Statistica Sinica Preprint No: SS-2020-0498	
Title	BOLT-SSI: A Statistical Approach to Screening Interaction Effects for Ultra-High Dimensional Data
Manuscript ID	SS-2020-0498
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0498
Complete List of Authors	Min Zhou, Mingwei Dai, Yuan Yao, Jin Liu, Can Yang and Heng Peng
Corresponding Author	Mingwei Dai
E-mail	daimw@swufe.edu.cn
Notice: Accepted version subject to English editing.	

BOLT-SSI: A Statistical Approach to Screening Interaction Effects for Ultra-High Dimensional Data

Min Zhou[1], Mingwei Dai[2], Yuan Yao[3], Jin Liu[4], Can Yang[5], Heng Peng[6]

Abstract: Detecting interaction effects among predictors on the response variable is a crucial step in various applications. In this paper, we first propose a simple method for sure screening interactions (SSI). Although its computation complexity is $O(p^2n)$, SSI works well for problems of moderate dimensionality (e.g., $p = 10^3 \sim 10^4$), without the heredity assumption. To ultra-high dimensional problems (e.g., $p = 10^6$), motivated by discretization associated Boolean representation and operations and the contingency table for discrete variables, we propose a fast algorithm, named "BOLT-SSI". The statistical theory has been established for SSI and BOLT-SSI, guaranteeing their sure screening property. The performance of SSI and BOLT-SSI are evaluated by comprehensive simulation and real case studies. Numerical results demonstrate that SSI and BOLT-SSI can often outperform their competitors in terms of computational efficiency and statistical accuracy. The proposed method can be applied for fully detecting interactions with more than 300,000 predictors. Based on this study, we believe that there is a great 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: Trade-off between statistical efficiency and computational complexity, Discretization, Sure independent screening for interaction detection, Ultra-high dimensionality, Package "BOLTSSIRR".

1. Introduction

The recent two decades are the golden age for the development of statistical science on high dimensional problems. A large number of innovative algorithms have been proposed to address the computational challenges in statistical inference for high dimensional problems. Despite a fruitful achievement in statistical science, there still exists a gap between the established statistical theory and computational performance of developed algorithms. On one hand, many statistical models can deal with the high dimensional problems under some theoretically mild conditions, but their computational cost can be too expensive to be affordable when dimensionality becomes extremely large. On the other hand, to address many real problems, many algorithms are not developed in a principled way, leading to computational results without statistical guarantees. As argued by Chandrasekaran & Jordan (2013), there is a great need to rethink the relationship between statistical accuracy and computational efficiency.

To bridge the gap, most statistical literatures focus on reducing the theoretical complexity of an algorithm, or simply using parallel computing to speed it up, without paying not enough attention to 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). In this paper, we use the interaction detection problem in high dimensional models as an example, to demonstrate that it is possible to design statistically guaranteed algorithms to overcome seemingly unaffordable computational cost by taking advantage of the computational

architecture.

1.1 Related work for interaction effect detection

The word "interaction", in Oxford English Dictionary, is illustrated as the reciprocal action, or influence of persons or things on each other. It is one kind of relationship among two or more objects, which have mutual influence upon one another. There is a long history of investigating the interaction effects in many different scientific fields (Wang & 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 neither of carbon and steel has much effect on the strength, but a combination of them has substantial effects. In medicine and pharmacology, the interaction effects of multiple drugs have been widely observed (Lees et al. 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, increasing interest has been focusing on detecting gene-gene interactions from genome-wide association studies (GWAS) (Cordell 2009, Wang & Chen 2018).

In this paper, we investigate the interaction effects from a statistical perspective, where the interaction effect is characterized by the statistical departure from the additive effects of two or more factors (see Fisher (1918), Cox (1984)). In the framework of high dimensional regression, it is common to use products of explanatory variables to study interaction effects of explanatory variables on response variables. Consider three explanatory variables X_i, X_j and X_k, their two-way interaction terms are X_jX_k, X_iX_j and X_iX_k. By 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 main effect term X_i, β_{jk} is the coefficient of interaction term X_jX_k, and ε is the independent error. For the 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 would be $p + p(p-1)/2$ if all two-way interaction terms are included. For example, in GWAS, there are millions of genotyped genetic variants, i.e., $p \approx 10^6$. The number of interaction terms goes up to an astronomical number at the order of 10^{12}. The computational cost of detecting interaction effects in such a scale becomes seemingly un-affordable, making the theoretical guarantees with mild conditions (e.g. sparsity assumptions) useless.

To reduce the computational cost, methods developed recently often make two types of heredity assumptions: the strong heredity assumption means that the interaction effect is important only if its both parent are significant, while the weak heredity assumption illustrates that the interaction term is important only if at least one of its parent is included in the model. To name a few, Choi et al. (2010) extended the LASSO method and identified the significant interaction terms in the linear model and generalized linear models under the strong heredity assumption. Choi et al. (2010) proved that their method possessed the oracle property (Fan & Li 2001, Fan & Peng 2004), that is, it performed well as if the true model was known in advance. The algorithm hierNet was developed by Bien et al. (2013) to select the interactions, which added a set of convex constraints to LASSO in the linear model and constructed the sparse interaction model with the strong or weak heredity assumptions. For the linear model, Hao & Zhang (2014) also proposed two algorithms iFORT and iFORM, and identified the interaction effects in a greedy fashion under the heredity assumption. Lim & Hastie (2015) introduced the method "glinternet" for learning pairwise interactions in a linear regression or logistic regression model with strong hierarchy constraint. Hao, Feng & Zhang (2018) further improved interaction detection by proposing a regularization algorithm under marginality principle (RAMP). The method "Backtracking" was developed by Shah (2016), which can be incorporated into many existing high-dimensional methods based on penalty functions, and works by building increasing sets of candidate interactions iteratively. She et al. (2018) proposed group regularized

