-Python*

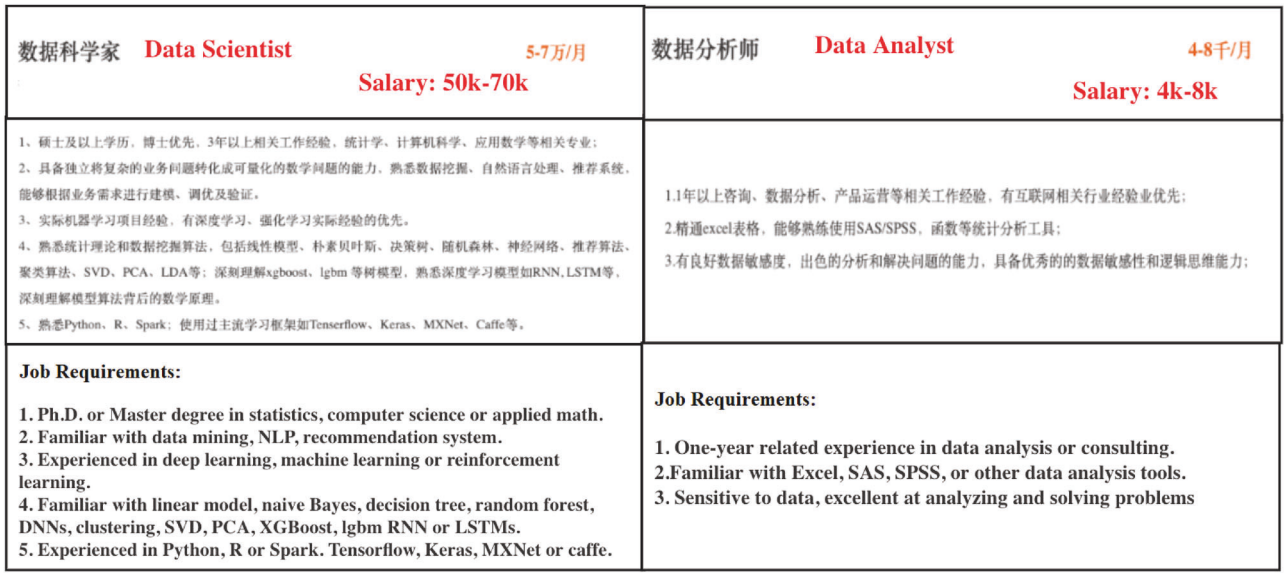

Figure 1. Two sample job advertisements.

consider a new dependence statistic between Y and the binary predictor X_k,

$$\omega_k \triangleq \int_{t\in\mathbb{R}} |F_{+k}(t) - F_{-k}(t)| dW(t), \tag{3.1}$$

where $W(t)$ is the cumulative distribution function (CDF) of some random variable T. For example, if $W(\cdot)$ is the CDF of the true response Y although it is generally unknown, then the statistic becomes $\omega_k = \int_{y\in\mathbb{R}} |F_{+k}(y) - F_{-k}(y)| dW(y) = E(|F_{+k}(Y) - F_{-k}(Y)|)$, which is the mean of $|F_{+k}(Y) - F_{-k}(Y)|$. In this article, we will choose a specific CDF for $W(\cdot)$ to define a new statistic with an appealing geometric interpretation in the next paragraph. It is worth noting that Y and X_k are statistically independent if and only if $\omega_k = 0$.

Next, we assume that the distribution of the unobserved response variable Y is arbitrary, but the joint distribution of the lower and upper bounds (Y^L, Y^U) is discrete. This assumption has been often used in the literature on interval censoring (Finkelstein and Wolfe 1985; Wellner and Zhan 1997; Yu et al. 1998). It is also reasonable in our empirical analysis because all the posted salary values, no matter the upper bounds or the lower bounds, are presented in the positive integers in the thousands which are at least 3k yuan/month and at most 80k yuan/month in our data. In other words, the number of all possible values of Y^L and Y^U can be assumed to be finite in practice. Let $\mathcal{A} = \{a \in \mathbb{R} : P(Y^L = a) + P(Y^U = a) > 0\} = \{y_1^*, y_2^*, \ldots, y_L^*\}$ be the union set of all possible values for the lower and upper bounds where $L < \infty$ is the cardinality of the set $\mathcal{A}$. Without loss of generality, we assume that $y_1^* < y_2^* < \cdots < y_L^*$. Then, we specifically set $W(t)$ be the CDF of a discrete random variable with the probability mass function,

$$p_i = P(T = y_i^*) = \frac{y_{i+1}^* - y_i^*}{y_L^* - y_1^*} = \frac{\Delta y_i^*}{y_L^* - y_1^*}, \quad i = 1, 2, \ldots, L, \tag{3.2}$$

where $\Delta y_i^* = y_i^* - y_{i-1}^*$ for $i = 2, 3, \ldots, L$ and $\Delta y_1^* = 0$. Using this specific CDF $W(t)$, the population level of the new dependence measure ω_k becomes

$$\omega_k = \sum_{i=1}^{L} |F_{+k}(y_i^*) - F_{-k}(y_i^*)| \frac{\Delta y_i^*}{y_L^* - y_1^*}. \tag{3.3}$$

A natural interpretation of the statistic ω_k defined in (3.3) is the area of the absolute difference between two conditional distribution functions, $F_{+k}(\cdot)$ and $F_{-k}(\cdot)$, of the response variable Y given $X_k = 1$ and $X_k = 0$, respectively, up to a constant $1/(y_L^* - y_1^*)$. Figure 2 illustrates the geometric interpretation of the statistic ω_k. The larger the statistic ω_k is, the more the discrepancy of two conditional distributions is, then the more important the kth predictor X_k is for the response. Because of this reason, we name this new dependence measure ω_k defined in (3.3) as the Absolute Distribution Difference (ADD) statistic between the response variable Y and the binary predictor X_k. We will use the ADD statistic to rank the importance of all X_k's for the response Y.

We remark that the proposed ADD statistic is related to the Kolmogorov-Smirnov (KS) statistic, $\sup_{y_i^*} |F_{+k}(y_i^*) - F_{-k}(y_i^*)|$. It has been used in Mai and Zou (2013) to filter the important continuous covariates for the binary classification problem with ultrahigh dimensional covariates. Note that the KS statistic tends to be more sensitive near the center of the distributions than at the tails. It only measures the supremum distance between two conditional distribution functions and ignores their differences at other points. However, the ADD statistic considers the differences between two conditional distribution functions at all possible points and geometrically quantifies the area of the absolute difference between two conditional distribution functions. Thus, it is expected that the ADD statistic is less sensitive to outliers and more powerful to measure the dependence between the response and the binary predictor.

Next, we estimate the ADD statistic ω_k in (3.3). The key is to estimate the conditional distribution functions, $F_{+k}(\cdot)$ and $F_{-k}(\cdot)$. If one can observe the exact values of Y, then the empirical cumulative distribution function (ECDF) can be directly used to estimate the conditional distribution functions and then ω_k. However, we only observe the interval-valued data which are treated as the case 2 interval censored data. In the literature,

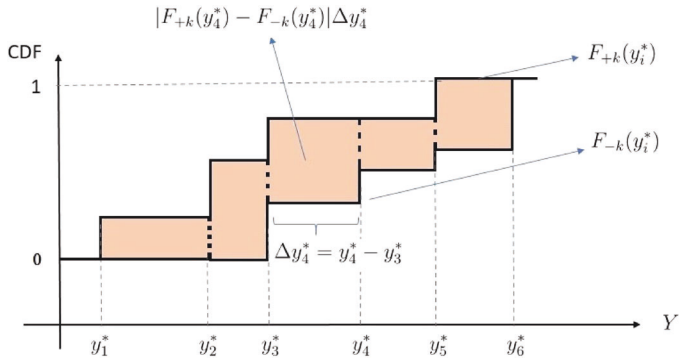

Figure 2. The geometric illustration of the ADD statistic.

the nonparametric maximum likelihood estimation (NPMLE) has been successfully used to estimate the distribution function for the interval censored data (Turnbull 1976; Groeneboom and Wellner 1992; Wellner and Zhan 1997; Yu et al. 1998).

