Statistica Sinica **33** (2023), 2327-2358
doi:https://doi.org/10.5705/ss.202020.0498

BOLT-SSI: A STATISTICAL APPROACH TO SCREENING INTERACTION EFFECTS FOR ULTRA-HIGH DIMENSIONAL DATA

Min Zhou, Mingwei Dai, Yuan Yao, Jin Liu, Can Yang and Heng Peng

Beijing Normal University-Hong Kong Baptist University United International College, Southwestern University of Finance and Economics, Victoria University of Wellington, Duke-NUS Medical School, The Hong Kong University of Science and Technology and Hong Kong Baptist University

Abstract: Detecting the interaction effects among the predictors on the response variable is a crucial step in numerous applications. We first propose a simple method for sure screening interactions (SSI). Although its computation complexity is $O(p^2n)$, the SSI method works well for problems of moderate dimensionality (e.g., $p = 10^3 \sim 10^4$), without the heredity assumption. For ultrahigh-dimensional problems (e.g., $p = 10^6$), motivated by a discretization associated Boolean representation and operations and a contingency table for discrete variables, we propose a fast algorithm, called "BOLT-SSI." The statistical theory is established for SSI and BOLT-SSI, guaranteeing their sure screening property. We evaluate the performance of SSI and BOLT-SSI using comprehensive simulations and real case studies. Our numerical results demonstrate that SSI and BOLT-SSI often outperform their competitors in terms of computational efficiency and statistical accuracy. The proposed method can be applied to fully detect interactions with more than 300,000 predictors. Based on our findings, we believe there is a need to rethink the relationship between statistical accuracy and computational efficiency. We have shown that the computational performance of a statistical method can often be greatly improved by exploring the advantages of computational architecture with a tolerable loss of statistical accuracy.

Keywords: Discretization, package "BOLTSSIRR", sure independent screening for interaction detection, trade-off between statistical efficiency and computational complexity, ultra-high dimensionality.

1. Introduction

In the past two decades, numerous innovative algorithms have been proposed to address the computational challenges of statistical inference in high-

Corresponding author: Mingwei Dai, Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan 610074, China. E-mail: daimw@swufe.edu.cn.

dimensional problems. However, there still exists a gap between established statistical theory and the computational performance of these algorithms. On the one hand, many statistical models can deal with high-dimensional problems under some theoretically mild conditions, but their computational cost becomes too expensive to be affordable when the dimensionality becomes extremely large. On the other hand, to address many real problems, some algorithms are not developed in a principled way, leading to computational results without statistical guarantees. As argued by Chandrasekaran and Jordan (2013), there is a great need to rethink the relationship between statistical accuracy and computational efficiency.

To bridge this gap, most studies focus on reducing the theoretical complexity of an algorithm or simply use parallel computing to speed it up, without taking advantage of the computational architecture. In fact, the computational performance of statistical models can often be greatly improved by designing new data structures or using hardware acceleration (e.g., graphical processing units for training deep neural networks). Here, we use the interaction detection problem in high-dimensional models to show that it is possible to design statistically guaranteed algorithms by taking advantage of the computational architecture.

1.1. Interaction effect detection

The Oxford English Dictionary defines the word "interaction" as a reciprocal action or influence of persons or things on each other. It is a relationship between two or more objects that have a mutual influence on one another. There is a long history of investigating interaction effects in different scientific fields (Wang and Chen (2020)). For example, in physical chemistry, the main topics are interactions between atoms and molecules. A simple example in the real world is that of carbon and steel. Neither has much effect on the strength, but the combination of the two has substantial effects. In medicine and pharmacology, the interaction effects of multiple drugs have been widely observed (Lees, Cunningham and Elliott (2004)). In genomics, gene-gene interactions and gene-environment interactions have been widely studied by bio-medical researchers since the seminal work of Bateson (1909). In recent years, there has been increasing interest on detecting gene-gene interactions from genome-wide association studies (GWAS) (Cordell (2009); Wang and Chen (2018)).

Here, we investigate interaction effects from a statistical perspective, where an interaction effect is characterized by a statistical departure from the additive effects of two or more factors (see Fisher (1918); Cox (1984)). In a high-dimensional regression framework, it is common to use products of explanatory

variables to study the interaction effects of the explanatory variables on the response variables. Consider three explanatory variables X_i, X_j, and X_k with two-way interaction terms X_jX_k, X_iX_j, and X_iX_k. After including these interaction terms, the standard linear regression model becomes $Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \sum_{1 \le j < k \le p} \beta_{jk} X_j X_k + \varepsilon$, where Y is the response variable, β_0 is the intercept term, β_i is the coefficient of the main effect term X_i, β_{jk} is the coefficient of the interaction term X_jX_k, and ε is an independent error. In high-dimensional data, the number of variables p can be much larger than the sample size n. Clearly, the number of parameters to be determined is $p + p(p-1)/2$ if all two-way interaction terms are included. For example, a GWAS may include millions of genotyped genetic variants, that is, $p \approx 10^6$. The number of interaction terms is of the order of 10^{12}. In this case, the computational cost of detecting such interaction effects becomes unaffordable, making theoretical guarantees with mild conditions (e.g., sparsity assumptions) useless.

To reduce the computational cost, methods often make two types of heredity assumptions. The strong heredity assumption means that the interaction effect is important only if both parents are significant, and the weak heredity assumption states that the interaction term is important only if at least one of its parents is included in the model. Choi, Li and Zhu (2010) extended the LASSO method to identify the significant interaction terms in a linear model and generalized linear models (GLMs) under the strong heredity assumption. They proved that their method possesses the oracle property (Fan and Li (2001); Fan and Peng (2004)), that is, it performs as though the true model was known in advance. The algorithm hierNet was developed by Bien, Taylor and Tibshirani (2013) to select interactions. They added a set of convex constraints to the LASSO in the linear model and constructed a sparse interaction model using the strong and weak heredity assumptions. For the linear model, Hao and Zhang (2014) proposed two algorithms, iFORT and iFORM, identifying the interaction effects in a greedy fashion under the heredity assumption. Lim and Hastie (2015) introduced the method "glinternet" for learning pairwise interactions in a linear regression or logistic regression model with a strong hierarchy constraint. Hao, Feng and Zhang (2018) improved the interaction detection by proposing a regularization algorithm under the marginality principle (RAMP). The "backtracking" method was developed by Shah (2016). It can be incorporated into many existing high-dimensional methods based on penalty functions, and works by iteratively building increasing sets of candidate interactions. She, Wang and Jiang (2018) proposed a group regularized estimation under a structural hierarchy about variable selection for models that include interactions. They provided the minimax lower bounds for

strong and weak hierarchical variable selection, and showed that the proposed estimators enjoy sharp rate oracle inequalities. Deviating from these heredity assumptions for interaction detection, Fan et al. (2016) (Li et al. (2021)) suggested a flexible sure screening procedure, called the interaction pursuit (IP), in ultrahigh-dimensional linear interaction models. The IP method selects the "active interaction variables" by first screening significant predictor variables using the strong Pearson correlation between X_j^2 and Y^2, and then detecting the interaction effects between the identified active interaction variables. The IP method is a good attempt to detect pure interaction effects in a model. Kong et al. (2017) extended the IP method to the ultrahigh-dimensional linear interaction model with multiple responses by identifying the active interactive variables using the distance correlation with X_j^2 and the multiple response $\boldsymbol{Y}^2$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$ is a q-dimensional vector of responses and $\boldsymbol{Y}^2 = (Y_1^2, \ldots, Y_q^2)$.

However, the heredity assumption may not be satisfied in practice because of the existence of pure interaction effects. In human genetics, many gene-gene interaction effects have been detected in the absence of main effects (Cordell (2009); Wan et al. (2010)). For instance, Ritchie et al. (2001) detected pure epistatic interactions among two or more loci in relatively small samples for common complex multifactorial human diseases. They proposed the method MDR to identify interactions, and applied it to a real-data example (sporadic breast cancer case-control data set) to demonstrate the existence of pure interactions. Culverhouse et al. (2002) discussed interaction models without main effects, and examined pure epistatic interactions with loci that did not display any single-locus effects. Cordell (2009) discussed detecting gene-gene interactions that underlie human diseases, and indicated that many existing methods miss pure interactions in the absence of main effects. In real applications, methods without the heredity constraint enjoy better flexibility and are more suitable for models with pure epistatic interactions. This motivated new methods of detecting interactions without any heredity assumptions. For example, Fan et al. (2015) proposed a two-stage procedure "IIS-SQDA" for detecting important interactions for two-class classification with possibly unequal covariance matrices in a high-dimensional setting. Li and Liu (2019) considered stepwise conditional likelihood variable selection for discriminant analysis (SODA) to detect both main and quadratic interaction effects in logistic regression and quadratic discriminant analysis models. Tang, Fang and Dong (2020) proposed a method for detecting the interaction effects in regression problems using a one-step penalized M-estimator, and used an ADMM-based algorithm to solve the estimator efficiently. A new algorithm xyz based on random projection was introduced by Thanei, Meinshausen and Shah (2018) to screen in-

teraction effects. This algorithm does not rely on the heredity assumption. Thus, it can detect interaction effects in the absence of corresponding main effects. However, based on our empirical observations, its performance in real applications is not entirely satisfactory, because its accuracy when detecting interaction effects depends largely on the number of random projections. However, we still lack computationally efficient algorithms with statistically guaranteed performance for interaction detection. The aforementioned methods were all developed under a linear or a logistic regression framework.

1.2. Our contribution

Our contribution is to develop a computationally efficient and statistically guaranteed method for interaction detection in high-dimensional problems:

a. We propose a new sure screening procedure (SSI) based on the increment of the log-likelihood function to fully detect significant interactions in high-dimensional GLMs. Furthermore, to reduce the computational burden, we take advantage of computer architecture such as parallel techniques and Boolean operations to construct a more computationally efficient algorithm, BOLT-SSI, and detect interaction effects in a large-scale data set. For example, for the data set Northern Finland Birth Cohort (NFBC) with $n = 5{,}123$ individuals and $p = 319{,}147$ SNPs, the number of interactions is about 5×10^{10}. BOLT-SSI can quickly screen all these interactions within a short time; see Section 6.

b. We investigate the sure screening properties of SSI and BOLT-SSI from theoretical insights, and show that our computationally efficient methods are statistically guaranteed. We provide implementations of both the core SSI algorithm and its extension BOLT-SSI in the R package BOLT-SSI, available on the authors' website (`https://github.com/daviddaigithub/BOLTSSIRR`).

c. More importantly, our work is a practical attempt to integrate the advantages of well-designed computer architecture and statistically rigorous methodology to promote the application of computational structure in statistical modeling and practice, especially in the era of "big data". We hope this example motivates more combinations of statistical methods and computational techniques, greatly improving the computational performance of statistical methods.

The rest of this paper is organized as follows. In Sections 2 and 3, we pro-

pose the sure screening algorithms SSI and BOLT-SSI for detecting interactions in a ultrahigh-dimensional generalized linear regression model, where we briefly introduce the Boolean representation and operations. The theoretical properties of sure screening for the proposed methods are investigated in Section 4. In Section 5, we examine the finite-sample performance of SSI and BOLT-SSI compared with that of the alternative methods RAMP, xyz-algorithm, and IP using simulation studies. In Section 6, three real data sets are used to demonstrate the utility of our approaches. Our findings and conclusions are summarized in Section 7. All proofs are available in the Supplementary Material.