estimation under structural hierarchy about variable selection for models including interactions, and provided the minimax lower bounds for strong and weak hierarchical variable selection and showed that the proposed estimators enjoy sharp rate oracle inequalities. To deviate 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 ultra-high dimensional linear interaction models. The idea of the IP method is to select the "active interaction variables" by screening significant predictor variables with the strong Pearson correlation between X_j^2 and Y^2 firstly, and then detect the interaction effects among those identified active interaction variables. IP is a good attempt to detect pure interaction effects in the model. Kong et al. (2017) extended IP to the ultra-high 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 $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$ be 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 due to the existence of pure interaction effects. In human genetics, many gene-gene interaction effects have been detected in the absence of their main effects (Cordell 2009, Wan et al. 2010). For instance, Ritchie et al. (2001) made an effort to detect pure epistatic interactions among two or more loci in relatively small samples for common complex multifactorial human diseases. They proposed one method MDR to identify interactions, which was applied to a real data example (sporadic breast cancer case-control data set) to demonstrate the existence of pure interactions. Culverhouse et al. (2002) discussed the interaction models without main effects and examined the pure epistatic interactions whose loci did not display any single-locus effects. Cordell (2009) discussed how to detect gene-gene interactions that underlie human diseases and indicated that the many existing methods would miss pure interactions in the absence of main effects. In real applications, the methods without the heredity constraint can enjoy better flexibility and

be more suitable for models with pure epistatic interactions. This motivates new methods to detect interactions without any heredity assumptions. For example, Fan et al. (2015) proposed a two-stage procedure "IIS-SQDA" to detect important interactions for two-class classification with possibly unequal covariance matrices in the high-dimensional setting. Li & 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 et al. (2020) proposed one method to detect the interaction effects in regression problems by a one-step penalized M-estimator and used ADMM based algorithm to solve the estimator efficiently. A new algorithm xyz based on random projection was introduced by Thanei et al. (2018) to screen interaction effects. This algorithm does not rely on the heredity assumption. Thus it can detect interaction effects in the absence of the corresponding main effects. However, based on our empirical observations, its performance in the real applications is not entirely satisfactory because its accuracy of detecting interaction effects largely depends on the number of random projections. Yet, computationally efficient algorithms with statistically guaranteed performance for interaction detection are still lacking. All of these methods were developed under the linear or 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 log-likelihood function to fully detect significant interactions for the high dimensional generalized linear models. Furthermore, in order to reduce the computational burden, we take the advantages of computer architecture such as parallel techniques and Boolean operations to construct more computationally efficient algorithm BOLT-SSI, and make available the

detection for 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 with a short time. The details can be seen in section 6.

b. Moreover, 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. We take it as an example to promote the application of computational structure in the statistical modeling and practice, especially in the era of "Big Data". We hope this example motivates more combination 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 propose the sure screening algorithms SSI and BOLT-SSI for detecting interactions in ultra-high 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 in comparison to alternative methods, RAMP, xyz-algorithm, and IP, through 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. The details of the proof are given in the Supplementary.

2. Sure Screening Methods for Interaction in GLM

2.1 Generalized linear models(GLM) 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, whose probability density function 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, since we only concentrate on the estimation of mean regression function. It is well known that the distributions in the exponential family include the Binomial, Gaussian, Gamma, Inverse-Gaussian and Poisson distributions.

We consider the following generalized linear model 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 coefficient 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 the ultra-high dimensional regression model, we usually assume that there is a sparse structure in the underlying model. It means that only a few of predictor variables or features are significantly correlated with 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 that the true parameter $\boldsymbol{\beta}^\star = (\boldsymbol{\beta}_{\mathcal{C}}^{\star T}, \boldsymbol{\beta}_{\mathcal{I}}^{\star T})^T$, 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 main

effects, and $\boldsymbol{\beta}^{\star}_{\mathcal{I}} = (\beta^{\star}_{12}, \beta^{\star}_{13}, \ldots, \beta^{\star}_{(p-1)p})^T \in \mathbb{R}^q$ with $q = \binom{p}{2} = p(p-1)/2$ for interactions. Let $\mathcal{N}_{\star} = \{(i,j) : \beta^{\star}_{ij} \neq 0,\ 1 \leq i < j \leq p\}$, and denote that $s_n = |\mathcal{N}_{\star}|$, then the non-sparsity size s_n is a relative small number compared to the dimension p of the model.

2.2 SSI for two-way interaction in GLM

The model (2.1) can be simply rewritten as an ordinary generalized linear regression model form $E(Y|\boldsymbol{X}) = b'(\theta(\boldsymbol{X})) = g^{-1}(\boldsymbol{X}^T\boldsymbol{\beta})$. Fan et al. (2009) suggested to select the important variables by sorting the marginal likelihood, and Fan & Song (2010) pointed out that such technique can be considered as the 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 other main effects from 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 et al. (2009) and Fan & Song (2010) to detect the significant interaction effects. But such a direct screening method ignores the main effects of X_i and X_j, as argued by Jaccard et al. (1990), it often leads 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).

Denote that the random samples $\{(\boldsymbol{X}^{(k)}, Y^{(k)}, k = 1, \ldots, n\}$ are 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 $\boldsymbol{X}_{i,j} = (1, X_i, X_j)^T$. And 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 to detect the Interaction effects (SSI) is to calculate the maximum marginal likelihood estimator $\hat{\boldsymbol{\beta}}^M_{ij}$ by the minimizer of the marginal regression $\hat{\boldsymbol{\beta}}^M_{ij} = \arg\min_{\boldsymbol{\beta}_{ij}} \mathbb{P}_n\{l(\boldsymbol{X}^T_{ij}\boldsymbol{\beta}_{ij}, Y)\}$ where $l(\theta, Y) = b(\theta) - \theta Y - c(Y)$ and $\mathbb{P}_n f(\boldsymbol{X}, Y) = n^{-1}\sum_{k=1}^n f(\boldsymbol{X}^{(k)}_i, Y^{(k)}_i)$ is the empirical measure. Similarly, we can calculate the maximum marginal likelihood estimator $\hat{\boldsymbol{\beta}}^M_{i,j}$ without the interaction effect by the minimizer of the marginal regression $\hat{\boldsymbol{\beta}}^M_{i,j} = \arg\min_{\boldsymbol{\beta}_{i,j}} \mathbb{P}_n\{l(\boldsymbol{X}^T_{i,j}\boldsymbol{\beta}_{i,j}, Y)\}$.