Let $\mathcal{F}^*$ be the set of discrete distributions that are piecewise constant between the points $y_1^* < y_2^* < \cdots < y_L^*$. Any function $F \in \mathcal{F}^*$ can have jumps only at these finite points. A function $F \in \mathcal{F}^*$ can be identified with a vector $(F(y_1^*), F(y_2^*), \ldots, F(y_L^*))$, where $F(y_i^*)$ is the value of F at the point y_i^*. For the observed interval-valued data $\{(y_i^L, y_i^U, \mathrm{I}(y_i \le y_i^L), \mathrm{I}(y_i^L < y_i \le y_i^U), \mathrm{I}(y_i > y_i^U)), i = 1, \ldots, n\}$, the likelihood function for F, the discrete distribution function Y defined in the set $\mathcal{A}$, can be represented as

$$\mathcal{L}_n(F) = \prod_{i=1}^{n} [F(y_i^L)]^{N_n^L} [F(y_i^U) - F(y_i^L)]^{N_n^I} [1 - F(y_i^U)]^{N_n^R}, \tag{3.4}$$

where $N_n^L = \sum_{j=1}^n \mathrm{I}(y_i \le y_i^L), N_n^I = \sum_{j=1}^n \mathrm{I}(y_i^L < y_i \le y_i^U), N_n^R = \sum_{j=1}^n \mathrm{I}(y_i > y_i^U)$. Note that the union set of all observed values for the lower and upper bounds, $\{y_i^L, y_i^U, i = 1, \ldots, n\}$, is a subset of $\mathcal{A}$. The nonparametric maximum likelihood estimator (NPMLE) $\widehat{F}_n$ is the maximizer of the likelihood function $\mathcal{L}_n(F)$ over the class of distribution functions. Turnbull (1976) introduced self-consistency equations for computing the maximum likelihood estimator of the survival function. But a self-consistent estimate is not necessarily the NPMLE in the case of double censoring (Gu and Zhang 1993). To this end, Wellner and Zhan (1997) designed a hybrid EM-ICM algorithm combining the expectation-maximization (EM) algorithm and the modified iterative convex minorant (ICM) algorithm (Groeneboom and Wellner 1992) for computation of the NPMLE from censored data. An advantage of using a composite mapping is that the EM iteration never destroys the ascent likelihood function in the modified ICM algorithm. The hybrid EM-ICM algorithm has been shown to converge to the NPMLE. For the details of the hybrid algorithm, one can refer to Section S.2 in our supplementary materials. In this article, we will use this algorithm to estimate the conditional distribution functions using the interval-valued data and then the ADD statistic. It can be easily implemented by the R function *EMICM* in the R package *Icens*.[2]

In the sample level, let $\widetilde{\mathcal{A}} = \{\widetilde{y}_1^*, \widetilde{y}_2^*, \ldots, \widetilde{y}_M^*\}$ be the union set of all values for the upper and lower bounds in the observed interval-valued data such as $\widetilde{y}_1^* < \widetilde{y}_2^* < \cdots < \widetilde{y}_M^*$. Then, the ADD statistic can be estimated by the plug-in method as

$$\hat{\omega}_k = \sum_{i=1}^{M} |\widehat{F}_{+k}(\widetilde{y}_i^*) - \widehat{F}_{-k}(\widetilde{y}_i^*)| \frac{\Delta \widetilde{y}_i^*}{\widetilde{y}_M^* - \widetilde{y}_1^*}, \tag{3.5}$$

where $\Delta \widetilde{y}_i^* = \widetilde{y}_i^* - \widetilde{y}_{i-1}^*$ for $i = 2, 3, \ldots, M$ and $\Delta \widetilde{y}_1^* = 0$, $\widehat{F}_{+k}(\cdot)$ and $\widehat{F}_{-k}(\cdot)$ are the estimated conditional distribution functions given $X_k = 1$ and $X_k = 0$, respectively, using the EM-ICM algorithm.

3.2. ADD Based Sure Independence Screening

Our goal is to select important skill words as predictors for the response variable in the ultrahigh dimensional data. We use the ADD statistic as the marginal utility to quantify the importance of each predictor for the response and introduce a new sure independence screening approach. Without specifying any regression model, we first define the active predictor subset in the population level by

$$\mathcal{D} = \{k : F(y|\mathbf{x}) \quad \text{functionally depends on} \quad X_k \quad \text{for some } y \in \mathbb{R},\ 1 \le k \le p\},$$

where $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is the vector of all predictors generated in the job advertisements. If $k \notin \mathcal{D}$, then X_k and Y are independent and thus $\omega_k = 0$. Otherwise, $\omega_k > 0$ and X_k are Y are dependent. Hence, ranking ω_k is a useful way to select the active subset. In the sample level, we propose to use $\widehat{\omega}_k$ in (3.5) to rank predictors and select a reduced model $\widehat{\mathcal{D}} = \{k : \widehat{\omega}_k \ge \gamma_n, \text{ for } 1 \le k \le p\}$, where γ_n is a prespecified threshold. In practice, one could choose the reduced model as the top d predictors. That is,

$$\widehat{\mathcal{D}}^* = \{k : \widehat{\omega}_k \text{ is among the top } d \text{ largest of all}\}, \tag{3.6}$$

where the submodel size d is chosen to be less than n, such as $n/\log(n)$ or $n - 1$ suggested by Fan and Lv (2008). We name this approach as the ADD based Sure Independence Screening, ADD-SIS for short.

Next, we establish the sure screening property of the ADD-SIS. This property is indispensable for all sure independence screening methods because it ensures that all the active predictors in $\mathcal{D}$ can be included in the reduced model $\widehat{\mathcal{D}}$ with the probability approaching one for ultrahigh dimensional data. We assume the following conditions.

(A1) There exists positive constants $c > 0$ and $0 \le \tau < 1/2$ such that $\min_{k \in D} \omega_k \ge 2cn^{-\tau}$.

(A2) Assume $\log(p) = o(n^{1-2\tau})$, where τ is defined in Assumption (A1).

Assumption (A1) requires that the minimum true signal can not be too small but can vanish to zero as the sample size goes to the infinity at the rate of $n^{-\tau}$. It is a critical assumption in the sure independence screening literature, such as Condition 3 in Fan and Lv (2008), Condition (C2) in Li, Zhong, and Zhu (2012) and Condition (C1) in Pan et al. (2019), etc. Assumption (A2) allows the dimension p of predictors diverge to the infinity at an exponential rate $O(n^{\alpha})$ where $\alpha < 1 - 2\tau$. The rate is same as Fan and Lv (2008), Pan et al. (2019), etc. The following theorem presents the sure screening property of the ADD-SIS.

[2] *http://www.bioconductor.org/packages/release/bioc/html/Icens.html*

Theorem 3.1. Under Assumptions (A1) and (A2), when n is sufficiently large, there exists a positive constant a,

$$\Pr\left\{\max_{1\le k\le p}|\hat{\omega}_k-\omega_k|\ge cn^{-\tau}\right\}\le p\mathcal{O}\{n^{\tau-1/2}\exp(-an^{1-2\tau})\}. \tag{3.7}$$

Furthermore, when n is sufficiently large and $\gamma_n = cn^{-\tau}$,

$$\Pr\{\mathcal{D}\subseteq\hat{\mathcal{D}}\}\ge 1-s_n\mathcal{O}\{n^{\tau-1/2}\exp(-an^{1-2\tau})\}, \tag{3.8}$$

where s_n is the cardinality of $\mathcal{D}$.

Theorem 3.1 justifies that the feature screening based on the ADD statistic $\hat{\omega}_k$ is able to select the important predictors for the response. The proof is based on the asymptotic normality of the NPMLE of the distribution functions using the interval censored data (Yu et al. 1998). The proof is relegated to the supplementary materials of the article.