2. Sure Screening Methods for Interaction in GLM

2.1. GLMs with two-way interaction

Assume that given the predictor vector $\boldsymbol{x}$, the conditional distribution of the random variable Y belongs to an exponential family, with a probability density function that has the canonical form $f_{Y|\boldsymbol{x}}(y|\boldsymbol{x}) = \exp[y\theta(\boldsymbol{x}) - b\{\theta(\boldsymbol{x})\} + c(y)]$, where $b(\cdot)$ and $c(\cdot)$ are some known functions and $\theta(\boldsymbol{x})$ is a canonical natural parameter. Here, we ignore the dispersion parameter ϕ in the canonical form, because we concentrate on the estimation of the mean regression function. It is well known that the exponential family includes the binomial, Gaussian, gamma, inverse-Gaussian and Poisson distributions.

We consider the following GLM with two-way interactions:

$$E(Y|\boldsymbol{X}) = b'\{\theta(\boldsymbol{X})\} = g^{-1}\left(\beta_0 + \sum_{i=1}^{p}\beta_i X_i + \sum_{i<j}\beta_{ij}X_iX_j\right) \tag{2.1}$$

for the canonical link function $g^{-1}(\cdot) = b'$, with $\theta(\boldsymbol{X}) = \beta_0 + \sum_{i=1}^{p}\beta_i X_i + \sum_{i<j}\beta_{ij}X_iX_j \hat{=} \beta_0 + \sum_{i=1}^{p}\beta_i X_i + \sum_{i<j}\beta_{ij}X_{ij}$, where $\boldsymbol{X} = (\boldsymbol{X}_{\mathcal{C}}^T, \boldsymbol{X}_{\mathcal{I}}^T)^T$ with $\boldsymbol{X}_{\mathcal{C}} = (X_0, X_1, X_2, X_3, \ldots, X_p)^T$ and $\boldsymbol{X}_{\mathcal{I}} = (X_{12}, X_{13}, \ldots, X_{(p-1)p})^T$. For simplicity, we assume that $X_0 = 1$ and each of the other predictor variables is standardized with zero mean and unit variance. The corresponding sets of coefficients are $\boldsymbol{\beta}_{\mathcal{C}} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T \in \mathbb{R}^p$ and $\boldsymbol{\beta}_{\mathcal{I}} = (\beta_{12}, \beta_{13}, \ldots, \beta_{(p-1)p})^T \in \mathbb{R}^q$, where $q = \binom{p}{2} = p(p-1)/2$.

In a ultrahigh-dimensional regression model, we usually assume there is a sparse structure in the underlying model. This means that only a few of predictor variables or features are significantly correlated with the response Y. Hence, for the above model with two-way interactions, we assume there are only a small number of interactions contributing to the response Y. Denote the true param-

eter $\boldsymbol{\beta}^\star = (\boldsymbol{\beta}_{\mathcal{C}}^{\star T}, \boldsymbol{\beta}_{\mathcal{I}}^{\star T})^T$, where $\boldsymbol{\beta}_{\mathcal{C}}^\star = (\beta_0^\star, \beta_1^\star, \beta_2^\star, \ldots, \beta_p^\star)^T \in \mathbb{R}^{p+1}$ for the main effects, and $\boldsymbol{\beta}_{\mathcal{I}}^\star = (\beta_{12}^\star, \beta_{13}^\star, \ldots, \beta_{(p-1)p}^\star)^T \in \mathbb{R}^q$ with $q = \binom{p}{2} = p(p-1)/2$ for the interactions. Let $\mathcal{N}_\star = \{(i,j) : \beta_{ij}^\star \neq 0,\ 1 \leq i < j \leq p\}$ and $s_n = |\mathcal{N}_\star|$. Then, the nonsparsity size s_n is a relative small number compared with the dimension p of the model.

2.2. SSI for two-way interaction in GLM

The model (2.1) can be simply rewritten in the ordinary generalized linear regression model form $E(Y|\boldsymbol{X}) = b'(\theta(\boldsymbol{X})) = g^{-1}(\boldsymbol{X}^T\boldsymbol{\beta})$. Fan, Samworth and Wu (2009) suggested selecting the important variables by sorting the marginal likelihood, and Fan and Song (2010) pointed out that this technique can be considered as marginal likelihood ratio screening, which builds on the difference between two marginal log-likelihood functions. If we regard the interaction variable X_{ij} the same as the other main effects from the predictor variables X_i, X_j, by considering the marginal likelihood of (X_{ij}, Y), we could directly apply the sure screening techniques of Fan, Samworth and Wu (2009) and Fan and Song (2010) to detect the significant interaction effects. However such a direct screening method ignores the main effects of X_i and X_j, as argued by Jaccard, Wan and Turrisi (1990), often leading to false discoveries for the pure significant interaction effects. Hence, we consider the following sure screening procedure to detect pure interaction effects in the model (2.1).

The random samples $\{(\boldsymbol{X}^{(k)}, Y^{(k)}, k = 1, \ldots, n\}$ are independent and identically distributed (i.i.d.) from the model (2.1) with the canonical link. Let $\boldsymbol{X}_{ij} = (1, X_i, X_j, X_{ij})^T$ and $\boldsymbol{X}_{i,j} = (1, X_i, X_j)^T$. Their coefficients are expressed as $\boldsymbol{\beta}_{ij} = (\beta_{ij0}, \beta_i, \beta_j, \beta_{ij})^T$ and $\boldsymbol{\beta}_{i,j} = (\beta_{i,j0}, \beta_{i,}, \beta_{j,})^T$, respectively. The first step of the sure screening procedure for detecting the interaction effects (SSI) is to calculate the maximum marginal likelihood estimator $\hat{\boldsymbol{\beta}}_{ij}^M$ using the minimizer of the marginal regression $\hat{\boldsymbol{\beta}}_{ij}^M = \operatorname{argmin}_{\boldsymbol{\beta}_{ij}} \mathbb{P}_n\{l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}, Y)\}$, where $l(\theta, Y) = b(\theta) - \theta Y - c(Y)$ and $\mathbb{P}_n f(\boldsymbol{X}, Y) = n^{-1}\sum_{k=1}^n f(\boldsymbol{X}_i^{(k)}, Y_i^{(k)})$ is the empirical measure. Similarly, we can calculate the maximum marginal likelihood estimator $\hat{\boldsymbol{\beta}}_{i,j}^M$ without the interaction effect using the minimizer of the marginal regression $\hat{\boldsymbol{\beta}}_{i,j}^M = \operatorname{argmin}_{\boldsymbol{\beta}_{i,j}} \mathbb{P}_n\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}, Y)\}$.

Correspondingly, let the population version of the above minimizers of the marginal regressions be $\boldsymbol{\beta}_{ij}^M = \operatorname{argmin}_{\boldsymbol{\beta}_{ij}} \mathrm{E}\{l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}, Y)\}$ and $\boldsymbol{\beta}_{i,j}^M = \operatorname{argmin}_{\boldsymbol{\beta}_{i,j}} \mathrm{E}\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}, Y)\}$. In fact, the coefficient β_{ij}^M measures the importance of the interaction terms from population insight. Though the real joint regression parameter $\beta_{ij}^\star$ is not the same as the marginal regression coefficient β_{ij}^M, we

still expect that, under mild conditions, $|\beta_{ij}^M|$ or the increment of the marginal log-likelihood function $L_{ij}^\star = \mathrm{E}\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M, Y) - l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M, Y)\}$ is large, if and only if $|\beta_{ij}^\star|$ is large. Hence, the second step of the SSI procedure is to calculate the increment of the empirical maximum marginal likelihood function, $L_{ij,n} = \mathbb{P}_n\{l(\boldsymbol{X}_{i,j}^T\hat{\boldsymbol{\beta}}_{i,j}^M, Y) - l(\boldsymbol{X}_{ij}^T\hat{\boldsymbol{\beta}}_{ij}^M, Y)\}$ and $\boldsymbol{L}_n = (L_{12,n}, \ldots, L_{(p-1)p,n})^T \in \mathbb{R}^q$. Then, $L_{ij,n}$ measures the strength of the interaction X_{ij} in the marginal model from the empirical version. A larger $L_{ij,n}$, similarly to L_{ij}^*, indicates that the interaction X_{ij} contributes more to the response Y. The final step of the SSI procedure is to sort the vector $\boldsymbol{L}_n$ in decreasing order, and given the threshold value γ_n, to select the interaction effect variables $\widehat{\mathcal{N}}_{\gamma_n} = \{(i,j) : L_{ij,n} \geq \gamma_n,\ 1 \leq i < j \leq p\}$ as the final candidates of the significant pure interaction effects.

Under regularized conditions and similarly to the classical approach, it is not difficult to show that SSI has the so-called "sure screening properties." Therefore, we relegate investigations of the SSI properties to the Supplementary Material. From practical insight, the proposed SSI procedure's computational complexity is $O(p^2 n)$. When p is of moderate size $(10^3 - 10^4)$, SSI can quickly screen all interaction terms. It can be accelerated further using parallel computing, because all the interaction terms can be evaluated independently.

3. BOLT-SSI

Despite the simplicity of SSI, it cannot be scaled up to handle the case that the dimensionality p is very large, for example, $p = 10^6$. In such a scenario, as in other methods, we can impose similar uncheckable heredity assumptions to shrink the screening space of SSI to detect the interaction effects. However for such an approach, some significant interaction effects may never be discovered. Hence, even with enough large observational samples, the method's efficiency could still be worst. The other approach is to use a rough, but fast algorithm or calculation method to approximate and accelerate SSI's speed to deal with ultrahigh-dimensional scenarios. From a theoretical perspective, this would not decrease the original SSI algorithm's complexity and has to sacrifice SSI stability. Such an approach would not lose much information about the data or miss essential discoveries. In particular, because the number of observations is sufficiently large, such an approach's statistical efficiency could be satisfied by the requirements of real applications. This is the trade-off between statistical efficiency and computational efficiency.

Using the computer's computational architecture, we follow the second approach and present a computationally efficient algorithm named "BOLT-SSI"

for detecting interactions in ultrahigh-dimensional problems. The BOLT-SSI algorithm is motivated by the following fact: when X_j, X_k, and Y are discrete variables, the interaction effects of X_j and X_k on Y measured by a logistic regression can be calculated exactly based on a few numbers in the contingency table of X_j, X_k, and Y. These numbers can be obtained efficiently by designing a new data structure and its associated operations, that is, a Boolean representation and Boolean operations. To handle continuous or countable variables, we propose discretization first, and then use the above strategy for screening. This section describes the BOLT-SSI algorithm, and the next section establishes the statistical theory that guarantees its performance.

3.1. Equivalence between logistic models and log-linear models

When all predictors and the response are categorical variables, we usually take the logistic model (for a binary response) or baseline-category logit models (for a response with several categories) to fit the data set. Actually, the logistic regression models or baseline-category logit models have corresponding log-linear regression models for the contingency table when the predictor and the response are categorical (see Agresti and Kateri (2011), Chapter 9, Section 9.5). Based on this equivalence, the significance of the interaction effects can be measured by the increment of the corresponding log-linear regression models.