Correspondingly, let the population version of the above minimizers of the marginal regressions be $\boldsymbol{\beta}_{ij}^M = \arg\min_{\boldsymbol{\beta}_{ij}} \mathrm{E}\{l(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}, Y)\}$ and $\boldsymbol{\beta}_{i,j}^M = \arg\min_{\boldsymbol{\beta}_{i,j}} \mathrm{E}\{l(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}, Y)\}$. In fact, the coefficient β_{ij}^M can measure the importance of the interaction terms from population insight. Though the real joint regression parameter $\beta_{ij}^\star$ would not be the same as the marginal regression coefficient β_{ij}^M, we could 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 some 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. The larger $L_{ij,n}$, similar to L_{ij}^*, the more the interaction X_{ij} contributes to the response Y. The final step of the SSI procedure is to sort the vector $\boldsymbol{L}_n$ in a decreasing order and given threshold value γ_n, select the following 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 similar as the classical approach, it is not difficult to show that SSI has the so-called "sure screening properties". So here we delegate those investigations of SSI properties to our supplementary file. From practical insight, the proposed SSI procedure's computational complexity is in the order of $O(p^2n)$. When p is of moderate size $(10^3 - 10^4)$, SSI can quickly screen all interaction terms. It can be further accelerated by parallel computing because all the interaction terms can be evaluated independently.

3. BOLT-SSI

Despite the simplicity of SSI, it can not be scaled up to handle the case that dimensionality p is very large, e.g., $p = 10^6$. To such a scenario, as other methods, we could impose similar uncheckable heredity assumptions to shrink the screening space of SSI to detect the interaction

effects. But for such an approach, some significant interaction effects could never be discovered. Hence, even though we could have 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 ultra-high dimensional scenarios. Though from theoretical insight, it 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 and miss essential discoveries. Especially, as the number of observations is large enough, such an approach's statistical efficiency could be satisfied by the requirement of real applications as our experience. It is the other kind of trade-off between statistical efficiency and computational efficiency.

In this paper, utilizing the computer's computational architecture, we follow the second approach and present a computationally efficient algorithm named "BOLT-SSI" to detect interactions in ultra-high dimensional problems. The BOLT-SSI algorithm is motivated by the following fact: when X_j, X_k and Y all are discrete variables, the interaction effects of X_j and X_k on Y measured by logistic regression can be exactly calculated based on a few numbers in the contingency table of X_j, X_k and Y. These numbers can be efficiently obtained by designing a new data structure and its associated operations, i.e., 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 details of BOLT-SSI algorithm and establishes statistical theory to guarantee its performance in the next section.

3.1 Equivalence between the logistic models and log-linear models

When all predictors and the response are categorical variables, we usually take the logistic model (for binary response) or baseline-category logit models (for the response with several categories) to fit the data set. Actually, the logistic regression models or baseline-category logit

models have their corresponding log-linear regression models for the contingency table when the predictor and the response are categorical (See Agresti & Kateri (2011), Chapter 9 Section 9.5). Based on this equivalence, the significance of interaction effects can be measured by the increment of the corresponding log-linear regression models.

Assume that we consider the following two logistic models with main effects and full model, respectively: $\text{logit}(P(Y=1|X,Z)) = \beta_0 + \beta_i^X + \beta_j^Z$ and $\text{logit}(P(Y=1|X,Z)) = \beta_0 + \beta_i^X + \beta_j^Z + \beta_{ij}^{XZ}$. Denote that $\widehat{l}_M$ and $\widehat{l}_F$ be the sample version of the negative maximum log-likelihood for the above logistic regression models with main effects and 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 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 $X = i$, $\pi_{ij+} = \sum_k \pi_{ijk}$ is the marginal probability of $X = i$ and $Z = j$ and $n_{ij+} = \sum_k n_{ijk}$ is the corresponding count. Similarly, $\pi_{+j+} = \sum_{ik} \pi_{ijk}$, $\pi_{++k} = \sum_{ij} \pi_{ijk}$, $\pi_{i+k} = \sum_j \pi_{ijk}$, $n_{i+k} = \sum_j n_{ijk}$, $\pi_{+jk} = \sum_i \pi_{ijk}$, $n_{+j+} = \sum_{ik} n_{ijk}$, $n_{++k} = \sum_{ij} n_{ijk}$, $n_{+jk} = \sum_i n_{ijk}$.

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 & Stephan (1940) is used to calculate the estimate

of u_{ijk} efficiently. Three steps are included in the first cycle of the IPF algorithm: $\mu_{ijk}^{(1)} = \mu_{ijk}^{(0)} \frac{n_{ij+}}{\mu_{ij+}^{(0)}}$, $\mu_{ijk}^{(2)} = \mu_{ijk}^{(1)} \frac{n_{i+k}}{\mu_{i+k}^{(1)}}$, $\mu_{ijk}^{(3)} = \mu_{ijk}^{(2)} \frac{n_{+jk}}{\mu_{+jk}^{(2)}}$, where $\mu_{ij+} = \sum_k \mu_{ijk}$, , $\mu_{i+k} = \sum_j \mu_{ijk}$, $\mu_{+jk} = \sum_i \mu_{ijk}$. This cycle does not stop until the process converges and the convergence property has been proved by Fienberg et al. (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 quickly constructed in a fast manner. In this way, the estimation $\widehat{l}_H$ will be obtained.

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

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

3.2 Discretization

In the case that 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 would be more reasonable to use the equal-frequency method by quantiles to split the domain of variables to several intervals. The number of intervals is called "arity" in the discretization context (See 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.