Remark. First, although the estimated ADD statistic $\hat{\omega}_k$ in (3.5) and its properties in Theorem 3.1 are specifically introduced for the interval censored data, the similar result is also valid for the complete data where the exact values of the response are observed. In this case, the ECDF is used to estimate the conditional distribution functions in the ADD statistic (3.1). The Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky, Kiefer, and Wolfowitz 1956) of the ECDF can be used to prove the uniform bound of $\hat{\omega}_k$ like (3.7). Second, the same ADD-SIS procedure can be also applied to ultrahigh dimensional binary classification problems like the FAIR in Fan and Fan (2008) and the Kolmogorov Filter (KF) in Mai and Zou (2013). The ADD-SIS is expected to be more robust to outliers and more powerful to detect the dependence between the binary response and the continuous covariates.

4. Numerical Simulation

Before analyzing our real data, we use the Monte Carlo simulations to assess the finite sample performance of the ADD-SIS compared with other existing methods including distance correlation based SIS (DC-SIS) in Li, Zhong, and Zhu (2012), the Kolmogorov Filter (KF) in Mai and Zou (2013) and the mean variance index based SIS (MV-SIS) in Cui, Li, and Zhong (2015).

We first generate the binary predictors. Let $\mathbf{x}^* = (X_1^*, X_2^*, \ldots, X_p^*)^{\mathrm{T}}$ generate from a multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$, where $\sigma_{ij} = \rho^{|i-j|}$, where $\rho = 0.2$ or $\rho = 0.5$. We consider two ways to generate binary predictors. (a) Balanced case, for each $1\le j\le p$, each normal variable X_j^* is dichotomized by its median value and the obtained binary variable $X_j = 0$ if X_j^* is lower than its median, and $X_j = 1$ if else. (b) Imbalanced case, we similarly take the 70th percentile of each predictor as the dichotomization threshold to generate the binary predictors. Denote by $\mathbf{x} = (X_1,\ldots,X_p)^{\mathrm{T}}$ the vector of the generated binary predictors.

Next, we generate the interval-valued response values based on the binary predictors. We first generate the true response value Y^* based three following models with five active predictors indexed by $\{j = 1,2,3,19,20\}$, respectively:

Example 1. Linear Model: $Y^* = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}+\epsilon$, $\epsilon\sim N(0,1)$. Similar to Fan and Lv (2008), we set the coefficients for the true active predictors to be $\beta_j = (-1)^{U_j}(a+|Z|)$ for $j = 1,2,3,19,20$, where $a = 2\log(n)/\sqrt{n}$, $Z\sim N(0,1)$, $U_j = 0$ for $j = 1,2,3$, and $U_j = 1$ for $j = 19,20$.

Example 2. Poisson Regression: Y^* is generated from a Poisson distribution where the intensity parameter λ is followed as $\log(\lambda) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. The vector of true coefficients is set to be $\boldsymbol{\beta} = (1, 0.8, 0.6, \mathbf{0}_{15}, -0.6, -0.8, \mathbf{0}_{p-20})^{\mathrm{T}}$.

Example 3. Cox Proportional Hazard Model: The hazard function for the ith individual is $h_i(y) = h_0(y)\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, where $h_0(t)$ is the baseline hazard. Thus, Y^* is generated from the distribution $F(y|\mathbf{x}_i) = 1-\exp[-H_0(y)\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})]$, for $y > 0$, where $H_0(y) = \int_0^y h_0(u)du$. For the baseline hazard, we choose the Weibull distribution with the shape parameter equal to 2 and the scale parameter equal to 0.15. Set $\boldsymbol{\beta} = (1.2, 1, 0.8, 0_{15}, -1, -1.2, 0_{p-20})^{\mathrm{T}}$.

Then, random disturbances are added to or subtracted from the true response Y^* to generate the interval-valued data (Y^L, Y^R) by setting $Y^L = [Y^* - U]$ and $Y^R = [Y^* + 2V]$, where $U\sim$ Poisson(2), $V\sim\chi^2(2)$ and $[a]$ indicates the integer part of a. Note that $E(2V) = 4 > E(U) = 2$. This unbalanced setting for the interval-valued data mimics the real situation in which employers commonly tend to set higher upper bounds for posted salaries to attract more applications. Taking the integer parts makes the simulated data analogous to the real application where the lower and upper bounds for post salaries are integers. Meanwhile, by taking the integer parts, the cardinality of all possible unique values is finite. Besides, to mimic the case 2 interval censoring, we also include left censored data and right censored data in the simulation. We use "I/R/L" to denote the percentage of the response that is interval censoring, right censoring and left censoring. We consider two censoring cases: Case 1, (I/R/L) = (80%/14%/6%) and Case 2, (I/R/L) = (80%/6%/14%). We remark that the ADD-SIS is able to make use of the interval information based on the NPMLE of interval censored data. However, the other screening methods can not be directly applied for the interval-value data. To make them applicable, we choose three different single points to represent the interval $[y_i^L, y_i^U]$ to accommodate the unknown asymmetry of the interval, that is, the lower one-third point $y_i^L+(y_i^U-y_i^L)/3$ (denoted by M1), the middle point $(y_i^U+y_i^L)/2$ (denoted by M2) and the upper one-third point $y_i^L+2(y_i^U-y_i^L)/3$ (denoted by M3). We directly take the lower bounds for right censored data and the upper bounds for left censored data as the alternative response values.

We set the sample size n to be 200 and the dimension of predictors p to be 1000. Each simulation is replicated 200 times and the evaluation criteria include: (a) the minimum model size (MMS) to include all the active predictors and its associated robust standard deviation (RSD=IQR/1.34), (b) the proportion that the kth active predictor is selected in the selected model of size $[n/\log(n)]$, denoted by $\mathcal{P}_k$, $k = 1,2,3,19,20$, (c) The proportion that all active predictors are selected in the selected model of size $[n/\log(n)]$, $[2n/\log(n)]$ or $[3n/\log(n)]$, represented by $\mathcal{P}_{a1}$, $\mathcal{P}_{a2}$, or $\mathcal{P}_{a3}$, respectively. Table 1 reports the

Table 1. Median minimum model sizes (MMS) with the associated robust standard deviation in the parentheses based on 200 simulations, the average with the standard deviation in parentheses of the implementation time per simulation for Examples 1–3.