We consider the following two logistic models with main effects and the full model: $\text{logit}\{P(Y = 1|X, Z)\} = \beta_0 + \beta_i^X + \beta_j^Z$ and $\text{logit}\{P(Y = 1|X, Z)\} = \beta_0 + \beta_i^X + \beta_j^Z + \beta_{ij}^{XZ}$. Denote $\widehat{l}_M$ and $\widehat{l}_F$ as the sample versions of the negative maximum log-likelihood for the above logistic regression models with main effects and the full model, respectively. The increment of the log-likelihood function is defined as $\widehat{l}_M - \widehat{l}_F$. The corresponding log-linear regression models can be expressed as the homogeneous association regression model $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Z + \lambda_k^Y + \lambda_{ij}^{XZ} + \lambda_{ik}^{XY} + \lambda_{jk}^{ZY}$ and the saturated model $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Z + \lambda_k^Y + \lambda_{ij}^{XZ} + \lambda_{ik}^{XY} + \lambda_{jk}^{ZY} + \lambda_{ijk}^{XZY}$. Let $\widehat{l}_H$ and $\widehat{l}_S$ be the sample version of the negative maximum log-likelihood for the homogeneous association regression model and the saturated model, respectively. $\widehat{l}_H - \widehat{l}_S$ is the corresponding increment of the log-likelihood function. Thus, we can take advantage of $\widehat{l}_H - \widehat{l}_S$ to screen the interaction terms instead of using $\widehat{l}_M - \widehat{l}_F$.

Now, we want to obtain the difference $\widehat{l}_H - \widehat{l}_S$. Suppose that we have one three-way $(I \times J \times K)$ table with cell counts $\{n_{ijk}\}$ of random variables X, Z, and Y. The kernel of the log-likelihood function for this contingency table is $L(\boldsymbol{\mu}) = \sum_{ijk} n_{ijk} \log(\mu_{ijk}) - \sum_{ijk} \mu_{ijk}$. Denote that $\pi_{i++} = \sum_{jk} \pi_{ijk}$ is the

marginal probability of $X = i$ and $n_{i++} = \sum_{jk} n_{ijk}$ is the number of samples with $X = i$, and $\pi_{ij+} = \sum_k \pi_{ijk}$ is the marginal probability of $X = i$ and $Z = j$ and $n_{ij+} = \sum_k n_{ijk}$ is the corresponding count. Similarly, $\pi_{+j+} = \sum_{ik} \pi_{ijk}$, $\pi_{++k} = \sum_{ij} \pi_{ijk}$, $\pi_{i+k} = \sum_j \pi_{ijk}$, $n_{i+k} = \sum_j n_{ijk}$, $\pi_{+jk} = \sum_i \pi_{ijk}$, $n_{+j+} = \sum_{ik} n_{ijk}$, $n_{++k} = \sum_{ij} n_{ijk}$, and $n_{+jk} = \sum_i n_{ijk}$.

For the saturated model, we know that $\widehat{\mu}_{ijk} = n_{ijk}$ and directly get the estimation $\widehat{l}_S = \sum_{ijk} n_{ijk} \log(n_{ijk}) - \sum_{ijk} n_{ijk}$. For the homogeneous association regression model, the iterative proportional fitting (IPF) algorithm Deming and Stephan (1940) is used to calculate the estimate of u_{ijk} efficiently. Three steps are included in the first cycle of the IPF algorithm: $\mu_{ijk}^{(1)} = \mu_{ijk}^{(0)}(n_{ij+}/\mu_{ij+}^{(0)})$, $\mu_{ijk}^{(2)} = \mu_{ijk}^{(1)}(n_{i+k}/\mu_{i+k}^{(1)})$, and $\mu_{ijk}^{(3)} = \mu_{ijk}^{(2)}(n_{+jk}/\mu_{+jk}^{(2)})$, where $\mu_{ij+} = \sum_k \mu_{ijk}$, and $\mu_{i+k} = \sum_j \mu_{ijk}$, $\mu_{+jk} = \sum_i \mu_{ijk}$. This cycle does not stop until the process converges. The convergence property has been proved by Fienberg (1970) and Haberman (1974). We count the number n_{ijk} by using the Boolean representation. Thus, the contingency table for X and Z given Y can be constructed in a fast manner. In this way, we obtain the estimation $\widehat{l}_H$.

Consequently, we can take advantage of this equivalence to efficiently estimate the corresponding increment of the log-likelihood function using the IPF algorithm when the predictors and the response are qualitative. If some variables are continuous, we can discretize them; see the next section. In Section 4, we show that our algorithm is still statistically guaranteed after discretization.

She and Tang (2019) revisited the IPF, showing that it can be modified slightly to deliver coefficient estimates. They also discovered an interesting connection between the IPF and majorization-minimization (MM) algorithms, and employed state-of-the-art optimization techniques to develop highly scalable IPF algorithms (IPS) (without using parallel computation). We do not use this version of the IPS algorithms because we consider a simple model with two main effects and one interaction term. However, it is possible to accelerate our algorithm by replacing the original IPF algorithm with the new IPS algorithm.

3.2. Discretization

If some of the predictors and/or response are continuous or countable, we suggest discretizing them, simply binned by equal width or frequency. Considering the variation of random observations, it is more reasonable to use the equal-frequency method by quantiles to split the domain of variables into several intervals. The number of intervals is called the "arity" in the discretization context (Liu et al. (2002)). Assume that the arity is denoted by l, and then $l - 1$ is

the maximum number of cut-points of the continuous features.

We follow the assumption of Fan and Song (2010), and consider variable or feature selection of the following GLM: $Y = b'(\boldsymbol{X}^T\boldsymbol{\beta}) + \varepsilon$, where $\boldsymbol{X} = (X_1,\ X_2, \ldots,\ X_p)^T$ is a $p \times 1$ random vector, $\boldsymbol{\beta} = \{\beta_1,\ \beta_2, \ldots,\ \beta_p\}$ is the parameter vector, Y is the response, $b'(\cdot)$ is the canonical link function, and we assume that $\mathcal{M}_\star = \{1 \le k \le p :\ \beta_k \neq 0\}$ is the set of indices of nonzero parameters. Define the marginal log-likelihood increment $L_k^\star = \mathrm{E}\{l(\beta_0^M, Y) - l(\boldsymbol{X}_k^T\boldsymbol{\beta}_k^M, Y)\}$, for $k = 1, 2, \ldots, p$, where $\beta_0^M = \mathrm{argmin}_{\beta_0}\,\mathrm{E}l(\beta_0, Y)$, $\boldsymbol{X}_k^T = \{1, X_k\}$, $\boldsymbol{\beta}_k^M = \{\beta_{k,0}, \beta_k^M\}^T$, and $\boldsymbol{\beta}_k^M = \mathrm{argmin}_{\boldsymbol{\beta}_k}\,\mathrm{E}l(\boldsymbol{X}_k^T\boldsymbol{\beta}_k, Y)$. Furthermore, $\mathrm{E}(Y) = \mathrm{E}(X_k) = 0$ and $\mathrm{E}(Y^2) = \mathrm{E}(X_k^2) = 1$, for $k = 1, 2, \ldots, p$. Let $\rho_k = \mathrm{Corr}(Y, X_k)$ and (Y_1, X_{1k}), (Y_2, X_{2k}) be independent copies of (Y, X_k).

Assume that S^{X_k} and S^Y are the support sets of the variables X_k and Y, respectively. Denote that $\{P_i^{X_k}\}_{i=1}^l$ and $\{P_j^Y\}_{j=1}^m$ are partitions of their supports, which means that $\bigcup_{i=1}^l P_i^{X_k} = S^{X_k}$ and $P_{i_1}^{X_k} \bigcap P_{i_2}^{X_k} = \emptyset$ for $i_1 \neq i_2$, and $\bigcup_{j=1}^m P_j^Y = S^Y$ and $P_{j_1}^Y \bigcap P_{j_2}^Y = \emptyset$ for $j_1 \neq j_2$, where l and m are two positive constants. Here, the l-quantiles and m-quantiles are considered as break points for the partitions of the variables X_k and Y. Define $\widetilde{X}_k = i - 1$ if $X_k \in P_i^{X_k}$, for $i = 1, \ldots, l$, and $\widetilde{Y} = j - 1$ if $Y \in P_j^Y$, for $j = 1, \ldots, m$. Then, the variables X_k and Y are discretized into two categorical variables, $\widetilde{X}_k$ and $\widetilde{Y}$, respectively. Furthermore, denote that $\widetilde{X}_{k_i} = I(X_k \in P_i^{X_k})$, for $1 \le i \le l$, and $\widetilde{Y}_j = I(Y \in P_j^Y)$, for $1 \le j \le m$, where $I(\cdot)$ is the indicator function. After discretization, we have the new increment of the log-likelihood function as $\widetilde{L}_k^\star = \mathrm{E}\{l(\widetilde{\beta}_0^M, \widetilde{Y}) - l(\widetilde{\boldsymbol{X}}_k^T\widetilde{\boldsymbol{\beta}}_k^M, \widetilde{Y})\}$, for $k = 1, 2, \ldots, p$.

Now, we consider the discretization for the marginal model with the interaction effect. Assume that S^{X_i}, S^{X_j}, and S^Y are the support sets of the variables X_i, X_j, and Y, respectively. Denote that $\{P_s^{X_i}\}_{s=1}^{l_1}$, $\{P_t^{X_j}\}_{t=1}^{l_2}$, and $\{P_k^Y\}_{k=1}^m$ are partitions of their supports, which means that $\bigcup_{s=1}^l P_s^{X_i} = S^{X_i}$ and $P_{s_1}^{X_i} \bigcap P_{s_2}^{X_i} = \emptyset$ for $s_1 \neq s_2$, $\bigcup_{t=1}^l P_t^{X_j} = S^{X_j}$ and $P_{t_1}^{X_j} \bigcap P_{t_2}^{X_j} = \emptyset$ for $t_1 \neq t_2$, and $\bigcup_{k=1}^m P_k^Y = S^Y$ and $P_{k_1}^Y \bigcap P_{k_2}^Y = \emptyset$ for $k_1 \neq k_2$, where l_1, l_2, and m are positive constants. Here, we still consider the l_1-quantiles, l_2-quantiles, and m-quantiles as the break points for the partitions of the variables X_i, X_j, and Y, respectively. Define $\widetilde{X}_i = s - 1$ if $X_i \in P_s^{X_i}$, for $s = 1, \ldots, l_1$, and $\widetilde{X}_j = t - 1$ if $X_j \in P_t^{X_j}$, for $t = 1, \ldots, l_2$. Furthermore, denote that $\widetilde{X}^{ij} = u - 1$ for $u = 1, \ldots, l_1 * l_2$, if $X_i \in P_s^{X_i}$ and $X_j \in P_t^{X_j}$. In addition, we define the discretized response $\widetilde{Y}$ as $\widetilde{Y} = j - 1$ if $Y \in P_j^Y$, for $j = 1, \ldots, m$. Hence, we have the new categorical predictor $\widetilde{X}_i$, $\widetilde{X}_j$, and the response $\widetilde{Y}$. We also get the new interaction variable $\widetilde{X}^{ij}$. Furthermore, denote that $\widetilde{X}_{st}^{ij} = I(\{X_i \in P_s^{X_i}\} \bigcap \{X_j \in P_t^{X_j}\})$, for $1 \le s \le l_1, 1 \le t \le l_2$, and

$\widetilde{Y}_j = I(Y \in P_j^Y)$, for $1 \le j \le m$, where $I(\cdot)$ is the indicator function. After discretization, the new increment of the log-likelihood function in the population version is defined as $\widetilde{L}_{ij}^{\star} = E\{l(\widetilde{\boldsymbol{X}}_{i,j}^T \widetilde{\boldsymbol{\beta}}_{i,j}^M, \widetilde{Y}) - l(\widetilde{\boldsymbol{X}}_{ij}^T \widetilde{\boldsymbol{\beta}}_{ij}^M, \widetilde{Y})\}, 1 \le i < j \le p$.