For more detail, we follow the assumption of Fan & Song (2010), and consider variable or feature selection of the generalized linear model: $Y = b'(\boldsymbol{X}^T\boldsymbol{\beta})+\varepsilon$, 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 assume that $\mathcal{M}_\star = \{1 \le k \le p :\ \beta_k \neq 0\}$ is the set of indexes of nonzero parameter. Define the marginal log-likelihood increment $L_k^\star = \mathrm{E}\{l(\beta_0^M, Y) - l(\boldsymbol{X}_k^T\boldsymbol{\beta}_k^M, Y)\}, k = 1, 2, \ldots, p$ where $\beta_0^M = \arg\min_{\beta_0} \mathrm{E}l(\beta_0, Y)$, $\boldsymbol{X}_k^T = \{1, X_k\}$, $\boldsymbol{\beta}_k^M = \{\beta_{k,0}, \beta_k^M\}^T$ and $\boldsymbol{\beta}_k^M = \arg\min_{\boldsymbol{\beta}_k} \mathrm{E}l(\boldsymbol{X}_k^T\boldsymbol{\beta}_k, Y)$. Furthermore, $\mathrm{E}(Y) = \mathrm{E}(X_k) = 0$ and $\mathrm{E}(Y^2) = \mathrm{E}(X_k^2) = 1,\ k = 1, 2, \ldots, p$. Let $\rho_k = \mathrm{Corr}(Y, X_k)$ and (Y_1, X_{1k}), (Y_2, X_{2k}) be the independent copies of (Y, X_k).

Assume that S^{X_k} and S^Y are the support sets of 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 the break points for the partitions of variables X_k and Y. Define $\widetilde{X}_k = i-1$ if $X_k \in P_i^{X_k}$, $i = 1, \ldots, l$; $\widetilde{Y} = j-1$ if $Y \in P_j^Y$, $j = 1, \ldots, m$. And then variables X_k and Y are discretized to two categorical variables $\widetilde{X}_k$ and $\widetilde{Y}$, respectively. Furthermore, denote that $\widetilde{X}_{k_i} = I(X_k \in P_i^{X_k})$, $1 \le i \le l$ and $\widetilde{Y}_j = I(Y \in P_j^Y)$, $1 \le j \le m$, where $I(\cdot)$ is the indicator function. After discretization, we have the new increment of log-likelihood function as $\widetilde{L}_k^\star = \mathrm{E}\{l(\widetilde{\beta}_0^M, \widetilde{Y}) - l(\widetilde{\boldsymbol{X}}_k^T\widetilde{\boldsymbol{\beta}}_k^M, \widetilde{Y})\}, \quad k = 1, 2, \ldots, p.$

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 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 variables X_i, X_j and Y, respectively. Define $\widetilde{X}_i = s-1$ if $X_i \in P_s^{X_i}$, $s = 1, \ldots, l_1$ and $\widetilde{X}_j = t-1$ if $X_j \in P_t^{X_j}$, $t = 1, \ldots, l_2$. Furthermore, denote that $\widetilde{X}^{ij} = u-1$, if $u = 1, \ldots, l_1 * l_2$ if $X_i \in P_s^{X_i}$ and $X_j \in P_t^{X_j}$. And also, we define the discretized response $\widetilde{Y}$ as $\widetilde{Y} = j-1$ if $Y \in P_j^Y$, $j = 1, \ldots, m$. Hence, we have the new categorical predictor $\widetilde{X}_i$, $\widetilde{X}_j$ and response $\widetilde{Y}$, respectively. And also, we get the new interaction variable $\widetilde{X}^{ij}$. Furthermore, denote that $\widetilde{X}_{st}^{ij} = I\left(\left\{X_i \in P_s^{X_i}\right\} \bigcap \left\{X_j \in P_t^{X_j}\right\}\right)$, $1 \le s \le l_1, 1 \le t \le l_2$ and $\widetilde{Y}_j = I(Y \in P_j^Y)$, $1 \le j \le m$, where $I(\cdot)$ is the indicator function. After discretization, the new increment of log-likelihood function in 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 screening procedures. Higher arity would lead to a more accurate sure screening. However, when the sample size of data is large enough, the relatively small arity l could also guarantee the accuracy of the screening procedure from our theoretical investigation and numerical studies. Hence though large l_i for different continuous features X_i can be also used. we recommend using $l = 2, 3$ to make a trade-off between the computation burden and efficiency of model estimation for our proposed BOLT-SSI when the sample size of the data is relatively large. Furthermore, if Y is a continuous response , similarly we also suggest to use 2-quantile (median) to split the response Y, that is, $m = 2$ and $\widetilde{Y} = 0$ if $Y \le M_d(Y)$; $\widetilde{Y} = 1$ if $Y > M_d(Y)$, where $M_d(Y)$ is the median of the response Y. What's more, if X and Y are countable, they can also 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 software is composed of hundreds of Boolean operations under a lower level of the computer computing. Hence if the Boolean operation can be directly applied to realize the proposed algorithm, the computational speed could be much improved.

Assume that the continuous data set $\boldsymbol{X}$ is one $n \times p$ matrix with n observations and p predictors, Y be the response. After discretizing 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 3 rows since 3 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 them with $\widetilde{Y} = 1$, and each bit can represent 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 2 predictors and 16 samples, where the first 8 columns represent samples with $\widetilde{Y} = 0$ and the others represent samples with $\widetilde{Y} = 1$:

$$\widetilde{\boldsymbol{X}}^T = \begin{array}{c} \\ \widetilde{Y} \\ \widetilde{X}_1 \\ \widetilde{X}_2 \end{array} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \vdots & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 2 & 3 & 1 & 2 & 3 & 2 & \vdots & 2 & 2 & 1 & 1 & 3 & 2 & 2 & 1 \\ 3 & 2 & 1 & 1 & 3 & 2 & 2 & 1 & \vdots & 2 & 3 & 2 & 3 & 1 & 2 & 3 & 2 \end{bmatrix}$$

and its Boolean representation is

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

From the Boolean representation $\widetilde{\boldsymbol{X}}_{bit}$, we can easily 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, we count the number of 1s in the final string "00000100", that is 1. As a result, it is more efficient by using $\widetilde{\boldsymbol{X}}_{bit}$ to construct the contingency table for any two discretized predictors. Since the fast logic operation with $\widetilde{\boldsymbol{X}}_{bit}$ is utilized, we can accelerate our computation for our algorithm.

Obviously, $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{X}}_{bit}$ are equivalent and they 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 $\widetilde{\boldsymbol{X}}$ would use 32×64 bits, 16 times of 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. So all of the large data could be directly uploaded into the RAM, or even be saved in the cache. The transferring amount of time for the data between hard disk and RAM, and between RAM and cache can be largely reduced. This is the other advantage of the Boolean representation or the discretization.