			(I/R/L) = Case 1		(I/R/L) = Case 2		
Example	ρ	Method	Balanced	Unbalanced	Balanced	Unbalanced	Time (s)
1	0.2	**ADD-SIS**	42.5(114.7)	51.5(137.7)	42.0(102.3)	56.0(142.3)	4.07(0.22)
		DC-SIS (M1)	44.5(92.8)	62.0(118.6)	40.0(98.0)	58.5(129.3)	5.43(0.09)
		KF (M1)	73.8(136.7)	88.5(162.2)	65.0(142.1)	82.5(164.0)	0.45(0.02)
		MV-SIS (M1)	46.0(103.2)	67.0(137.5)	42.5(104.3)	64.5(113.0)	1.71(0.07)
		DC-SIS (M2)	24.0(54.5)	34.0(71.7)	33.0(72.6)	39.0(104.5)	5.42(0.09)
		KF (M2)	43.0(96.5)	47.5(101.6)	50.8(121.3)	59.5(116.4)	0.47(0.02)
		MV-SIS (M2)	25.5(61.7)	35.0(71.0)	32.0(77.3)	39.0(105.1)	1.70(0.07)
		DC-SIS (M3)	84.0(162.9)	121.5(184.2)	59.0(139.9)	98.0(152.3)	5.38(0.08)
		KF (M3)	118.5(169.5)	155.0(232.0)	90.5(197.2)	140.0(203.7)	0.46(0.02)
		MV-SIS (M3)	85.5(152.7)	127.5(180.9)	63.0(163.8)	95.5(149.2)	1.69(0.07)
	0.5	**ADD-SIS**	6.0(9.6)	7.0(17.8)	6.0(8.9)	8.0(18.5)	4.14(0.19)
		DC-SIS (M1)	7.0(13.4)	10.5(24.8)	9.0(15.6)	12.5(31.8)	5.43(0.10)
		KF (M1)	12.0(31.3)	16.5(54.9)	13.3(32.2)	19.0(43.4)	0.45(0.02)
		MV-SIS (M1)	8.0(15.6)	10.0(26.1)	8.5(18.7)	13.0(30.4)	1.70(0.08)
		DC-SIS (M2)	6.0(5.2)	7.0(10.4)	8.0(9.6)	8.5(18.7)	5.41(0.08)
		KF (M2)	9.0(15.6)	9.0(28.7)	10.5(23.7)	15.0(33.5)	0.46(0.02)
		MV-SIS (M2)	6.0(6.7)	7.0(13.5)	8.0(11.1)	8.0(20.8)	1.68(0.07)
		DC-SIS (M3)	11.0(27.8)	22.0(57.8)	11.0(32.6)	19.0(52.3)	5.37(0.08)
		KF (M3)	19.5(53.2)	37.5(81.9)	19.0(44.8)	31.0(82.3)	0.46(0.02)
		MV-SIS (M3)	12.0(29.8)	22.0(64.1)	12.0(33.5)	20.0(51.9)	1.68(0.07)
2	0.2	**ADD-SIS**	11.0(29.1)	12.0(15.8)	15.5(30.6)	15.0(22.6)	4.34(0.31)
		DC-SIS (M1)	40.5(78.0)	41.0(73.8)	53.0(124.0)	46.0(95.1)	5.59(0.15)
		KF (M1)	80.3(132.9)	74.5(138.3)	98.3(156.5)	76.0(148.3)	0.45(0.03)
		MV-SIS (M1)	50.0(97.3)	49.5(83.0)	71.0(145.1)	50.5(84.5)	1.71(0.06)
		DC-SIS (M2)	21.5(29.5)	17.0(25.6)	36.0(84.5)	27.0(62.1)	5.54(0.13)
		KF (M2)	50.5(59.5)	35.8(53.6)	69.0(108.9)	50.0(74.9)	0.46(0.02)
		MV-SIS (M2)	26.5(45.2)	22.0(38.5)	46.5(94.5)	32.0(61.3)	1.68(0.07)
		DC-SIS (M3)	82.0(131.4)	97.0(148.4)	84.5(164.2)	87.5(153.3)	5.49(0.11)
		KF (M3)	146.3(170.4)	144.5(167.2)	160.0(188.3)	127.0(182.0)	0.46(0.02)
		MV-SIS (M3)	94.5(137.9)	111.5(166.2)	103.0(195.0)	85.5(143.4)	1.68(0.07)
	0.5	**ADD-SIS**	5.0(1.5)	5.0(1.5)	5.0(2.2)	6.0(2.2)	4.46(0.29)
		DC-SIS (M1)	8.0(12.0)	8.0(12.6)	9.5(16.3)	9.0(16.3)	5.58(0.10)
		KF (M1)	21.0(40.1)	18.0(33.9)	24.0(52.4)	22.0(44.5)	0.45(0.02)
		MV-SIS (M1)	12.0(29.1)	11.0(15.8)	13.0(28.7)	10.5(18.5)	1.71(0.06)
		DC-SIS (M2)	6.0(3.7)	6.0(3.2)	8.0(8.3)	7.5(8.2)	5.53(0.08)
		KF (M2)	11.5(20.2)	10.0(13.5)	15.5(33.4)	13.0(23.0)	0.46(0.02)
		MV-SIS (M2)	8.0(11.1)	7.0(5.9)	9.0(17.2)	8.0(8.9)	1.68(0.07)
		DC-SIS (M3)	15.0(38.0)	17.0(26.9)	15.5(32.1)	14.0(27.4)	5.49(0.07)
		KF (M3)	37.5(73.2)	40.0(56.0)	38.5(74.3)	33.5(63.9)	0.46(0.01)
		MV-SIS (M3)	24.0(58.0)	21.0(31.3)	23.0(43.2)	16.5(32.6)	1.68(0.06)
3	0.2	**ADD-SIS**	20.5(52.4)	45.5(83.8)	24.5(50.8)	42.0(83.2)	4.40(0.39)
		DC-SIS (M1)	49.0(89.1)	92.0(153.4)	46.5(97.3)	105.5(188.7)	5.51(0.07)
		KF (M1)	84.5(157.3)	149.5(204.9)	88.0(139.9)	162.3(217.3)	0.44(0.02)
		MV-SIS (M1)	65.0(116.8)	97.0(153.3)	66.0(115.8)	119.0(181.4)	1.69(0.07)
		DC-SIS (M2)	25.0(33.0)	47.5(111.6)	40.0(86.4)	76.5(140.1)	5.49(0.07)
		KF (M2)	56.5(70.8)	89.0(137.2)	69.3(111.2)	121.0(178.3)	0.46(0.10)
		MV-SIS (M2)	34.5(49.1)	54.5(125.8)	45.5(81.2)	76.0(134.9)	1.69(0.06)
		DC-SIS (M3)	109.5(169.9)	158.5(231.5)	86.5(128.4)	156.5(220.0)	5.49(0.08)
		KF (M3)	164.5(183.1)	219.0(261.3)	149.8(156.5)	219.0(231.1)	0.47(0.03)
		MV-SIS (M3)	136.5(160.9)	173.0(229.2)	96.5(144.9)	165.5(222.9)	1.69(0.06)
	0.5	**ADD-SIS**	6.0(3.9)	11.0(15.6)	6.0(5.9)	11.0(14.1)	4.55(0.39)
		DC-SIS (M1)	12.5(24.6)	20.0(37.1)	18.0(26.9)	26.0(48.6)	5.52(0.08)
		KF (M1)	30.8(61.9)	45.8(71.3)	34.5(64.2)	55.0(81.2)	0.45(0.02)
		MV-SIS (M1)	19.5(34.8)	23.5(47.6)	20.5(34.3)	30.0(46.3)	1.68(0.06)
		DC-SIS (M2)	8.0(8.2)	11.0(15.9)	12.0(16.5)	17.0(37.1)	5.50(0.07)
		KF (M2)	17.0(29.9)	21.0(39.5)	24.5(33.9)	30.5(48.9)	0.46(0.03)
		MV-SIS (M2)	10.0(15.6)	12.0(20.0)	13.0(20.0)	17.5(29.1)	1.68(0.06)
		DC-SIS (M3)	26.0(53.4)	46.0(68.4)	28.0(49.1)	51.5(69.1)	5.50(0.08)
		KF (M3)	64.5(105.6)	77.5(128.4)	49.5(94.5)	89.0(101.2)	0.46(0.01)
		MV-SIS (M3)	41.5(71.3)	53.0(87.1)	34.0(58.3)	51.5(73.6)	1.68(0.05)

median of minimum model sizes (MMS) with the associated robust standard deviations in the parentheses out of 200 simulations, and the average with the standard deviation in parentheses of the implementation time per simulation. Tables S.8–S.10 in the supplementary materials of this article summarize the sample proportions of single active predictors and all active predictors out of 200 simulations. According to simulation results, we conclude that the ADD-SIS requires smaller minimum model sizes to include all active predictors in most scenarios. The reason could be that the ADD-SIS makes use of the interval