Remark 1. Actually, there is a trade-off between the arity l and the accuracy of the screening procedures. Higher arity leads to a more accurate sure screening. However, when the sample size of the data is large enough, a relatively small arity l can also guarantee the accuracy of the screening procedure, from our theoretical investigation and numerical studies. Hence, large l_i for different continuous features X_i can also be used. We recommend using $l = 2, 3$ for the trade-off between the computational burden and the efficiency of the model estimation for our proposed BOLT-SSI when the sample size of the data is relatively large. Furthermore, if Y is a continuous response, we suggest using the two-quantile (median) to split the response Y, that is, $m = 2$ and $\widetilde{Y} = 0$ if $Y \le M_d(Y)$, and $\widetilde{Y} = 1$ if $Y > M_d(Y)$, where $M_d(Y)$ is the median of the response Y. Furthermore, if X and Y are countable, they can be discretized more like the continuous case, because they are counting data with an order.

3.3. Boolean representation and logical operations

After discretization, the Boolean operation can be used to speed up the SSI procedure, especially the algorithm to calculate $\widetilde{L}_k^*$. The Boolean representation and its operations is a classical and fundamental computer computing technique. A standard floating computation that provides a basic operation for many statistical applications comprises of hundreds of Boolean operations under a lower level of computing. Hence, if the Boolean operation can be directly applied to realize the proposed algorithm, the computational speed can be much improved.

Assume that the continuous data set $\boldsymbol{X}$ is one $n \times p$ matrix with n observations and p predictors, and Y is the response. After discretizing the data set $\boldsymbol{X}$ and response Y, each predictor $\widetilde{X}_i$ has l levels and $\widetilde{Y}$ has m categories. Here, we take $l = 3$ and $m = 2$ as an example. Assuming that $\widetilde{Y}$ has two values (0 and 1), then instead of using one row for each predictor $\widetilde{X}_i$, the new representation uses three rows, because three levels are included in each $\widetilde{X}_i$. Each row consists of two-bit strings, one for samples with $\widetilde{Y} = 0$, and the other for those with $\widetilde{Y} = 1$. Each bit represents one sample in the string. The values (0 and 1) illustrate whether the sample belongs to such a categorical level for each predictor X_i. For instance, we have one discretized data set $\widetilde{\boldsymbol{X}}$ with two predictors and 16 samples, where the first eight columns represent samples with $\widetilde{Y} = 0$, and the others represent samples with $\widetilde{Y} = 1$:

$$\widetilde{\boldsymbol{X}}^T = \begin{matrix} \widetilde{Y} \\ \widetilde{X}_1 \\ \widetilde{X}_2 \end{matrix} \begin{bmatrix} 0\,0\,0\,0\,0\,0\,0\,0 \vdots 1\,1\,1\,1\,1\,1\,1\,1 \\ 1\,3\,2\,3\,1\,2\,3\,2 \vdots 2\,2\,1\,1\,3\,2\,2\,1 \\ 3\,2\,1\,1\,3\,2\,2\,1 \vdots 2\,3\,2\,3\,1\,2\,3\,2 \end{bmatrix}.$$

Its Boolean representation is

$$\widetilde{\boldsymbol{X}}_{bit}^T = \begin{matrix} \\ \widetilde{X}_1 = 1 \\ \widetilde{X}_1 = 2 \\ \widetilde{X}_1 = 3 \\ \widetilde{X}_2 = 1 \\ \widetilde{X}_2 = 2 \\ \widetilde{X}_2 = 3 \end{matrix} \begin{bmatrix} \widetilde{Y} = 0 & \widetilde{Y} = 1 \\ 10001000 & 00110001 \\ 00100101 & 11000110 \\ 01010010 & 00001000 \\ 00110001 & 00001000 \\ 01000110 & 10100101 \\ 10001000 & 01010010 \end{bmatrix}.$$

From the Boolean representation $\widetilde{\boldsymbol{X}}_{bit}$, we find that the first sample belongs to the first category of X_1 and the third category of X_2. Further, we can quickly obtain the number of observations that belong to any two categories by taking the logic operation. For example, if we want to calculate the number of samples with $\widetilde{X}_1 = 2$ and $\widetilde{X}_2 = 2$ in the category $\widetilde{Y} = 0$, we just conduct the logical **AND** operation: "00100101 **AND** 01000110 = 00000100," and then count the number of 1s in the final string "00000100", that is, one. As a result, it is more efficient to use $\widetilde{\boldsymbol{X}}_{bit}$ to construct the contingency table for any two discretized predictors. Because we use the fast logic operation with $\widetilde{\boldsymbol{X}}_{bit}$, we can accelerate our computation for our algorithm.

Obviously, $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{X}}_{bit}$ are equivalent, and store the same amount of information. Because one byte is composed of 8 bits, $\widetilde{\boldsymbol{X}}_{bit}$ uses 128 bits to save the data, but $\widetilde{\boldsymbol{X}}$ uses 32×64 bits, 16 times the space of $\widetilde{\boldsymbol{X}}_{bit}$, to save the same data if our computer is a 64-bit computer system. As a result, the Boolean representation could dramatically reduce the storage space of the data, and the large data can be uploaded directly into the RAM, or even saved in the cache. The transferring time for the data between the hard disk and the RAM, and that between the RAM and the cache, can be reduced significantly. This is the other advantage of the Boolean representation or the discretization.

3.4. New algorithm "BOLT-SSI"

In this section, we discuss our algorithm BOLT-SSI. For our ultrahigh-dimensional GLM (2.1), instead of calculating the increment $\widetilde{L}_{ij,n} = \widehat{l}_{M_{ij}} - \widehat{l}_{F_{ij}}$ for any pair of $\widetilde{X}_i$ and $\widetilde{X}_j$, we compute the new increment of the log-likelihood function

$\widetilde{L}'_{ij,n} = \widehat{l}_{H_{ij}} - \widehat{l}_{S_{ij}}$ using the IPF method. Then, by taking the thresholding value γ_n or choosing the large $d = \lfloor n/\log n \rfloor$ or $\max(n, p)$, we obtain the selected sure screening set $\widehat{\mathcal{N}}_{\gamma_n}$. Our algorithm BOLT-SSI is summarized as follows:

Step 1. For any pair of continuous variables X_i and X_j, for $1 \le i < j \le p$, transform them to the corresponding discretized variables $\widetilde{X}_i$ with level l_i and $\widetilde{X}_j$ with level l_j, and change the response Y to a categorical variable $\widetilde{Y}$, if necessary.

Step 2. Directly calculate $\widehat{l}_{S_{ij}}$, and use the IPF algorithm to approximately estimate $\widetilde{l}_{H_{ij}}$. Then compute $\widetilde{L}'_{ij,n} = \widehat{l}_{H_{ij}} - \widehat{l}_{S_{ij}}$ for all pairs of X_i and X_j.

Step 3. Choose the threshold γ_n and select the following interactions: $\widetilde{\mathcal{N}}_{\gamma_n} = \{(i, j) : \widetilde{L}'_{ij,n} \ge \gamma_n,\ 1 \le i < j \le p\}$. Usually, we select the d largest $L_{ij,n}$, where $d = \max(n, p)$.

Sometimes, the dimension p is very large and can be in the order of tens of millions. The IPF method may be time consuming when computing all $\widehat{l}_{H_{ij}}$. Here, we propose using an approximation tool to prune the interaction terms in the second step. For the homogeneous association regression model in Section 3.1, we use the Kirkwood Superposition Approximation (KSA), first proposed by Kirkwood (1935), to provide an estimator for μ_{ijk} in this model. That is, $\widehat{\mu}_{ijk}^{KSA} = (n/\eta)\{\widehat{\pi}_{ij+}\widehat{\pi}_{i+k}\widehat{\pi}_{+jk}/(\widehat{\pi}_{i++}\widehat{\pi}_{+j+}\widehat{\pi}_{++k})\}$, where $\eta = \sum_{ijk}\{\widehat{\pi}_{ij+}\widehat{\pi}_{i+k}\widehat{\pi}_{+jk}/(\widehat{\pi}_{i++}\widehat{\pi}_{+j+}\widehat{\pi}_{++k})\}$ is a normalization term, $n = \sum_{ijk} n_{ijk}$. Then, we get the approximation $\widehat{l}_{KSA}$ for $\widehat{l}_{H_{ij}}$. Wan et al. (2010) shows that $\widehat{l}_{KSA} - \widehat{l}_S$ is an upper bound of $\widehat{l}_H - \widehat{l}_S$, that is, $0 \le \widehat{l}_H - \widehat{l}_S \le \widehat{l}_{KSA} - \widehat{l}_S$. Based on this boundary and by setting up one threshold γ_{KSA}, in the second step, we can filter out many insignificant interaction terms quickly, and reduce the size of the pool of all interaction effects. The value γ_{KSA} can be defined by the conservative Bonferroni correction or specified by the user. Obviously, if $\gamma_{KSA} = 0$, no interaction term is deleted in this step. In the final step, for the remaining interaction terms, we compute $\widetilde{L}'_{ij,n}$ using the IPF algorithm. Then, select the d largest $\widetilde{L}'_{ij,n}$, where $d = \max(n, p)$ or $\lfloor n/\log n \rfloor$, or take the thresholding value γ_n to obtain the sure screening set $\widehat{\mathcal{N}}_{\gamma_n}$. The term γ_n can be taken as the Bonferroni correction $100*(1-0.05*p(p-1)/2)\%$ percentile decided by the χ^2 test with degrees of freedom $(l_i - 1)(l_j - 1)$ for any one interaction between $\widetilde{X}_i$ and $\widetilde{X}_j$. In summary, our algorithm BOLT-SSI with KSA is summarized as follows:

Step 1. For any pairs of continuous variables X_i and X_j, for $1 \le i < j \le p$, transform them to corresponding discretized variables $\widetilde{X}_i$ with level l_i and

$\widetilde{X}_j$ with level l_j, and change the response Y to a categorical variable $\widetilde{Y}$, if necessary.

Step 2. By using the KSA to approximate $\widetilde{l}_{H_{ij}}$ of the IPF algorithm for all pairs of X_i and X_j, we compute $\widehat{l}_{KSA_{ij}} - \widehat{l}_{S_{ij}}$ and set up the threshold γ_{KSA} to remove part of the interaction terms.

Step 3. For the remaining interaction effects, we compute $\widetilde{L}'_{ij,n} = \widehat{l}_{H_{ij}} - \widehat{l}_{S_{ij}}$ and further identify the important interaction effects using the χ^2-test with degrees of freedom $(l_i - 1)(l_j - 1)$, or directly selecting the d largest $\widetilde{L}'_{ij,n}$.

So far, we have specified the procedures of our new algorithm BOLT-SSI. Apparently, the new method BOLT-SSI is much faster than the original method SSI. Even though BOLT-SSI loses some statistical efficiency by discretizing the predictor variables or response variable, its sure screening properties can still be guaranteed for moderate or large sample sizes. Moreover, compared with other screening methods, BOLT-SSI does not rely on hierarchy assumptions, but screens significant two-way interactions for all pairs among the predictors.

4. Sure Screening Properties of BOLT-SSI

In this section, we derive the sure screening properties of BOLT-SSI by discussing SIS's relationship and discretization. SIS was first proposed by Fan and Lv (2008) for screening features. Later, works discussed this issue further, such as Fan, Samworth and Wu (2009); Fan and Song (2010); Fan, Feng and Song (2011); Chang, Tang and Wu (2013); Chen, Weng and Chu (2013); Saldana and Feng (2018), and Pan et al. (2018). The details of the sure screening properties of SSI are available in Section 1 of the Supplementary Material. We also demonstrate the efficiency loss by discretization in the last part of this section.