3.4 New algorithm "BOLT-SSI"

Now, we illustrate our algorithm BOLT-SSI in details. For our ultra-high dimensional generalized linear model (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 $\widetilde{L}'_{ij,n} = \widehat{l}_{H_{ij}} - \widehat{l}_{S_{ij}}$ by the IPF method. Then, by taking the thresholding value γ_n or choosing the large $d = \left\lfloor \frac{n}{\log n} \right\rfloor$ or $\max(n, p)$, the selected sure screening set $\widehat{\mathcal{N}}_{\gamma_n}$ is obtained. Our algorithm BOLT-SSI is summarized as follows:

Step 1. For any pair of the continuous variables X_i and X_j, $1 \le i < j \le p$, 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 $\widetilde{l}_{H_{ij}}$, and 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: $\widetilde{\mathcal{N}}_{\gamma_n} = \{(i,j) : \widetilde{L}'_{ij,n} \geq \gamma_n,\ 1 \leq i < j \leq p\}$. 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 for computing all $\widehat{l}_{H_{ij}}$. Here, we propose to use an approximation tool to prune interaction terms in the second step. For the homogeneous association regression model in section 3.1, Kirkwood Superposition Approximation (KSA), which was firstly proposed by Kirkwood (1935), is utilized to provide an estimator for μ_{ijk} in this model. That is, $\widehat{\mu}_{ijk}^{KSA} = \frac{n}{\eta}\frac{\widehat{\pi}_{ij+}\widehat{\pi}_{i+k}\widehat{\pi}_{+jk}}{\widehat{\pi}_{i++}\widehat{\pi}_{+j+}\widehat{\pi}_{++k}}$, where $\eta = \sum_{ijk}\frac{\widehat{\pi}_{ij+}\widehat{\pi}_{i+k}\widehat{\pi}_{+jk}}{\widehat{\pi}_{i++}\widehat{\pi}_{+j+}\widehat{\pi}_{++k}}$ is a normalization term, $n = \sum_{ijk} n_{ijk}$. And 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 $\widehat{l}_H - \widehat{l}_S$, i.e., $0 \leq \widehat{l}_H - \widehat{l}_S \leq \widehat{l}_{KSA} - \widehat{l}_S$. 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 then reduce the size of a pool of all interaction effects. The value γ_{KSA} can be defined by the conservative Bonferroni correction or specified by 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 their $\widetilde{L}'_{ij,n}$ by the IPF algorithm. Then select the d largest $\widetilde{L}'_{ij,n}$, where $d = \max(n,p)$ or $\left\lfloor \frac{n}{\log n} \right\rfloor$, 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 degree 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, $1 \leq i < j \leq p$, 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 a part of 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 by χ^2-test with degree freedom $(l_i - 1)(l_j - 1)$, or directly select 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" will be much faster than the original method "SSI". Even though BOLT-SSI loses some statistical efficiency by discretizing predictor variables or response variable; its sure screening properties can still be guaranteed for moderate or large sample sizes. Moreover, compared to other screening methods, BOLT-SSI does not rely on hierarchy assumptions but screen 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. SIS was firstly proposed by Fan & Lv (2008) for screening features. Later, many works have discussed further this issue such as Fan et al. (2009), Fan & Song (2010), Fan et al. (2011), Chang et al. (2013), Chen et al. (2013), Saldana & Feng (2018), Pan et al. (2018). The details of sure screening properties of SSI can be seen in section 1 of the Supplementary. And also we 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 discrization of the predictor variables and response variables, i.e., the connection between SIS and Discretized SIS. As discussed in Section 3.1, after discretization we have such new increment of log-likelihood function $\widetilde{L}_k^\star = E\{l(\widetilde{\beta}_0^M, \widetilde{Y}) - l(\widetilde{\boldsymbol{X}}_k^T\widetilde{\boldsymbol{\beta}}_k^M, \widetilde{Y})\}$, $k = 1, 2, \ldots, p.$ with $m = 2$ and $l \geq 2$. We need some marginally symmetric conditions for further studies. Those conditions are used to investigate sure screening properties of a rank robust SIS procedure by Li et al. (2012).

(M1) Let (Y_1, X_{1k}), (Y_2, X_{2k}) be the 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, i.e., $f_{\Delta\varepsilon_k|\Delta X_k}(t) = \pi_{0k} f_0(t, \sigma_0^2|\Delta X_k) + (1 - \pi_{0k}) f_1(t, \sigma_1^2|\Delta X_k)$, where $f_0(t, \sigma_0^2|\Delta X_k)$ is symmetric unimodal probability distribution and $f_1(t, \sigma_1^2|\Delta X_k)$ is a symmetric probability distribution function and σ_0^2, σ_1^2 are conditional variances related to ΔX_k, $k \in \mathcal{M}_\star$. 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, $i = 1, 2, \ldots, n$.

Theorem 1. *Under the marginally symmetric condition (M1)-(M3) and the condition of Theorem 3 in Fan and Song (2010), i.e., 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 2-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 corre-

sponding increments of the log-likelihood after discretization.

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

4.2 Properties of BOLT-SSI

Similar to above, we need the following some 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 $(Y_1, X_{1i}, X_{1j}, X_{1ij}, \zeta_{1ij})$, $(Y_2, X_{2i}, X_{2j}, X_{2ij}, \zeta_{2ij})$ be the independent copies of $(Y, X_i, X_j, X_{ij}, \zeta_{ij})$. We further centralize ζ_{ij} and denote that $\rho_{ij} = \frac{\mathrm{Cov}(\zeta_{ij}, X_{ij})}{\sqrt{\mathrm{Var}(\zeta_{ij})\mathrm{Var}(X_{ij})}}$.