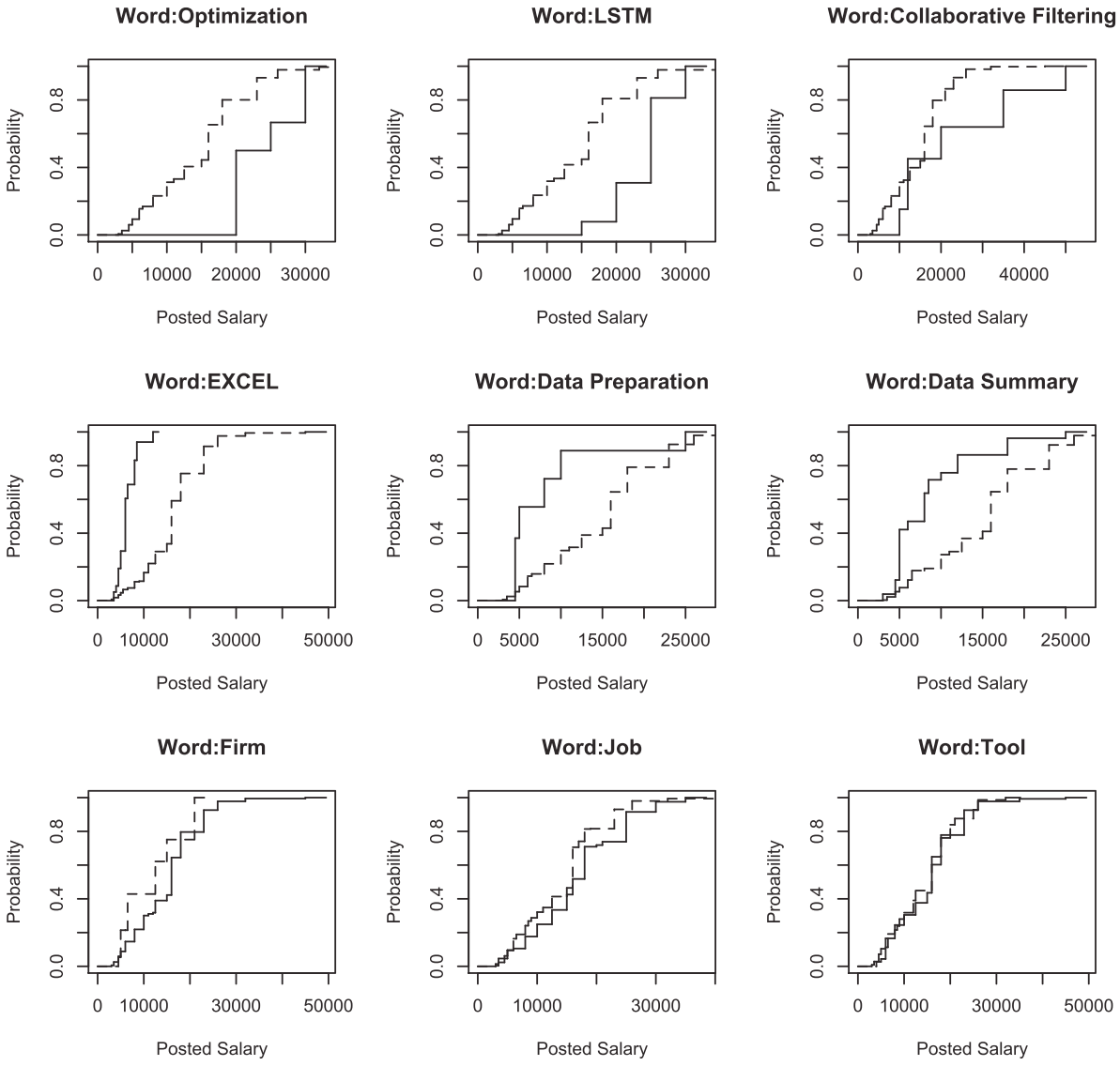

Figure 3. Estimated conditional distribution functions of the posted salary with (solid curves) and without (dashed curves) the different words.

information in the data while the single-point based methods ignore the interval information and may loss some powers to detect the relationship between the response and predictors.

5. An Application

In this section, we conduct an empirical analysis of the online job advertisements data for data scientists and data analysts in a major China's online job posting website. We first use the ADD-SIS to select important words generated by word segmentation and then explore the regression relationship between the posted salary and the important selected words using an accelerated failure time (AFT) model. Note that the proposed method can be also applied to any other job category. In the supplementary materials, we also consider another real data analysis on job postings for web developers.

5.1. Feature Screening

After generating the potential skill words as predictors by word segmentation and dropping out all the stop words, we have collected $p = 1043$ predictors in our data. However, some words have very low frequencies in the job samples of size $n = 497$. Thus, we filter out the low-frequency words whose frequencies are less than 5 and obtain the remaining predictors of the dimension $p = 677$ in our analysis. Then, we applied the ADD-SIS to rank all predictors and select the important predictors.

To gain a first insight into the empirical results, we first plot the estimated conditional distribution functions of the posted salary using the EM-ICM algorithm with or without a given word in Figure 3. We select nine words including three words (Optimization, LSTM, Collaborative Filtering) which are ranked in the top by the ADD-SIS and positively contribute to the salary, three words (Excel, Data Preparation, Data Summary) which are also ranked in the top by the ADD-SIS but negatively contribute to the salary and three general words (Firm, Job, Tool) which are not ranked in the top and seem uncorrelated with the salary. In each plot, the solid curve and the dashed curve represent the estimated conditional distribution functions of the posted salary with and without the given word, respectively. The plots provide an intuitive insight about the importance of each word for the salary. For example, in the left

Table 2. Top words selected by the ADD-SIS which are positively or negatively correlated to the posted salary.

Top words that are *positively* correlated to the posted salary			
Graduate degree	Parallel computation	MapReduce	Data structure
Audio algorithms	Collaborative filtering	ICCV	Semantic segmentation
Object detection	CVPR	Massive Data	Risk management
Optimization	System configuration	OCR	Research achievements
Image classification	POS tagging	LSTM	Self-learning ability
Software engineering	Semantic analysis	Embedding	Deep learning
Automatic translation	CNN	Sentiment analysis	RNN
Top words that are *negatively* correlated to the posted salary			
Excel	Office softwares	Data collection	College degree
Report writing	Data summary	Data preparation	VBA
Data management	ERP system	PPT	Microsoft access

plot of the first row, we can see that the estimated median of the salary for the jobs which require Optimization (the solid curve) is about 20–25k yuan/month while the estimated median of the salary for the job postings without Optimization (the dashed curve) is about 15k yuan/month. The estimated ADD statistic is useful to measure the area of the distribution discrepancy between two estimated conditional distribution functions of the posted salary with and without the given word.

It is noteworthy that $\hat{\omega}$ in (3.5) itself can't determine the direction of the association between the posted salary and the selected words. As shown in Figure 3, the estimated conditional distribution functions of the posted salary with and without the selected words helps us determine the direction of the association. We list the top words selected by the ADD-SIS which are positively or negatively correlated to the posted salary in Table 2. Generally speaking, the jobs with higher salaries require the higher degrees with modern machine learning and statistical modeling skills. These top words which are positively correlated with the salary can be categorized into several topic groups. (a) Algorithm and computation related words such as MapReduce, parallel computation, optimization, LSTM (Long Short-Term Memory), etc; (b) Natural language processing (NLP)/text mining related words such as POS (part-of-speech) tagging, semantic segmentation, sentiment analysis, automatic translation etc; (c) Computer vision/image analysis related words such as object detection, CNN, ICCV, image classification, etc; (d) Degree and personality related words such as graduate degree, self-learning ability, etc. On the other hand, some basic techniques of data analysis such as Excel, PPT, VBA, Microsoft Access, data summary are only needed in the jobs with lower posted salaries. This finding could provide some guidance for people who are seeking data scientist jobs to design a study plan to meet the necessary job requirements. It also suggests that programs in applied statistics can tailor the curriculums via adding some courses on deep learning algorithms, NLP/text mining, computational algorithms for massive data and so on, which could help students to meet the job requirements of data scientists.

In addition, for the purpose of comparison, we also consider other feature screening methods including DC-SIS (Li, Zhong, and Zhu 2012), KF (Mai and Zou 2013) and MV-SIS (Cui, Li, and Zhong 2015). To make them applicable for the interval-valued data, we use the single point scheme for these screening methods like simulations. Then, we conduct the comparison based on the regression analysis in the next section.