4.1. Properties of discretization SIS

First, without considering interaction effects, we investigate the connection between the marginal likelihood and the marginal likelihood after the discrization of the predictor variables and response variables, that is, the connection between SIS and discretized SIS. As discussed in Section 3.1, after discretization, we have a new increment of the log-likelihood function $\widetilde{L}_k^\star = E\{l(\widetilde{\beta}_0^M, \widetilde{Y}) - l(\widetilde{\boldsymbol{X}}_k^T \widetilde{\boldsymbol{\beta}}_k^M, \widetilde{Y})\}$, for $k = 1, 2, \ldots, p$, with $m = 2$ and $l \geq 2$. First, we need some marginally symmetric conditions. These conditions are used to investigate the sure screening properties of the rank robust SIS procedure of Li et al. (2012).

(M1) Let (Y_1, X_{1k}), (Y_2, X_{2k}) be independent copies of (Y, X_k). Denote $\Delta\varepsilon_k =$

$Y_1 - Y_2 - \rho_k(X_{1k} - X_{2k})$ and $\Delta X_k = X_{1k} - X_{2k}$, where $\rho_k = corr(Y, X_k)$. The conditional distribution of $\Delta\varepsilon_k$ given ΔX_k is a symmetric finite mixture distribution, that is, $f_{\Delta\varepsilon_k|\Delta X_k}(t) = \pi_{0k} f_0(t, \sigma_0^2|\Delta X_k) + (1 - \pi_{0k}) f_1(t, \sigma_1^2|\Delta X_k)$, where $f_0(t, \sigma_0^2|\Delta X_k)$ is a symmetric unimodal probability distribution, $f_1(t, \sigma_1^2|\Delta X_k)$ is a symmetric probability distribution function, and σ_0^2, σ_1^2 are conditional variances related to ΔX_k, for $k \in \mathcal{M}_\star$. Furthermore, there exists a given positive constant $\pi^\star \in (0, 1]$ such that $\pi_{0k} \geq \pi^\star$, for any $k \in \mathcal{M}_\star$.

(M2) $c_{\mathcal{M}_\star} = \min_{k \in \mathcal{M}_\star} E|X_k|$ is a positive constant and is free of p.

(M3) The predictors $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ and the error term ε_i are independent, for $i = 1, 2, \ldots, n$.

Theorem 1. *Under the marginally symmetric conditions* (M1)−(M3) *and the condition of Theorem* 3 *in Fan and Song (*2010*), that is, for* $k \in \mathcal{M}_\star$, $|Cov(b'(\boldsymbol{X}^T\boldsymbol{\beta}^\star), X_k)| \geq C_1 n^{-\kappa}$, *where* C_1 *is a positive constant and* $\kappa < 1/2$. *After using the two-quantile and l-quantiles to discretize the response* Y *and the predictor* X_k, *we have*

(1) at least one $\widetilde{X}_{k_i}$, *such that* $|Cov(\widetilde{Y}, \widetilde{X}_{k_i})| \geq C_2 n^{-\kappa}$, *for some positive constant* C_2.

(2) Furthermore, $\min_{k \in \mathcal{M}_\star} \widetilde{L}_k^\star \geq C_3 n^{-2\kappa}$, *for some positive constant* C_3, *and* $\widetilde{L}_k^\star$ *is the corresponding increment of the log-likelihood after discretization.*

Theorem 1 ensures that if the predictor variables in the original scale are associated with the response, they are also related to each other after discretization. Therefore, as in our argument above, combining the Boolean representation, logical operation, and discretization could provide a fast way of screening the predictor variables in high-dimensional GLMs without losing much efficiency. This stimulates us to apply discretization to the interaction pursuit. Based on the results above, we obtain a similar connection between SSI and discretized SSI (BOLT-SSI).

4.2. Properties of BOLT-SSI

As before, we need the following marginally symmetric conditions to investigate the screening properties of BOLT-SSI.

Let $\zeta_{ij} = Y - b'(\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M)$ and $(Y_1, X_{1i}, X_{1j}, X_{1ij}, \zeta_{1ij})$, and $(Y_2, X_{2i}, X_{2j}, X_{2ij}, \zeta_{2ij})$ be independent copies of $(Y, X_i, X_j, X_{ij}, \zeta_{ij})$. We further centralize ζ_{ij} and denote that $\rho_{ij} = \mathrm{Cov}(\zeta_{ij}, X_{ij}) / \sqrt{\mathrm{Var}(\zeta_{ij})\mathrm{Var}(X_{ij})}$.

(M1′) Denote $\Delta\varepsilon_{ij} = \zeta_{1ij} - \zeta_{2ij} - \rho_{ij}(X_{1ij} - X_{2ij})$ and $\Delta X_{ij} = X_{1ij} - X_{2ij}$. Then the conditional distribution of $\Delta\varepsilon_{ij}$ given ΔX_{ij} is a symmetric finite mixture distribution, that is, $f_{\Delta\varepsilon_{ij}|\Delta X_{ij}}(t) = \pi_{0ij} f_0(t, \sigma_0^2|\Delta X_{ij}) + (1 - \pi_{0ij}) f_1(t, \sigma_1^2|\Delta X_{ij})$, where $f_0(t, \sigma_0^2|\Delta X_{ij})$ is a symmetric unimodal probability distribution, $f_1(t, \sigma_1^2|\Delta X_{ij})$ is a symmetric probability distribution function, and σ_0^2, σ_1^2 are conditional variances related to ΔX_{ij}, for $i, j \in \mathcal{N}_\star$. Furthermore, there exists a constant $\pi^\star \in (0, 1]$ such that $\pi_{0ij} \geq \pi^\star$, for any $i, j \in \mathcal{N}_\star$.

(M2′) $c_{\mathcal{N}_\star} = \min_{i,j \in \mathcal{N}_\star} E|X_{ij}|$ is a positive constant and is free of p.

(M3′) The predictors $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ and the error term ε are independent.

Remark 2. In fact, the marginally symmetric condition (M1′) is also easily satisfied. Denote that $\varepsilon_{ij} = \zeta_{ij} - \rho_{ij} X_{ij}$. A special case is that under the linear model, the conditional distribution of ε_{ij} given X_{ij} does not depend on X_{ij} and it has K modes, where K is finite. This implies that the conditional distribution $\varepsilon_{ij}|X_{ij}$ is the same as the distribution of ε_{ij}. Suppose that ε_{1ij} and ε_{2ij} follow a distribution $f_\varepsilon(t)$ with K modes, that is, $f_\varepsilon(t) = \sum_{k=1}^K \pi_k f_k(t)$, where $\pi_k \geq 0$ and $\sum_{k=1}^K \pi_k = 1$. Moreover, assume that $f_{lm}^\star(t)$, for $1 \leq l, m \leq K$, are the distributions of the difference $Z_l - Z_m$, where Z_l and Z_m are independent and follow the distributions $f_l(t)$ and $f_m(t)$, respectively. Therefore, the distribution of $\Delta\varepsilon_{ij} = \varepsilon_{1ij} - \varepsilon_{2ij}$ can be expressed as

$$
\begin{aligned}
f_{\Delta\varepsilon}(t) &= \sum_l \sum_m \pi_l \pi_m f_{lm}^\star(t) = \sum_l \pi_l^2 f_{ll}^\star(t) + \sum_{l \neq m} \pi_l \pi_m f_{lm}^\star(t) \\
&= \left(\sum_l \pi_l^2 \right) \sum_l \frac{\pi_l^2}{\sum_l \pi_l^2} f_{ll}^\star(t) + \left(1 - \sum_l \pi_l^2 \right) \sum_{l \neq m} \frac{\pi_l \pi_m}{1 - \sum_l \pi_l^2} f_{lm}^\star(t) \\
&\triangleq \pi_0^\star f_0^\star(t) + (1 - \pi_0^\star) f_1^\star(t).
\end{aligned}
$$

Obviously, $f_{ll}^\star(t)$ are symmetric unimodal distributions because of the unimodal distributions $f_l(t)$, and then $f_0^\star(t)$ is symmetric and unimodal. Furthermore, $f_1^\star(t)$ is a symmetric and multimodal density function. Moreover, $\pi_0^\star = \sum_l \pi_l^2 \geq (\sum_l \pi_l^2)^2 / K = 1/K$.

As the definition of the conditional linear expectation, provided by Barut, Fan and Verhasselt (2016), denote that $E_L(Y|\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_{ij}^M) = b'(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_{ij}^M)$, $E_L(Y|\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M) = b'(\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M)$, and $\mathrm{Cov}_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M) \equiv E\{X_{ij} - E_L(X_{ij}|\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M)\} \{Y - E_L(Y|\boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_{i,j}^M)\}$.

Theorem 2. *Under the marginally symmetric conditions* (M1′)−(M3′) *and the condition: for* $i, j \in \mathcal{N}_\star$ *with* $|Cov_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)| \geq c_1 n^{-\kappa}$, *where* c_1 *is a positive constant and* $\kappa < 1/4$, *after using the two-quantile,* l_1*-quantiles, and* l_2*-quantiles to discretize the response* Y *and the predictors* X_i *and* X_j, *we have*

(1) at least one $\widetilde{X}_{st}^{ij}$ *such that* $|Cov_L(\widetilde{Y}, \widetilde{X}_{st}^{ij}|\widetilde{\boldsymbol{X}}_{i,j}^T\widetilde{\boldsymbol{\beta}}_{i,j}^M)| \geq c_{10} n^{-\kappa}$, *for some positive constant* c_{10}.

(2) Furthermore, $\min_{i,j\in\mathcal{N}_\star} \widetilde{L}_{ij}^\star \geq c_{11} n^{-2\kappa}$, *for some positive constant* c_{11}, *and* $\widetilde{L}_{ij}^\star$ *is the corresponding increment of the log-likelihood after discretization.*

Theorem 2 claims that important interaction terms are still significant after discretization. Consequently, similarly to the sure screening properties of SSI, we can also show the sure screening properties of BOLT-SSI, that is, it can detect significant interaction effects with large probability, even when the dimension of the model is ultrahigh.

4.3. Discussion of efficiency loss by discretization

By Theorem 1 and Theorem 2, and the steps in Theorems A.5 and A.6 in the Supplementary Material, the sure screening properties of discretization SIS and BOLT-SIS can be guaranteed as the sample size n tends to infinity. However, there is information loss by discretization, and the efficiency of the proposed screening procedure could be much reduced, especially when the arity $l, m = 2$ or 3.

To simplify our analysis of the efficiency loss by discretization, we compare the estimation efficiency of the Pearson correlation ρ between the sample correlation estimate and the estimate by our discretization for the bivariate normal random vector $(X, Y)^T \sim N\left\{(0,0)^T, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right\}$. To discretize X and Y, we consider the worst disretization with the largest information loss, that is, $m = l = 2$, and $\widetilde{X} = I\{X > M_d(X)\}$ and $\widetilde{Y} = I\{Y > M_d(Y)\}$. Then, based on the proof of Theroem 4.1 in the Supplementary Material, we have $\tilde{\rho} = \text{Corr}(\widetilde{X}, \widetilde{Y})$ $= 4E\{I(X_2 > X_1)I(Y_2 > Y_1)\} - 1 = \tau = (2/\pi)\arcsin\rho$, where τ, is the Kendall rank correlation of the bivariate normal random vector (X, Y). It is well known that $\tau = (2/\pi)\arcsin\rho$ for the bivariate normal population. Hence, if we have the estimate $\hat{\tau}$ of the Kendall rank correlation, then the Pearson correlation of the bivariate normal random vector can be estimated as $\hat{\rho}_\tau = \sin(\pi/2)\hat{\tau}$.