(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, i.e., $f_{\Delta\varepsilon_{ij}|\Delta X_{ij}}(t) = \pi_{0ij} f_0(t, \sigma_0^2|\Delta X_{ij}) + (1 - \pi_{0ij}) f_1(t, \sigma_1^2|\Delta X_{ij})$, where $f_0(t, \sigma_0^2|\Delta X_{ij})$ is symmetric unimodal probability distribution and $f_1(t, \sigma_1^2|\Delta X_{ij})$ is a symmetric probability distribution function and σ_0^2, σ_1^2 are conditional variances related to ΔX_{ij}, $i, j \in \mathcal{N}_\star$. 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. It implies that the conditional distribution $\varepsilon_{ij}|X_{ij}$ is the same as the distribution of ε_{ij}. Suppose that ε_{1ij}, 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 $f^\star_{lm}(t)$, $1 \leq l, m \leq K$, 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

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

Obviously, $f^\star_{ll}(t)$ are symmetric unimodal distributions because of the unimodal distributions $f_l(t)$, and then $f_0^\star(t)$ is symmetric and unimodal. And $f_1^\star(t)$ 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 et al. (2016), denote that $E_L(Y|\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M) = b'(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{ij}^M)$, $E_L(Y|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M) = b'(\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M)$, $\mathrm{Cov}_L(Y, X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M) \equiv E(X_{ij} - E_L(X_{ij}|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M))(Y - E_L(Y|\boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_{i,j}^M))$.

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 2-quantile, l_1-quantiles and l_2-quantiles to discretize the response Y and the predictors X_i, 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, $\min_{i,j\in\mathcal{N}_\star} \widetilde{L}^\star_{ij} \geq c_{11}n^{-2\kappa}$ *for some positive constant* c_{11} *and* $\widetilde{L}^\star_{ij}$ *is the corresponding increments of the log-likelihood after discretization.*

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

4.3 Discussion of Efficiency Loss by Discretization

By Theorem 1 and Theorem 2, and following steps in both Theorem A.5 and A.6 in Supplementary, 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 to obtain the intuition about such efficiency loss by discretization, we just 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 discretization with the largest information loss, i.e. $m = l = 2$, 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, we have $\tilde{\rho} = \mathrm{Corr}(\widetilde{X}, \widetilde{Y}) = 4E[I(X_2 > X_1)I(Y_2 > Y_1)] - 1 = \tau = \frac{2}{\pi}\arcsin\rho$, where τ, in fact, is the kendall rank correlation of the bivariate normal random vector (X,Y). It is well known that $\tau = \frac{2}{\pi}\arcsin\rho$ for the bivariate normal population, and 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 $\hat{\rho}_\tau = \sin\frac{\pi}{2}\hat{\tau}$.

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) has shown that the asymptotic property of $\hat{\rho}_s$ under normal assumption should be, $\sqrt{n}(\hat{\rho}_s - \rho) \sim N\left(0, (1-\rho^2)^2\right)$, 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 discussion above, in fact 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 normaility of U-statistics (Lee 2019), the asymptotic distribution of the estimate $\hat{\tau}$ is $\sqrt{n}(\hat{\tau} - \tau) \sim N\left(0, 4\left[\frac{1}{9} - \left(\frac{2}{\pi}\arcsin\frac{\rho}{2}\right)^2\right]\right)$. Then with Delta method and by simple calculation, the asymptotic normality of $\hat{\rho}_\tau$ should be $\sqrt{n}(\hat{\rho}_\tau - \rho) \sim N\left(0, 4\left[\frac{1}{9} - \left(\frac{2}{\pi}\arcsin\frac{\rho}{2}\right)^2\right] * \frac{\pi^2}{4}(1-\rho^2)\right)$, 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 $\frac{\mathrm{Var}(\hat{\rho}_\tau)}{\mathrm{Var}(\hat{\rho}_s)} = 4\left[\frac{1}{9} - \left(\frac{2}{\pi}\arcsin\frac{\rho}{2}\right)^2\right] * \frac{\pi^2}{4}\frac{1}{1-\rho^2}$.

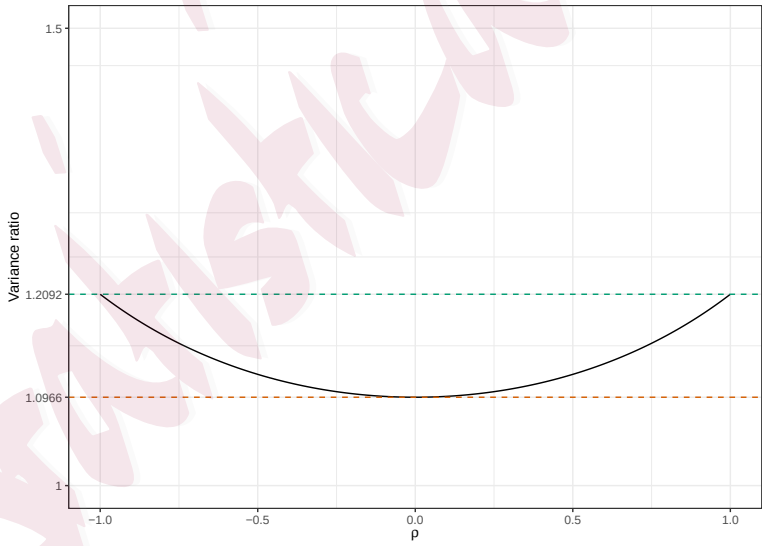

Figure 1: Relative efficiency of $\hat{\rho}_\tau$ and $\hat{\rho}_s$

As shown by the 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. It means we do not need much more samples to get the same accurate estimate of ρ as our discretized estimate $\hat{\rho}_\tau$ compared to 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 bivariate normal population, if (X, Y) follows other bivariate distribution, by monotonic transformation, we could transfer (X, Y) to one of bivariate normal random vectors. Usually, under general conditions, such a monotonic transformation would not change the Pearson correlation between X and Y much under general conditions. Furthermore, the discretized estimate $\hat{\rho}_\tau$ is invariant. Hence in some sense, as the sample size of 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 such worst 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 by numerical studies. By default, we use BOLT-SSI with KSA in our simulation studies. The methods, hierNet (Bien et al. 2013), glinternet(Lim & Hastie 2015), IP(Fan et al. 2016), RAMP (Hao, Feng & Zhang 2018) and xyz(Thanei et al. 2018) are employed in comparisons with respect to the performance on 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 $\log(\frac{\pi}{1-\pi}) = \sum_{i=1}^{p} X_i\beta_i + \sum_{j<k} X_jX_k\beta_{jk}$. 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 by the above linear model and logistic model. For all settings, the set of the 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 interactions, and let 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 by the following interaction effect settings for linear regression model or logistic model. For Poisson regression, we discuss the performance of BOLT-SSI in our Supplementary part.