5.2. Regression Analysis

Next, we use the parametric interval censoring regression models to explore the joint relationship between the posted salary and the selected predictors. Odell, Anderson, and D'Agostino (1992) studied the Weibull distribution based accelerated failure time model for interval censored regression with maximum likelihood estimation model. The parametric interval regression model is appealing because of its simplicity, computational efficiency and interpretability. We consider an accelerated failure time (AFT) model to establish the relationship between the posted salary and the selected predictors, denoted by $\mathbf{x}_{\widehat{\mathcal{D}}}$,

$$\log Y = \mu + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{\widehat{\mathcal{D}}} + \varepsilon, \tag{5.1}$$

where ε is the error term. To avoid potential model misspecification, we consider three different probability distributions for ε: normal, logistic and Weibull. The maximum likelihood estimators can be obtained using the R function *survreg* in the R package *survival*[3] in which we specify the data type as *interval2*. To compare the adequacy of model fitting based on different subsets of selected predictors, we consider two goodness-of-fit summary statistics: the log-likelihood statistic and the Chi-squared test statistic. The latter one is based on the likelihood ratio statistic to compare the discrepancy between the model of interest and the null model. Table 3 summaries these two goodness-of-fit statistics for different models. Two goodness-of-fit statistics of the submodel based on the ADD-SIS are the largest of all. It demonstrates that the predictors selected by the ADD-SIS with the consideration of the interval information are more relevant to interpret the interval-valued response.

Next, we evaluate the predictability of the accelerated failure time (AFT) model with the selected predictors for the posted salary. We consider the leave-one-out cross-validation (LOOCV) prediction error as the evaluation criterion to assess the different models. Specifically, for each $i = 1, \ldots, n$, we train an AFT regression model using the $n-1$ observations by deleting the ith observation and then obtain the ith predicted response value, denoted by $\hat{y}_i^{\star}$, based on the trained AFT model. Then, we compute the empirical cumulative distribution function (ECDF) of predicted response values $\{\hat{y}_1^{\star}, \hat{y}_2^{\star}, \ldots, \hat{y}_n^{\star}\}$, denoted by $\widehat{F}_{\text{prediction}}(\cdot)$. On the other hand, we can estimate the true distribution function of the interval-valued response using the EM-ICM algorithm, denoted by $\widehat{F}_{\text{interval}}(\cdot)$. To gain

[3] *https://cran.r-project.org/web/packages/survival/index.html*

Table 3. Goodness-of-fit summary statistics of the accelerated failure time model based on the selected predictors of size $d = 60$ by different screening methods.

	Log-likelihood statistic			Chi-squared test statistic			Time(sec)
Error distribution	Normal	Logistic	Weibull	Normal	Logistic	Weibull	
ADD-SIS	−464.9	−465.5	−464.0	682.2	693.9	640.8	14.2
KF(M1)	−500.7	−499.0	−507.7	610.7	626.9	553.3	0.4
DC-SIS (M1)	−490.8	−493.8	−494.3	630.4	637.3	580.2	34.8
MV-SIS (M1)	−493.3	−495.0	−494.8	625.3	634.9	579.1	4.4
KF (M2)	−497.2	−495.5	−503.2	617.5	634.0	562.4	0.3
DC-SIS (M2)	−490.9	−494.0	−494.3	630.3	636.9	580.2	32.0
MV-SIS (M2)	−493.3	−494.8	−495.0	625.5	635.3	578.8	4.7
KF (M3)	−497.2	−495.5	−503.2	617.5	633.9	562.3	0.3
DC-SIS (M3)	−493.4	−494.9	−494.9	625.3	635.0	579.1	33.0
MV-SIS (M3)	−493.4	−494.9	−494.9	625.3	635.0	579.1	4.3

NOTE: The last column reports the implementation time of each feature screening method.

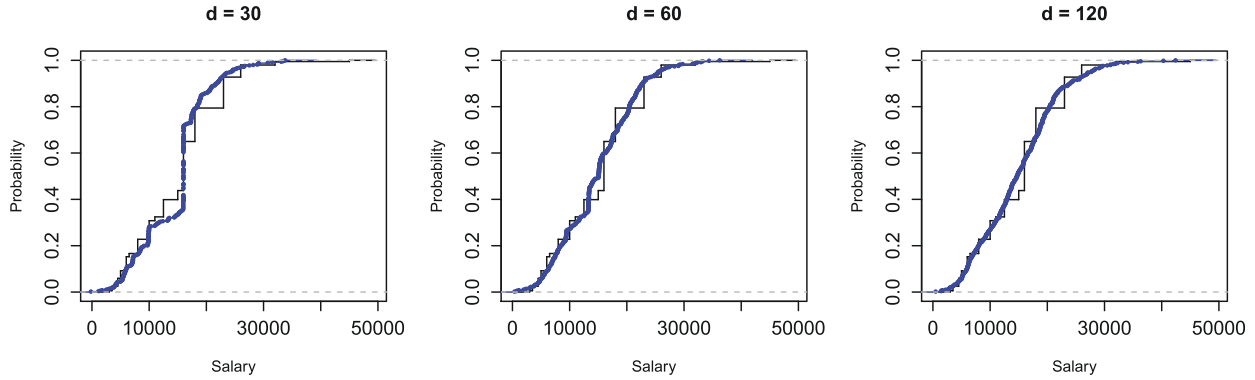

Figure 4. Prediction performance of the accelerated failure time (AFT) model with normal errors using the predictor subsets of size $d = 30, 60, 120$ selected by the ADD-SIS. The blue thick dotted lines denote $\widehat{F}_{\text{prediction}}(\cdot)$, the empirical cumulative distribution functions (ECDF) of predicted response values and the black thin solid lines denote $\widehat{F}_{\text{interval}}(\cdot)$, the estimated distribution function of the interval-valued response using the EM-ICM algorithm.

an intuition about their comparison, we plot $\widehat{F}_{\text{prediction}}(\cdot)$ (blue thick dotted lines) and $\widehat{F}_{\text{interval}}(\cdot)$ (black thin solid lines) in Figure 4 based on the AFT model with normal errors using the predictor subsets of size $d = 30, 60, 120$ selected by the ADD-SIS. We can observe that when the selected model size $d = 60$ or 120, the curves of $\widehat{F}_{\text{prediction}}(\cdot)$ based on the predicted values are quit close to the step functions of $\widehat{F}_{\text{interval}}(\cdot)$. It illustrates that the AFT model can be used to predict the posted salaries based on the important skill words selected by the ADD-SIS.

Then, we consider two two-sample statistics to measure the distribution discrepancy between $\widehat{F}_{\text{prediction}}(\cdot)$ and $\widehat{F}_{\text{interval}}(\cdot)$: the two-sample Kolmogorov-Smirnov (KS_{pred}) statistic and the two-sample Cramer-von Mises (CVM_{pred}) statistic. Define the pooled set $\mathcal{S} = \{\hat{y}_1^\star, \hat{y}_2^\star, \ldots, \hat{y}_n^\star\} \bigcup \widetilde{\mathcal{A}}$, where $\widetilde{\mathcal{A}} = \{\widetilde{y}_1^*, \widetilde{y}_2^*, \ldots, \widetilde{y}_M^*\}$ is the union set of all values for the upper and lower bounds in the observed interval-valued responses. We let $\mathcal{S} = \{z_1, z_2, \ldots, z_{n+M}\}$ such that $z_1 < z_2 < \cdots < z_{n+M}$. Two two-sample statistics are defined as, respectively,

$$\text{KS}_{\text{pred}} = \sup_{z_i \in \mathcal{S}} \left| \widehat{F}_{\text{prediction}}(z_i) - \widehat{F}_{\text{interval}}(z_i) \right|,$$

$$\text{CVM}_{\text{pred}} = \frac{nM}{(n+M)^2} \sum_{i=1}^{n+M} \left[\widehat{F}_{\text{prediction}}(z_i) - \widehat{F}_{\text{interval}}(z_i) \right]^2.$$

Table 4 summarizes the two two-sample statistics for the AFT model with normal errors using the predictor subsets of size $d = 30, 60, 120$ selected by the ADD-SIS. The AFT models with less top predictors (e.g., $d = 30$) might be not enough to fit the data, which leads to poor predictability. On the other hand, if too many predictors are included in the regression model (e.g., $d = 120$), some irrelevant predictors may be contained and the prediction power is also discounted. Among three different choice of the model size, the AFT model of a moderate size ($d = 60$) gives the best prediction performance.