Let $\hat{\rho}_s$ be the sample Pearson correlation of X and Y, which is the optimal estimate of the Pearson correlation ρ. Hotelling (1953) shows that the asymptotic

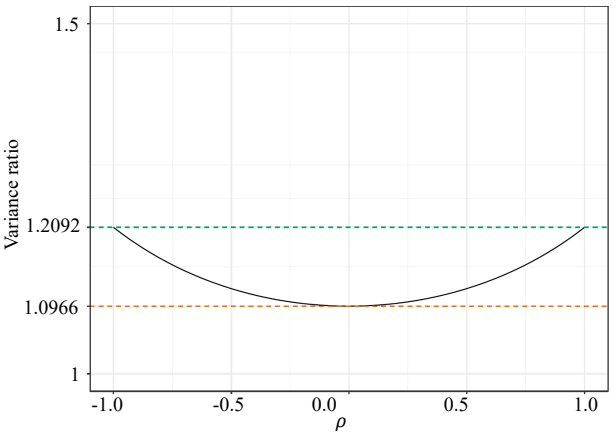

Figure 1. Relative efficiency of $\hat{\rho}_\tau$ and $\hat{\rho}_s$.

property of $\hat{\rho}_s$ under the normal assumption is $\sqrt{n}(\hat{\rho}_s - \rho) \sim N\{0, (1-\rho^2)^2\}$, which implies that $\sqrt{n}\hat{\rho}_s \sim N(0,1)$ when X and Y are independent.

Next, let $\hat{\tau}$ be the sample correlation of $\widetilde{X}$ and $\widetilde{Y}$. As discussed above, it is an estimate of the Kendall rank correlation τ. By the results of Esscher (1924) and Kendall (1949) under the normal assumption, and based on the asymptotic normality of U-statistics (Lee (1990)), the asymptotic distribution of the estimate $\hat{\tau}$ is $\sqrt{n}(\hat{\tau} - \tau) \sim N(0, 4[1/9 - \{(2/\pi)\arcsin(\rho/2)\}^2])$. Then, using the delta method and a simple calculation, the asymptotic normality of $\hat{\rho}_\tau$ is $\sqrt{n}(\hat{\rho}_\tau - \rho) \sim N(0, 4[1/9 - \{(2/\pi)\arcsin(\rho/2)\}^2] * (\pi^2/4)(1-\rho^2))$, that is, $\sqrt{n}\hat{\rho}_\tau \sim N(0, \pi^2/9)$ when $\rho = 0$. Therefore, the relative efficiency of these two procedures is $\mathrm{Var}(\hat{\rho}_\tau)/\mathrm{Var}(\hat{\rho}_s) = 4[1/9 - \{(2/\pi)\arcsin(\rho/2)\}^2] * (\pi^2/4)\{1/(1-\rho^2)\}$.

As shown in Figure 1, such relative efficiency is bounded between $\pi^2/9 \approx 1.0966$ at $\rho = 0$, and $2\sqrt{3}\pi/9 \approx 1.2092$ at $\rho = 1$ or -1. Therefore, we do not need many more samples to get the same accurate estimate of ρ as our discretized estimate $\hat{\rho}_\tau$, compared with the sample Pearson correlation estimate $\hat{\rho}_s$, which is the optimal estimate of ρ in some sense.

Though the above discussion is based on the assumption that (X, Y) follows a bivariate normal population, if (X, Y) follows some other bivariate distribution, by monotonic transformation, we can transfer (X, Y) to one of the bivariate normal random vectors. Usually, under general conditions, such a monotonic transformation would not much change the Pearson correlation between X and Y under general conditions. Furthermore, the discretized estimate $\hat{\rho}_\tau$ is invariant. Hence, in some sense, because the sample size of the data is relatively large, $\hat{\rho}_\tau$ can be used to screen the relationship between X and Y, without losing much

efficiency.

The above discussion is based on the worst discretization that the arity $m = l = 2$. In this case, it has been shown that the statistical efficiency loss is relatively small, but as shown by our numerical studies, the computational complexity is reduced dramatically. Hence, the discretization approach is an appropriate way to balance the trade-off between statistical efficiency and computational complexity. The statistical efficiency loss by discretization can be tolerated, as long as the sample size of the data is relatively large.

5. Numerical Studies

In this section, we investigate the performance of the proposed SSI and BOLT-SSI using numerical studies. By default, we use BOLT-SSI with KSA in our simulation studies. The methods hierNet (Bien, Taylor and Tibshirani (2013)), glinternet (Lim and Hastie (2015)), IP (Fan et al. (2016)), RAMP (Hao, Feng and Zhang (2018)), and xyz (Thanei, Meinshausen and Shah (2018)) are used to compare the performance of the estimation and prediction.

We consider the linear model $y = \sum_{i=1}^{p} X_i\beta_i + \sum_{j<k} X_jX_k\beta_{jk} + \epsilon$ and logistic model $\log\{\pi/(1-\pi)\} = \sum_{i=1}^{p} X_i\beta_i + \sum_{j<k} X_jX_k\beta_{jk}$. We generate the covariates $\{x_i\}_{i=1}^{n} \sim N(0, \Sigma)$ with $\Sigma_{jk} = \rho^{|j-k|}$, where ρ varies in $[0, 0.5]$, and then generate the response y using the above linear model and logistic model. For all settings, the set of important main effects is $S = \{1, 2, \ldots, 10\}$, with the true coefficients $\beta_S = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)^T$. For the linear model, the error term $\epsilon \sim N(0, \sigma^2)$ with $\sigma \in \{2, 3, 4\}$ for different signal-to-noise ratio (SNR) situations. For the logistic model, we change the values of the coefficients of the interactions, and let the significant interaction effect coefficient $\beta_{ij} = 1, 2, 3$ to obtain the different SNR. We consider different heredity structures, including strong heredity, weak heredity, and anti-heredity, using the following interaction effect settings for the linear regression model or logistic model. For the Poisson regression, we discuss the performance of BOLT-SSI in our Supplementary Material.

Example 1. (Linear Model with Strong Heredity). The set of 10 important interaction effects is defined as $T = \{(1, 2), (1, 3), (2, 3), (2, 5), (3, 4), (6, 8), (6, 10), (7, 8), (7, 9), (9, 10)\}$ with corresponding coefficients (2,2,2,2,2,2,2,2,2,2).

Example 2. (Linear Model with Weak Heredity). The set of 10 important interaction effects is defined as $T = \{(1, 2), (1, 13), (2, 3), (2, 15), (3, 4), (6, 10), (6, 18), (7, 9), (7, 18), (10, 19)\}$ with corresponding coefficients (2,2,2,2,2,2,2,2,2,2).

Example 3. (Linear Model with Anti-Heredity). The set of 10 important interaction effects is $T = \{(11,12),(11,13),(12,13),(12,15),(13,14),(16,18),(16,20),(17,18),(17,19),(19,20)\}$ with corresponding coefficients (2,2,2,2,2,2,2,2,2,2)

Example 4. (Linear Model with Mixed Heredity). The set of 10 important interaction effects is $T = \{(1,2),(1,3),(2,3),(2,15),(6,18),(7,18),(16,20),(17,18),(17,19),(19,20)\}$ with corresponding coefficients (2,2,2,2,2,2,2,2,2,2).

Example 5. (Logistic Model with Strong Heredity). Logistic Model with Strong Heredity. The set of 10 important interaction effects is $T = \{(1,2),(1,3),(2,3),(2,5),(3,4),(6,8),(6,10),(7,8),(7,9),(9,10)\}$.

Example 6. (Logistic Model with Weak Heredity]). The set of 10 important interaction effects is $T = \{(1,2),(1,13),(2,3),(2,15),(3,4),(6,10),(6,18),(7,9),(7,18),(10,19)\}$.

Example 7. (Logistic Model with Anti-Heredity). The set of 10 important interaction effects is $T = \{(11,12),(11,13),(12,13),(12,15),(13,14),(16,18),(16,20),(17,18),(17,19),(19,20)\}$.

Example 8. (Logistic Model with Mixed Heredity). The set of 10 important interaction effects is $T = \{(1,2),(1,3),(2,3),(2,15),(6,18),(7,18),(16,20),(17,18),(17,19),(19,20)\}$.

We investigate the screening performance and post-screening performance of the interaction effect screening and variable selection methods under different examples.

Let T with cardinality $t = |T|$ denote the significant interaction effects in the model, that is, $T = \{(j,k) : \beta_{j,k} \neq 0\}$. For each scenario, we run $M = 100$ Monte Carlo simulations for each method. For the mth simulation, denote the estimated interaction subsets as $\widehat{T}_m$. We evaluate the performance in terms of variable selection and model prediction based on the following criteria:

- The average coverage rate (ACR): the percentage of all true interactions included in the selected models.
- Average model size (AMS): $M^{-1}\sum_{m=1}^{M} MS_m$, where MS_m is the model size of the interaction effect predictors selected by the screening methods or post-model selection method in the mth simulation.
- The average out-of-sample R^2 for linear regression model: $R^2 = 100\% \times \left\{1 - \sum(Y_i^* - \boldsymbol{X}_i^{*T}\hat{\boldsymbol{\beta}})^2 / \sum(Y_i^* - \bar{Y}^*)^2\right\}$, where $(\boldsymbol{X}_i^*, Y_i^*)$ is the testing data and $\hat{\boldsymbol{\beta}}$ is the estimate of the coefficient based on the training data.

- Predictive misclassification rate (PMR) for the logistic model: $PMR = I(Y_i^* \neq \hat{Y})$, where Y_i^* is the true value of the testing data, and $\hat{Y}$ is the predictive value of the testing data based on the training model.

5.1. Screening performance

For the screening procedures, we consider SSI, BOLT-SSI, IP and xyz for the linear model and logistic model. For the method xyz, we choose the top 500 interaction terms screened by it (actually, 500 is the largest number of interactions that the package "xyz" can select by screening), and let the projection time L of "xyz" be $10, 100, 1000$, respectively. For the method IP, we choose the top $n-1$ variables as the active set. For our method SSI, the top $n-1$ interaction effect terms are selected into the active set. For BOLT-SSI, we consider two cases: keeping the top $n-1$, or the top $\max\{n, p\}$ significant interaction predictors as the screening selected active set. Because the methods IP and xyz are not available for the logistic model, we only investigate the screening properties of SSI and BOLT-SSI for Examples 5−8.

From the results in Tables 1 and 2, the coverage rate decreases when the SNR is relatively small. The proposed SSI has a high coverage percentage in screening interaction effects for different heredity structures. The methods xyz and IP have a lower converge percentage, except for the strong heredity setting compared with SSI. For the proposed BOLT-SSI, though its performance is not better than that of SSI, its coverage rate is better than the other two methods when the top p significant interaction effects are considered as the screening active set. By discretization, the data lose some information, and hence BOLT-SSI is not as efficient as SSI, even though its speed is much faster. Hence, it would increase the probability of keeping the true active interaction effect predictors in the screened model by keeping the p top significant interaction effect predictors in the active set after screening. All in all, the screening performance of SSI and BOLT-SSI(p) is more stable than that of the other methods.

5.2. Post-screening performance

In this subsection, we compare the final model selection and prediction of existing methods (RAMP, xyz, hierNet, glinternet) with the Lasso after screening by our proposed SSI and BOLT-SSI. For the method RAMP, the tuning parameter is selected using EBIC with $\gamma = 1$, because the EBIC tends to work best among of the settings, as shown by Hao, Feng and Zhang (2018). For the method xyz, we consider the projection time L as 100, 500 and use five-fold cross-validation (CV) to select the tuning parameter for the post-screening selection.