- **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. Suppose that 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. Consider 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. Denote that 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. Assume that 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. Suppose that 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 those interaction effect screening and variable selection methods under different examples.

Let T with cardinality $t = |T|$ denote the significant interaction effects in the model, i.e., $T = \{(j,k) : \beta_{j,k} \neq 0\}$. For each scenario, we run $M = 100$ Monte-Carlo simulations for each method. For the m-th simulation, denote that the estimated interaction subsets as $\widehat{T}_m$. We evaluate the performance on 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 interaction effect predictors selected by the screening methods or post-model selection method in the m-th simulation.

- The average out-of-sample R^2 for linear regression model: $R^2 = 100\% \times \left\{1 - \frac{\sum(Y_i^* - \boldsymbol{X}_i^{*T}\hat{\boldsymbol{\beta}})^2}{\sum(Y_i^* - \bar{Y}^*)^2}\right\}$, 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 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 testing data based on the training model.

5.1 Screening Performance

For the screening procedures, we consider SSI, BOLT-SSI, and the employed methods, IP and xyz for the linear model and logistic model. For the method xyz, we choose top 500 interaction terms screened by it (Actually, 500 is the largest number of interactions that the package "xyz" can be selected 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, similarly 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. Since the methods IP and xyz are not available for the logistic model, we only investigate the screening properties of SSI and BOLT-SSI for Example 5-8.

From the results shown by Tables 1 and 2, the coverage rate will decrease when the signal-noise ratio is relatively small. The proposed SSI has a high coverage percentage in screening interaction effects for different heredity structures. For the methods xyz and IP, they have a lower converge percentage except for the strong heredity setting compared to SSI. For the proposed BOLT-SSI, though its performance is not better than 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 would lose some information, and hence BOLT-SSI would not be as efficient as SSI even though its speed is much faster than SSI. Hence

Table 1: Screening results for Linear Models when $p = 5000$

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

Table 2: Screening results for Logistic Models with $n = 400$ and $p = 2000$

Methods	β_{jk}	SSI	BOLT-SSI	BOLT-SSI(p)	SSI	BOLT-SSI	BOLT-SSI(p)
			$\rho = 0$			$\rho = 0.5$	
	1	0.02	0.00	0.35	0.53	0.08	0.76
Example 5	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
	1	0.02	0.00	0.28	0.00	0.00	0.39
Example 6	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
	1	0.02	0.00	0.35	0.53	0.08	0.76
Example 7	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
	1	0.00	0.00	0.28	0.04	0.00	0.43
Example 8	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

it would increase much probability to keep 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 performances of SSI and BOLT-SSI(p) are more stable than the performance of 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 by using EBIC with $\gamma = 1$ since the EBIC tends to work the best among most of the settings as shown by Hao et al. (2018). For the method xyz, we consider the projection time L as 100, 500 and use 5-fold cross-validation (CV) to select the tuning parameter for the post-screening selection. For our methods SSI and BOLT-SSI, we use 5-folds CV and LASSO to further refine the model selection after screening. All of the simulation settings are the same as the Example 1-8 above. Especially, We set $\rho = 0.5$ for all the studies. To compare the prediction, for every simulation, we let $n_1 = 0.75 * n$ of the data set as the training data and the remaining data is considered as the testing data. Note that firstly we let p be relatively small so that it is possible to compare the performance of hierNet(Bien et al. 2013) and glinternet(Lim & Hastie 2015) in Tables 2-3 of Supplementary, 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 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. Especially, we set $\sigma = 2$ for linear models, and $\beta_{ij} = 3$ for logistic models. All results of

different methods with $(n, p) = (1000, 5000)$ are summarized in Table 3. It is shown that our method still has a good performance in the high dimensional feature space. Furthermore, we also take Examples 5 and 8 to illustrate the patterns of our method. The results are shown in Figures 1-4 in the Supplementary. Obviously, as sample size n increases, the performances of all methods become better, as shown in Figure 1 and Figure 3 in the Supplementary., and our method has the best performance. In Figures 2 and 4 in the Supplementary., though the performance of our method degrades as the dimension p increases, its performance is still much better than others. The method RAMP is influenced by the heredity assumption, especially if the anti-heredity exists, the result of RAMP is worst.

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

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 equips 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, hierNet-w, based on the 50 simulated data sets by the screening procedure and post-screening procedure, where "w" and "s" stand 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. and Table 4 summarize the average computation time (seconds per run) for each procedure. Since 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 the algorithm does not rely on the heredity assumption of the data structure.

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	5000	34.75	—	—	30.38	155.22	68.20	658.42
200	1000	1.62	—	—	3.58	18.35	6.69	53.35
400	1000	2.26	—	—	4.15	20.69	57.68	107.11
800	1000	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	1105.96	37.16	—	—	177.84	207.08
500	500	0.74	—	511.21	—	—	311.87	368.86
500	5000	27.15	—	—	—	—	127.52	1281.45
200	1000	1.10	—	—	—	—	12.34	83.98
400	1000	1.38	—	—	—	—	94.48	273.06
800	1000	2.18	—	—	—	—	588.62	820.87

In summary, compared to the other methods, our proposed SSI and BOLT-SSI(p) have a stably high coverage rate in terms of the screening performance. When the dimension of data p is not too large, by fine coding, SSI can also finish the screening task in a limited time. After discretization, some data information would be lost, and hence BOLT-SSI can not use all of the information for screening, and hence it 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 huge dimensional data in a relatively small time period. 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 to other interaction screening and variable selection methods. Especially, when computational cost becomes unaffordable for SSI, we believe that BOLT-SSI is a valuable tool for high-dimensional or ultra-high dimensional interaction screening.

6. Real Data

6.1 Supermarket Data

The supermarket data was collected from a major supermarket located in northern China and has been analyzed by Wang (2009) and Hao et al. (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 is about 2×10^7, much larger than the number of interaction effects to model the Residential Building Data (See Supplementary).