Table 4. The distribution discrepancy between $\widehat{F}_{\text{prediction}}(\cdot)$ and $\widehat{F}_{\text{interval}}(\cdot)$ based on the AFT model with normal errors using the predictor subsets of size $d = 30, 60, 120$ selected by the ADD-SIS.

	KS_{pred} Statistic	CVM_{pred} Statistic
$d = 30$	0.272	0.292
$d = 60$	0.108	0.059
$d = 120$	0.128	0.073

Furthermore, we also compare the prediction powers of the AFT regression model based on the predictor subsets selected by different feature screening methods. Table 5 summarizes the above two-sample statistics to measure the distribution discrepancy between $\widehat{F}_{\text{prediction}}(\cdot)$ and $\widehat{F}_{\text{interval}}(\cdot)$ based on the AFT model using the submodel size $d = 60$ selected by different screening methods. We can see that the regression models selected by the ADD-SIS can provide the better prediction powers than the other three screening methods. In addition, we generate the 95% confidence bound by setting the confidence bounds to be [2.5%, 97.5%] quantiles for each of the predictions using the R package *survival*. Similar to the F_1 score as an accuracy measure of prediction in a binary classification procedure, we define the following criterion to evaluate the prediction performance of the predicted confidence interval compared with the original salary interval in the job posts, denoted by F_1^{Interval},

$$F_1^{\text{Interval}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

Table 5. The distribution discrepancy between $\widehat{F}_{\text{prediction}}(\cdot)$ and $\widehat{F}_{\text{interval}}(\cdot)$ as well as the defined F_1^{Interval} score to evaluate the performance of the predicted interval based on the AFT model using the selected model size $d = 60$.

Error	Method	KS_{pred}	CVM_{pred}	Recall	Precision	F_1^{Interval}
Normal	**ADD-SIS**	**0.108**	**0.059**	**0.809**	**0.455**	**0.582**
	KF (M1)	0.168	0.125	0.784	0.416	0.544
	DC-SIS (M1)	0.114	0.081	0.775	0.411	0.537
	MV-SIS (M1)	0.114	0.081	0.775	0.412	0.537
	KF (M2)	0.162	0.123	0.786	0.417	0.545
	DC-SIS (M2)	0.114	0.081	0.775	0.411	0.537
	MV-SIS (M2)	0.118	0.080	0.774	0.410	0.536
	KF (M3)	0.158	0.124	0.786	0.418	0.546
	DC-SIS (M3)	0.120	0.083	0.770	0.412	0.537
	MV-SIS (M3)	0.118	0.080	0.774	0.410	0.536
Logistic	**ADD-SIS**	**0.104**	**0.063**	**0.825**	**0.443**	**0.577**
	KF (M1)	0.123	0.091	0.787	0.411	0.540
	DC-SIS (M1)	0.119	0.076	0.783	0.408	0.537
	MV-SIS (M1)	0.119	0.076	0.783	0.408	0.537
	KF (M2)	0.122	0.088	0.788	0.412	0.541
	DC-SIS (M2)	0.119	0.076	0.783	0.408	0.537
	MV-SIS (M2)	0.119	0.074	0.782	0.407	0.536
	KF (M3)	0.116	0.087	0.788	0.413	0.542
	DC-SIS (M3)	0.117	0.077	0.782	0.409	0.537
	MV-SIS (M3)	0.119	0.074	0.782	0.407	0.535
Weibull	**ADD-SIS**	**0.154**	**0.162**	0.796	**0.431**	**0.560**
	KF (M1)	0.263	0.423	0.819	0.398	0.535
	DC-SIS (M1)	0.219	0.306	0.807	0.401	0.536
	MV-SIS (M1)	0.219	0.304	0.807	0.401	0.536
	KF (M2)	0.272	0.430	**0.821**	0.398	0.536
	DC-SIS (M2)	0.219	0.306	0.807	0.401	0.536
	MV-SIS (M2)	0.219	0.304	0.805	0.400	0.535
	KF (M3)	0.274	0.444	0.821	0.399	0.537
	DC-SIS (M3)	0.213	0.294	0.801	0.403	0.536
	MV-SIS (M3)	0.219	0.304	0.805	0.400	0.535

NOTE: Numbers in boldface mean the corresponding procedure is the best under the criterion.

where

$$\text{Precision} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|[Y_L^i, Y_R^i] \cap [\widehat{Y_L^i}, \widehat{Y_R^i}]\right|}{\left|[\widehat{Y_L^i}, \widehat{Y_R^i}]\right|},$$

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|[Y_L^i, Y_R^i] \cap [\widehat{Y_L^i}, \widehat{Y_R^i}]\right|}{\left|[Y_L^i, Y_R^i]\right|},$$

$[Y_L^i, Y_R^i]$ is the true interval of the response for the ith observation, $[\widehat{Y_L^i}, \widehat{Y_R^i}]$ is the corresponding [2.5%, 97.5%] confidence bound for predictions and $|I|$ denotes the length of the interval I. Note that there are several cases where the true salary interval is left censoring or right censoring. In this case, we choose 0 as the lower bound for the left censored data and the maximum salary, 80k yuan/month, as the upper bound of the right censored data, respectively. The higher the value of F_1^{Interval} is, the better prediction accuracy this model has. We also summarize the values of F_1^{Interval} in Table 5. Our method ADD-SIS outperforms all other benchmark approaches in terms of the F_1^{Interval} score as an accuracy measure of predictability for the prediction confidence bound.

6. Conclusion and Discussion

To explore the important skills of the job requirements, we study the relationship between the posted salary and the job requirements in online job markets. We have handled with two challenges of this problem. First, we treated the interval-valued form of the salary data as the case 2 interval censoring and used the EM-ICM algorithm to estimate the distribution function of the salary. We proposed a new dependence measure, Absolute Distribution Difference (ADD), between an interval-valued response and each binary predictor. Second, we introduced a new ADD based sure independence screening (ADD-SIS) to select important skill words from ultrahigh dimensional potential words generated by word segmentation from the job requirements. We further apply the newly proposed procedures for an empirical analysis of job advertisements on data scientists and data analysts. Our empirical analysis implies that words or phrases such as "graduate degree," "parallel computation," "MapReduce," "Optimization," "LSTM," "CNN" listed in top panel of Table 2 are positively associated with the posted salary of data scientists or data analysts, while words or phrases such as "Excel," "Office softwares," "data collection" listed in the bottom panel of Table 2 are negatively correlated to the posted salary of data scientists or data analysts. In practice, one can use the proposed ADD-SIS method to explore important skill words for different types of jobs. The newly proposed procedures are applicable for other types of job advertisements. They are also broadly applicable for other research areas where people collect interval-valued responses along with ultrahigh dimensional predictors. We remark that an accurate salary prediction model can benefit both employees and employers in the job markets. For employees, especially new graduates and people switching jobs, knowing the expected salary based on their experienced job skills can help them learn the current market value to find satisfying employment. It can help employers set an appropriate and competitive salary for new postings and intelligently optimize hiring plans. It also benefits the job posting platform, like LinkedIn, which can use the results to improve the better matching between employees and employers. As mentioned by 2021 Nobel Prize winer David Card in his recent work (Card 2022), accurate wage setting based on associated skills can lead to better evaluation on minimum wage policies and antitrust regulations. In conclusion, this return-to-skills association model could be able to reduce the information bias and increase the matching efficiency.