Table 1. Screening results for linear models when $p = 5{,}000$.

Methods	σ	SSI	BOLT-SSI	BOLT-SSI(p)	IP	xyz-L10	xyz-L100	xyz-L1000
				(n, p, ρ)=(500, 5000, 0)				
	2	0.98	0.03	0.64	0.73	0.00	0.01	0.76
Example 1	3	0.94	0.00	0.60	0.70	0.00	0.04	0.73
	4	0.80	0.00	0.48	0.59	0.00	0.01	0.55
				(n, p, ρ)=(500, 5000, 0.5)				
	2	1.00	0.80	0.98	0.99	0.29	0.52	0.52
Example 1	3	1.00	0.58	0.94	0.99	0.22	0.51	0.52
	4	1.00	0.43	0.88	0.98	0.14	0.50	0.50
				(n, p, ρ)=(500, 5000, 0)				
	2	0.90	0.01	0.38	0.03	0.00	0.04	0.56
Example 2	3	0.82	0.01	0.36	0.01	0.00	0.00	0.41
	4	0.73	0.00	0.00	0.01	0.00	0.01	0.31
				(n, p, ρ)=(500, 5000, 0.5)				
	2	0.73	0.03	0.60	0.00	0.00	0.00	0.00
Example 2	3	0.71	0.02	0.57	0.01	0.00	0.00	0.00
	4	0.67	0.00	0.45	0.00	0.00	0.00	0.00
				(n, p, ρ)=(500, 5000, 0)				
	2	0.89	0.03	0.62	0.03	0.00	0.02	0.56
Example 3	3	0.82	0.03	0.44	0.02	0.00	0.01	0.53
	4	0.73	0.00	0.45	0.01	0.00	0.00	0.46
				(n, p, ρ)=(500, 5000, 0.5)				
	2	1.00	0.33	0.81	0.74	0.28	0.53	0.53
Example 3	3	1.00	0.23	0.74	0.72	0.25	0.50	0.50
	4	1.00	0.11	0.73	0.68	0.14	0.51	0.51
				(n, p, ρ)=(500, 5000, 0)				
	2	0.91	0.00	0.44	0.06	0.00	0.03	0.47
Example 4	3	0.82	0.00	0.42	0.05	0.00	0.03	0.48
	4	0.69	0.00	0.23	0.03	0.00	0.00	0.34
				(n, p, ρ)=(500, 5000, 0.5)				
	2	0.80	0.07	0.75	0.27	0.00	0.01	0.01
Example 4	3	0.78	0.05	0.73	0.28	0.00	0.01	0.01
	4	0.76	0.02	0.66	0.28	0.00	0.01	0.01

For our methods SSI and BOLT-SSI, we use five-fold CV and the LASSO to further refine the model selection after screening. All of the simulation settings are the same as those in Examples 1 to 8. We set $\rho = 0.5$ for all the studies. To compare the prediction, for every simulation, we let $n_1 = 0.75n$ of the data set as the training data, and the remaining data are the testing data. Note that we first let p be relatively small so that it is possible to compare the performance of hierNet (Bien, Taylor and Tibshirani (2013)) and glinternet (Lim and Hastie (2015)) in Tables 2−3 of the Supplementary Material, where "w" stands for weak heredity.

Note that the computation time for hierNet-s and glinternet is very large for a single replicate. As a result, we omit the comparisons with hierNet and glinternet

Table 2. Screening results for logistic models with $n = 400$ and $p = 2{,}000$.

Methods	β_{jk}	SSI	BOLT-SSI	BOLT-SSI(p)	SSI	BOLT-SSI	BOLT-SSI(p)
		$\rho = 0$			$\rho = 0.5$		
Example 5	1	0.02	0.00	0.35	0.53	0.08	0.76
	2	0.40	0.04	0.56	0.84	0.30	0.86
	3	0.77	0.12	0.66	0.83	0.27	0.86
Example 6	1	0.02	0.00	0.28	0.00	0.00	0.39
	2	0.31	0.02	0.34	0.32	0.01	0.49
	3	0.56	0.06	0.63	0.44	0.05	0.66
Example 7	1	0.02	0.00	0.35	0.53	0.08	0.76
	2	0.40	0.04	0.56	0.84	0.30	0.86
	3	0.77	0.12	0.66	0.83	0.27	0.86
Example 8	1	0.00	0.00	0.28	0.04	0.00	0.43
	2	0.33	0.05	0.57	0.24	0.04	0.63
	3	0.52	0.05	0.70	0.41	0.13	0.68

for the other higher dimensional examples. In the high dimensional settings, we consider $(n, p) = (500, 5000)$, $(1000, 5000)$, $(1500, 5000)$, $(2000, 5000)$, $(1500, 10000)$, $(1500, 20000)$, and compare the performance of BOLT-SSI, RAMP, and xyz. Other methods are very time consuming, and are not considered in this setting. We set $\sigma = 2$ for the linear models, and $\beta_{ij} = 3$ for the logistic models. All results of the methods with $(n, p) = (1000, 5000)$ are summarized in Table 3. It is shown that our method still has good performance in the high-dimensional feature space. Furthermore, we take Examples 5 and 8 to illustrate the patterns of our method. The results are shown in Figures 1−4 in the Supplementary Material. Obviously, as sample size n increases, all of the methods perform better, as shown in Figure 1 and Figure 3 in the Supplementary Material, and our method performs best. In Figures 2 and 4 in the Supplementary Material, though the performance of our method degrades as the dimension p increases, its performance is still much better than that of others. The method RAMP is influenced by the heredity assumption, especially if the anti-heredity exists, and so the result of RAMP is worst.

5.3. Efficiency comparison

Here, we use Example 1 and Example 5 to study the efficiency of all the above methods. The machine we used is an Intel (R) Xenon(R) CPU E5-1603 v4 @ 2.80GHZ with 8.00 GB RAM. We compare the average computation time of variable selection among the following methods: SSI, BOLT-SSI, xyz, RAMP-s, RAMP-w, hierNet-s, and hierNet-w, based on the 50 simulated data sets by the screening procedure and the post-screening procedure, where "w" and "s" stand

Table 3. Selection and prediction results (standard error) with $(n, p) = (1000, 5000)$. The standard errors are shown in parentheses.

Assumption	Methods	ACR	AMS	R^2	PMR
Example 1	BOLT-SSI	0.98	53.91(2.5)	94.52(0.22)	—
	RAMP	0.16	21.67(0.7)	76.29(1.60)	—
	xyz-L100	0.73	28.10(0.7)	58.46(0.95)	—
	xyz-L500	1	23.94(0.2)	60.07(0.82)	—
Example 2	BOLT-SSI	0.62	45.80(2.3)	87.16(0.62)	—
	RAMP	1.00	20.35(0.1)	95.34(0.01)	—
	xyz-L100	0.23	72.70(2.9)	58.5 (1.16)	—
	xyz-L500	0.97	35.64(0.5)	76.43 (0.56)	—
Example 3	BOLT-SSI	0.93	47.61(1.8)	90.94(0.33)	—
	RAMP	0.00	4.5 (0.6)	13.96(0.11)	—
	xyz-L100	0.80	27.85(7.1)	58.48(1.31)	—
	xyz-L500	1	23.94(0.2)	59.36(1.20)	—
Example 4	BOLT-SSI	0.53	49.38(1.9)	88.53(0.50)	—
	RAMP	0.00	15.54(0.6)	61.83(0.79)	—
	xyz-L100	0.34	47.26(1.8)	59.53(1.05)	—
	xyz-L500	1	28.47(0.5)	68.44(0.89)	—
Example 5	BOLT-SSI	0.53	36.09(4.0)	-	23.26(0.32)
	RAMP	0.00	0.14(0.1)	-	25.62(0.03)
Example 6	BOLT-SSI	0.42	47.75(4.9)	-	26.73(0.62)
	RAMP	0.00	6.80(0.5)	-	28.15(0.60)
Example 7	BOLT-SSI	0.62	79.80(5.0)	-	20.98(0.31)
	RAMP	0.00	2.97(0.2)	-	28.67(0.31)
Example 8	BOLT-SSI	0.53	79.26(5.1)	-	22.85(0.41)
	RAMP	0.00	1.69(0.1)	-	25.34(0.24)

for weak heredity and strong heredity, respectively. To make fair comparisons, we do not consider the selection of tuning parameters in modeling. Figures 5−6 in the Supplementary Material. and Table 4 summarize the average computation time (seconds per run) for each procedure. Because the differences of computation time are relative small for various σ and ρ, we only present the results when $\sigma = 2$, $\beta_{jk} = 2$, and $\rho = 0.5$. It is clear that the method hierNet spends much time on the computation, no matter under the strong or weak heredity assumption, and the method RAMP with weak heredity is also very slow. BOLT-SSI is consistently fast and its screening of the algorithm does not rely on the heredity assumption of the data structure.

In summary, compared with other methods, our proposed SSI and BOLT-SSI(p) have a stably high coverage rate in terms of screening performance. When the dimension of the data p is not too large, by fine coding, SSI can also finish the screening task in a limited time. After discretization, some data informa-

Table 4. Average computation time of post-screening procedure for linear and logistic models.

n	p	BOLT-SSI	hierNet-s	hierNet-w	xyz-L100	xyz-L500	RAMP-s	RAMP-w
Linear Regression Models								
500	50	1.13	75.26	4.92	0.22	0.86	25.00	28.85
500	100	2.55	321.88	22.43	0.39	1.61	33.11	42.44
500	500	1.66	—	669.99	2.10	10.07	60.65	106.82
500	5,000	34.75	—	—	30.38	155.22	68.20	658.42
200	1,000	1.62	—	—	3.58	18.35	6.69	53.35
400	1,000	2.26	—	—	4.15	20.69	57.68	107.11
800	1,000	4.02	—	—	5.32	25.52	54.18	230.20
Logistic Regression Models								
500	50	0.44	306.91	11.53	—	—	139.52	147.16
500	100	0.82	1,105.96	37.16	—	—	177.84	207.08
500	500	0.74	—	511.21	—	—	311.87	368.86
500	5,000	27.15	—	—	—	—	127.52	1,281.45
200	1,000	1.10	—	—	—	—	12.34	83.98
400	1,000	1.38	—	—	—	—	94.48	273.06
800	1,000	2.18	—	—	—	—	588.62	820.87

tion is lost, and hence BOLT-SSI cannot use all of the information for screening, and hence is not as efficient as SSI. However, it is much faster than SSI and most of the other screening methods, and can finish screening for ultrahigh-dimensional data in a relatively short time. In fact, from our numerical studies, it is shown that BOLT-SSI makes a good trade-off between the computation complexity and the efficiency of screening. Consequently, SSI and BOLT-SSI have absolute competitiveness compared with other interaction screening and variable selection methods. In particular, when the computational cost becomes unaffordable for SSI, we believe that BOLT-SSI is a valuable tool for high-dimensional or ultrahigh-dimensional interaction screening.