Besides feature selection, there is another way to deal with high dimensionality or low frequency of rare words in text mining via aggregating similar words together to achieve dimension reduction and generate low-dimensional predictors. For instance, Li et al. (2021) grouped similar words into several topics with the nonnegative matrix factorization. Yan and Bien (2021) focused on aggregating rare words with low frequency to have more reliable features. To use the information of these words with very low frequencies, we could follow the idea of Yan and Bien (2021) to aggregate these low-frequency words into denser features and then apply the proposed ADD-SIS to the new aggregated features and words with enough frequencies. It might avoid the risk of losing important information of rare words.

Some extensions can be considered in the future studies. The ADD-SIS approach can be extended to deal with multi-categorial and continuous predictors using the slicing-and-fusion idea of the Fused Kolmogorov filter proposed in Mai and Zou (2015). Specially, when the predictor is multi-categorial, we can compute the ADD statistic for each pair of categories

and then take the supreme of all pairwise ADD statistics. For the continuous predictor, we can first slice it into multiple slices using quantiles and then compute the supreme of the ADD statistics over all pairs of slices. In addition, to make it insensitive to the slicing scheme, we can take the sum of the supreme values of the ADD statistics over different ways of slicing as the marginal utility to measure the importance of the continuous predictor for interval-valued response. The similar slicing method has also been used in Yan et al. (2018) for a fused mean-variance filter. In this article, we focus on feature screening for interval-valued data with ultrahigh dimensional predictors. In practice, people may encounter interval-valued data with large dimensional predictors. Thus, it is of interest to develop regularization methods for interval-valued data. This would be a good topic for future research.

Supplementary Materials

The EM-ICM algorithm, the proof of Theorem 3.1 and additional numerical results are included in the online supplementary materials.

Acknowledgments

We thank the Editor, the Associate Editor, and two referees for their insightful comments which have substantially improved the article.

Data Availability Statement

All numerical studies were conducted by using R code. The data and R code are available at webpage: *https://github.com/tsienchen/ADD-SIS*.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Zhong's research was supported by the National Natural Science Foundation of China (NSFC) grants (11922117, 12231011, 71988101), National Key R&D Program of China 2022YFA10038002 and National Statistical Science Research Program of China (2022LD08). Zhu's research was supported by NSFC (12225113 and 12171477) and Renmin University of China (22XNA026). Li's research was supported by National Science Foundation (NSF) DMS-1820702, and NIH grants R01AI136664 and R01AI170249. The content is solely the responsibility of the authors and does not necessarily represent the official views of NSFC, NSF, or NIH.

ORCID

Runze Li http://orcid.org/0000-0002-0154-2202

References

Aragón, J., and Eberly, D. (1992), "On Convergence of Convex Minorant Algorithms for Distribution Estimation with Interval-Censored Data," *Journal of Computational and Graphical Statistics*, 1, 129–140. [806,807]

Beaudry, P., Green, D. A., and Sand, B. M. (2016), "The Great Reversal in the Demand for Skill and Cognitive Tasks," *Journal of Labor Economics*, 34, S199–S247. [805]

Card, D. (2022), "Who Set Your Wage?" Technical report, National Bureau of Economic Research. [815]

Card, D., and DiNardo, J. E. (2002), "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles," *Journal of Labor Economics*, 20, 733–783. [805]

Cui, H., Li, R., and Zhong, W. (2015), "Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis," *Journal of the American Statistical Association*, 110, 630–641. [806,810,813]

Deming, D., and Kahn, L. B. (2018), "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals," *Journal of Labor Economics*, 36, S337–S369. [806]

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956), "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator," *The Annals of Mathematical Statistics*, 27, 642–669. [810]

Fan, J., and Fan, Y. (2008), "High-Dimensional Classification using Features Annealed Independence Rules," *The Annals of Statistics*, 36, 2605–2637. [810]

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557. [806]

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. [806,809,810]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with np-dimensionality," *The Annals of Statistics*, 38, 3567–3604. [806]

Finkelstein, D. M., and Wolfe, R. A. (1985), "A Semiparametric Model for Regression Analysis of Interval-Censored Failure Time Data," *Biometrics*, 41, 933–945. [808]

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H., and Rahwan, I. (2019), "Toward Understanding the Impact of Artificial Intelligence on Labor," *Proceedings of the National Academy of Sciences*, 116, 6531–6539. [805]

Geskus, R., and Groeneboom, P. (1999), "Asymptotically Optimal Estimation of Smooth Functionals for Interval Censoring, Case 2," *The Annals of Statistics*, 27, 627–674. [806,807]

Groeneboom, P., and Wellner, J. A. (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation* (Vol. 19), Basel: Springer. [809]

Gu, M. G., and Zhang, C. H. (1993), "Asymptotic Properties of Self-Consistent Estimators Based on Doubly Censored Data," *The Annals of Statistics*, 21, 611–624. [809]

Hershbein, B., and Kahn, L. B. (2018), "Do Recessions Accelerate Routine-biased Technological Change? Evidence from Vacancy Postings," *American Economic Review*, 108, 1737–1772. [805,806]

Kuhn, P., and Shen, K. (2013), "Gender Discrimination in Job Ads: Evidence from China," *The Quarterly Journal of Economics*, 128, 287–336. [805]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [806,809,810,813]

Li, Y., Zhu, R., Qu, A., Ye, H., and Sun, Z. (2021), "Topic Modeling on Triage Notes with Semiorthogonal Nonnegative Matrix Factorization," *Journal of the American Statistical Association*, 116, 1609–1624. [815]

Mai, Q., and Zou, H. (2013), "The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification," *Biometrika*, 100, 229–234. [806,808,810,813]

——— (2015), "The Fused Kolmogorov Filter: A Nonparametric Model-Free Screening Method," *The Annals of Statistics*, 43, 1471–1497. [815]

Marinescu, I., and Wolthoff, R. (2020), "Opening the Black Box of the Matching Function: The Power of Words," *Journal of Labor Economics*, 38, 535–568. [805,806]

Modestino, A. S., Shoag, D., and Ballance, J. (2016), "Downskilling: Changes in Employer Skill Requirements over the Business Cycle," *Labour Economics*, 41, 333–347. [806]

Odell, P. M., Anderson, K. M., and D'Agostino, R. B. (1992), "Maximum Likelihood Estimation for Interval-Censored Data using a Weibull-Based Accelerated Failure Time Model," *Biometrics*, 48, 951–959. [813]

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2019), "A Generic Sure Independence Screening Procedure," *Journal of the American Statistical Association*, 114, 928–937. [806,809]

Shao, X., and Zhang, J. (2014), "Martingale Difference Correlation and its Use in High-Dimensional Variable Screening," *Journal of the American Statistical Association*, 109, 1302–1318. [806]

Sun, Y., Han, A., Hong, Y., and Wang, S. (2018), "Threshold Autoregressive Models for Interval-Valued Time Series Data," *Journal of Econometrics*, 206, 414–446. [806]

Turnbull, B. W. (1976), "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data," *Journal of the Royal Statistical Society*, Series B, 38, 290–295. [809]

Wellner, J. A., and Zhan, Y. (1997), "A Hybrid Algorithm for Computation of the Nonparametric Maximum Likelihood Estimator from Censored Data," *Journal of the American Statistical Association*, 92, 945–959. [806,807,808,809]

Yan, X., and Bien, J. (2021), "Rare Feature Selection in High Dimensions," *Journal of the American Statistical Association*, 116, 887–900. [815]

Yan, X., Tang, N., Xie, J., Ding, X., and Wang, Z. (2018), "Fused Mean-Variance Filter for Feature Screening," *Computational Statistics & Data Analysis*, 122, 18–32. [816]

Yu, Q., Schick, A., Li, L., and Wong, G. Y. (1998), "Asymptotic Properties of the GMLE with Case 2 Interval-Censored Data," *Statistics & Probability Letters*, 37, 223–228. [807,808,809,810]