6. Real Data

The real data was collected from a major supermarket located in northern China and has been analyzed by Wang (2009) and Hao, Feng and Zhang (2018), which includes 6,398 predictors and 464 observations. The response is the number of customers on a particular day, and each predictor is the corresponding sale volume of the product. The supermarket manager wonders which products would be more associated with the number of customers, which means that he or she wants to select the most informative products to predict the response. Note that here, the total number of interaction terms for the supermarket data in modeling

Table 5. Average results and the standard errors (in parentheses) on the supermarket data set.

	main size	inter size	R^2(%)
BOLT-SSI	196.19(3.79)	42.43(1.13)	93.95(0.15)
SSI	107.70(0.73)	10.90(0.37)	92.73(0.14)
xyz-L10	37.80(0.26)	12.61(0.25)	87.03(0.26)
xyz-L100	35.54(0.24)	14.40(0.23)	86.94(0.22)
xyz-L500	35.26(0.25)	14.84(0.24)	86.59(0.28)
RAMP-AIC	229.18(1.68)	94.53(1.06)	90.48(0.23)
RAMP-BIC	101.17(3.25)	34.36(1.65)	91.18(0.20)
RAMP-EBIC	29.27(1.01)	3.07(0.29)	89.67(0.31)
RAMP-GIC	30.71(0.92)	3.20(0.30)	90.08(0.28)
iFORT	———	———	88.91(0.17)
iFORM	———	———	88.66(0.18)
LASSO-AIC	264.28(0.91)	0(0)	92.04(0.18)
LASSO-BIC	63.47(0.77)	0(0)	90.76(0.20)
LASSO-EBIC	15.62(0.46)	0(0)	72.09(0.53)
LASSO-GIC	19.19(0.74)	0(0)	75.05(0.58)
LASSO-AIC-m	30.72(0.61)	———	82.65(0.40)
LASSO-BIC-m	13.21(0.22)	———	69.58(0.48)

is about 2×10^7, much larger than the number of interaction effects to model the residential building data; see the Supplementary Material.

Here, we randomly select 400 observations as the training data and the remaining 64 observations as the testing data, and then use the out-of-sample R^2 to evaluate the prediction performance of our methods based on 100 random splits. The settings of all methods are the same as those of the above example. The average performance is summarized in Table 5, which includes the average sizes of the main effects and interaction effects, the average out-of-sample R^2, and their standard errors over 100 experiments. In addition to the results of our methods, Table 5 displays the out-of-sample R^2 by other methods, including RAMP-AIC, RAMP-BIC, RAMP-EBIC, RAMP-GIC, iFORT & iFORM, and RAMP. The corresponding results are extracted directly from the respective papers. We extract the results of LASSO-AIC, LASSO-BIC, LASSO-EBIC, and LASSO-GIC, from Hao, Feng and Zhang (2018) (RAMP). For LASSO-AIC-m and LASSO-BIC-m, we only consider the main effects. From the results in Table 5, BOLT-SSI demonstrates the best performance, with a mean out-of-sample $R^2 = 93.95\%$. Although BOLT-SSI selects more products, and it is a challenging task for the supermarket manager to interpret them, more products can improve

Table 6. Average computation time on the supermarket data set.

Methods	BOLT-SSI	SSI	xyz-L10	xyz-L100	xyz-L500	RAMPs	RAMPw
Time(s)	98.81	431.55	59.09	463.15	2,252.95	33.75	NULL

the supermarket's profit. Therefore, our method is helpful for the supermarket manager to make a decision.

To fairly assess the efficiency of the methods BOLT-SSI, SSI, xyz, and RAMP on this real data set, we use the computer as previously. Time(s) is the average computation time of five experiments, including variable selection and prediction. The results are listed in Table 6. Here, the result "NULL" means that the error exists. When we only run one time using "RAMP" with the weak heredity assumption, the error "cannot allocate vector of size 1.1 Gb" appears, which implies that the method "RAMP" may not be widely used on some ordinary computers when the dimension of the data set is huge. From the above two tables, in the first step of our screening methods, we use only marginal information of the data, or even sacrifice some information for the method BOLT-SSI. However, the advantages of computational efficiency are evident. For BOLT-SSI, the sacrifice of the data information can be ignored, which is consistent with our theoretical investigation.

7. Conclusion

We have presented a screening method for detecting important significant interaction effects in the high-dimensional GLMs. A new and straightforward procedure SSI and its extension BOLT-SSI are proposed. In contrast to most other screening or variable selection methods for detecting interaction effects, our proposed methods do not depend on the heredity assumption. The proposed screening methods conduct a full screening search for all of the interaction effects among the data. For ultrahigh-dimensional data, in some sense, such a task seems impossible. Here, we show that, by taking advantage of the computational structure, seemly impossible tasks can be done using a standard personal computer. Importantly, the statistical property of the proposed method is guaranteed by our established theory.

Our numerical studies consider only screening interaction effects for our method, even if p is ultrahigh. In real problems, if p is ultrahigh and the regularization methods cannot obtain a reasonable optimal solution in a limited time and with limited computing resources, we should also screen the main effects and interaction effects simultaneously.

In general, most data analysis projects are similar to engineering projects. Though most of the theoretical research would be beneficial to such projects, the requirements and expectations of engineering projects differ from those of theoretical studies. How to combine the advantages of engineering techniques to complete such projects under practical requirements and expectations requires further investigation.

Supplementary Material

To conserve space, all discussions and the sure screening properties of SSI and their proofs are relegated to Sections 1 and 2 of the Supplementary Material. In addition, Section 2 includes proofs of Theorems 1 and 2 about the sure screening properties of BOLT-SSI. Section 3 also contains part of the simulation for the data set with a small dimension and a dicussion on how to choose between SSI and BOLT-SSI. Two additional case studies are presented in Section 4. In one, the data dimension is huge, with $p = 319{,}156$. The total number of interaction terms is about 5×10^{10}. The R-package "BOLTSSIRR" contains the code for algorithms decribed in the article.

Acknowledgments

The authors thank the associate editor and two anonymous referees for their helpful comments. This work was supported in part by the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College (2022B1212010006), Guangdong Higher Education Upgrading Plan (2021-2025) (UIC R0400001-22), the Hong Kong Research Grant Council [16307818, 16301419, 16308120, 12303618], the RGC Collaborative Research Fund: C6021-19EF, the Initiation Grant for Faculty Niche Research Areas RC-FNRA-IG/20-21/SCI/05 from Hong Kong Baptist University, Grant R-913-200-098-263 from Duke-NUS Medical School, and AcRF Tier 2 (MOE2018-T2-1-046, MOE2018-T2-2-006) from the Ministry of Education, Singapore.

References

Agresti, A. and Kateri, M. (2011). *Categorical Data Analysis*. 3rd Edition. John Wiley & Sons, Hoboken.

Barut, E., Fan, J. and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association* **111**, 1266–1277.

Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge.

Bien, J., Taylor, J. and Tibshirani, R. (2013). A Lasso for hierarchical interactions. *The Annals of Statistics* **41**, 1111.

Chandrasekaran, V. and Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences* **110**, E1181–E1190.

Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics* **41**.

Chen, R.-B., Weng, J.-Z. and Chu, C.-H. (2013). Screening procedure for supersaturated designs using a Bayesian variable selection method. *Quality and Reliability Engineering International* **29**, 89–101.

Choi, N. H., Li, W. and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association* **105**, 354–364.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics* **10**, 392–404.

Cox, D. R. (1984). Interaction. *International Statistical Review/Revue Internationale de Statistique*, 1–24.

Culverhouse, R., Suarez, B. K., Lin, J. and Reich, T. (2002). A perspective on epistasis: Limits of models displaying no main effect. *The American Journal of Human Genetics* **70**, 461–471.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics* **11**, 427–444.

Esscher, F. (1924). On a method of determining correlation from the ranks of the variates. *Scandinavian Actuarial Journal* **1924**, 201–219.

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.

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.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32**, 928–961.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research* **10**, 2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604.

Fan, Y., Kong, Y., Li, D. and Lv, J. (2016). Interaction pursuit with feature screening and selection. *arXiv:1605.08933*.

Fan, Y., Kong, Y., Li, D. and Zheng, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics* **43**, 1243–1272.

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics* **41**, 907–917.

Fisher, R. A. (1918). XV.—The correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transaction of the Royal Society of Edinburgh* **52**, 399–433.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. The University of Chicago Press, Chicago.

Hao, N., Feng, Y. and Zhang, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association* **113**, 615–625.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **109**, 1285–1301.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)* **15**, 193–232.

Jaccard, J., Wan, C. K. and Turrisi, R. (1990). The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate Behavioral Research* **25**, 467–478.

Kendall, M. G. (1949). Rank and product-moment correlation. *Biometrika* **36**, 177–193.

Kirkwood, J. G. (1935). Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics* **3**, 300–313.

Kong, Y., Li, D., Fan, Y. and Lv, J. (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics* **45**, 897–922.

Lee, A. J. (1990). *U-statistics: Theory and Practice*. Routledge, New York.

Lees, P., Cunningham, F. and Elliott, J. (2004). Principles of pharmacodynamics and their applications in veterinary pharmacology. *Journal of Veterinary Pharmacology and Therapeutics* **27**, 397–414.

Li, D., Kong, Y., Fan, Y. and Lv, J. (2021). High-dimensional interaction detection with false sign rate control. *Journal of Business & Economic Statistics* **40**, 1234–1245.

Li, G., Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40**, 1846–1877.

Li, Y. and Liu, J. S. (2019). Robust variable and interaction selection for logistic regression and general index models. *Journal of the American Statistical Association* **114**, 271–286.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-Lasso regularization. *Journal of Computational and Graphical Statistics* **24**, 627–654.

Liu, H., Hussain, F., Tan, C. L. and Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery* **6**, 393–423.

Pan, W., Wang, X., Xiao, W. and Zhu, H. (2018). A generic sure independence screening procedure. *Journal of the American Statistical Association*.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F. et al. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics* **69**, 138–147.

Saldana, D. F. and Feng, Y. (2018). SIS: An R package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software* **83**, 1–25.

Shah, R. D. (2016). Modelling interactions in high-dimensional data with backtracking. *Journal of Machine Learning Research* **17**, 1–31.

She, Y. and Tang, S. (2019). Iterative proportional scaling revisited: A modern optimization perspective. *Journal of Computational and Graphical Statistics* **28**, 48–60.

She, Y., Wang, Z. and Jiang, H. (2018). Group regularized estimation under structural hierarchy. *Journal of the American Statistical Association* **113**, 445–454.

Tang, C. Y., Fang, E. X. and Dong, Y. (2020). High-dimensional interactions detection with sparse principal Hessian matrix. *The Journal of Machine Learning Research* **21**, 665–689.

Thanei, G.-A., Meinshausen, N. and Shah, R. D. (2018). The xyz algorithm for fast interaction search in high-dimensional data. *The Journal of Machine Learning Research* **19**, 1343–1384.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. et al. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics* **87**, 325–340.

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

Wang, J.-H. and Chen, Y.-H. (2018). Overlapping group screening for detection of gene-gene interactions: Application to gene expression profiles with survival trait. *BMC Bioinformatics* **19**, 335.

Wang, J.-H. and Chen, Y.-H. (2020). Interaction screening by Kendall's partial correlation for ultrahigh-dimensional data with survival trait. *Bioinformatics* **36**, 2763–2769.

Min Zhou

Beijing Normal University–Hong Kong Baptist University United International College, Zhuhai, Guangdong 519088, China.

E-mail: minzhou@uic.edu.cn

Mingwei Dai

Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan 610074, China.

E-mail: daimw@swufe.edu.cn

Yuan Yao

School of Mathematics and Statistics, Victoria University of Wellington, Kelburn, Wellington 6012, New Zealand.

E-mail: yuan.yao@vuw.ac.nz

Jin Liu

Duke-NUS Graduate Medical School, Singapore 169857.

E-mail: jin.liu@duke-nus.edu.sg

Can Yang

The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

E-mail: macyang@ust.hk

Heng Peng

Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong.

E-mail: hpeng@math.hkbu.edu.hk

(Received February 2021; accepted January 2022)