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. And the settings of all methods are the same as that of the above example. The average performance is summarized in Table 5, which includes the average sizes of main effects and interaction effects, the average out-of-

sample R^2 and their standard errors over 100 experiments. Besides 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 their papers. For the results of LASSO-AIC, LASSO-BIC, LASSO-EBIC, LASSO-GIC, we extract them from the paper of Hao et al. (2018) (RAMP). For LASSO-AIC-m, LASSO-BIC-m, we only consider the main effects. From the results in Table 5, the BOLT-SSI

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)

demonstrates the best performance, with the mean out-of-sample $R^2 = 93.95\%$. Although the products selected by BOLT-SSI are a few more, and it is a challenging task for the supermarket manager to interpret them, more products can improve the whole 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 still use the machine that equips Intel (R) Xenon(R) CPU E5-1603 v4 @ 2.80GHZ with 8.00 GB RAM. Time(s) is the average computation time of 5 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 by "RAMP" with weak

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	2252.95	33.75	NULL

heredity assumption in the above machine, the following error will appear, that is, "can not allocate vector of size 1.1 Gb", 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, at the first step of our screening methods, we only use marginal information of the data or even sacrifice some information for the method BOLT-SSI. However, the advantages of computational efficiency are much evident, and especially for BOLT-SSI, the sacrifice of the data information can be ignored, which is consistent with our theoretical investigation.

7. Conclusion and Discussion

In this paper, we study the screening method to detect important significant interaction effects in the generalized high dimensional linear model. A new and straightforward procedure SSI and its extension BOLT-SSI are proposed. Different from most of the other screening or variable selection methods for the interaction effects detecting, 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 ultra-high dimensional data, in some sense, such a task seems to be impossible to be completed. Here we show that, by taking advantage of computational structure, seemly impossible tasks can be done using a standard personal computer. Importantly, the statistical property of the proposed way is guaranteed by our established theory.

Our numerical studies only consider 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 under a limited time and limited computing resource, we should

also screen the main effects and interaction effects simultaneously.

Generally speaking, most of the data analysis projects are similar to engineer projects. Though most of the theoretical research would be beneficial to projects, the requirement and expectation of the engineering projects are different from those of theoretical studies. How to combine the advantages of engineering techniques to complete those projects under the requirement and expectation of practice needs further investigation. Our study here attempts to pursuit such a direction by a small step.

Acknowledgments

The authors thank the Associate Editor and two anonymous referees for their helpful comments. This work was supported in part by the Hong Kong Research Grant Council[16307818, 16301419, 16308120, 12303618], the RGC Collaborative Research Fund: C6021-19EF, 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, AcRF Tier 2(MOE2018-T2-1-046, MOE2018-T2-2-006) from the Ministry of Education, Singapore.

SUPPLEMENTARY MATERIAL

Due to space constraints, all of the discussion and the sure screening properties of SSI and their proofs are relegated to Section 1 and 2 in the supplementary material. In addition to these proofs, Section 2 includes the proofs of Theorem 1 and 2 about sure screening porperties of BOLT-SSI. Part of the simulation about the data set with small dimension and dicussion about how to choose between SSI and BOLT-SSI can also be found in Section 3. Furthermore, two more case studies are illustrated in Section 4. One of them is huge dimensional data, whose dimension is $p = 319,156$. The total number of interaction terms is about 5×10^{10}. Moreover, R-package "BOLTSSIRR" contains the codes to perform the algorithms decribed in the article.

1. Beijing Normal University–Hong Kong Baptist University United International College

E-mail: (minzhou@uic.edu.cn)

2. Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics

E-mail: (daimw@swufe.edu.cn)

3. Victoria University of Wellington, School of Mathematics and Statistics,

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

4. Duke-NUS Graduate Medical School

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

5. The Hong Kong University of Science and Technology

E-mail: (macyang@ust.hk)

6. Hong Kong Baptist University

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

References

Agresti, A. & Kateri, M. (2011), *Categorical data analysis*, Springer.

Barut, E., Fan, J. & Verhasselt, A. (2016), 'Conditional sure independence screening', *Journal of the American Statistical Association* **111**(515), 1266–1277.

Bateson, W. (1909), 'Mendel's principles of heredity'.

Bien, J., Taylor, J. & Tibshirani, R. (2013), 'A lasso for hierarchical interactions', *Annals of statistics* **41**(3), 1111.

Chandrasekaran, V. & Jordan, M. I. (2013), 'Computational and statistical tradeoffs via convex relaxation', *Proceedings of the National Academy of Sciences* **110**(13), E1181–E1190.

Chang, J., Tang, C. Y. & Wu, Y. (2013), 'Marginal empirical likelihood and sure independence feature screening', *Annals of statistics* **41**(4).

Chen, R.-B., Weng, J.-Z. & Chu, C.-H. (2013), 'Screening procedure for supersaturated designs using a bayesian variable selection method', *Quality and Reliability Engineering International* **29**(1), 89–101.

Choi, N. H., Li, W. & Zhu, J. (2010), 'Variable selection with the strong heredity constraint and its oracle property', *Journal of the American Statistical Association* **105**(489), 354–364.

Cordell, H. J. (2009), 'Detecting gene–gene interactions that underlie human diseases', *Nature Reviews Genetics* **10**(6), 392.

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

Culverhouse, R., Suarez, B. K., Lin, J. & Reich, T. (2002), 'A perspective on epistasis: limits of models displaying no main effect', *The American Journal of Human Genetics* **70**(2), 461–471.

Deming, W. E. & 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**(4), 427–444.

Esscher, F. (1924), 'On a method of determining correlation from the ranks of the variates', *Scandinavian Actuarial Journal* **1924**(1), 201–219.

Fan, J., Feng, Y. & Song, R. (2011), 'Nonparametric independence screening in sparse ultra-high-dimensional additive models', *Journal of the American Statistical Association* **106**(494), 544–557.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American statistical Association* **96**(456), 1348–1360.

Fan, J. & Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

Fan, J. & Peng, H. (2004), 'Nonconcave penalized likelihood with a diverging number of parameters', *The Annals of Statistics* **32**(3), 928–961.

Fan, J., Samworth, R. & Wu, Y. (2009), 'Ultrahigh dimensional feature selection: beyond the linear model', *Journal of machine learning research* **10**(Sep), 2013–2038.

Fan, J. & Song, R. (2010), 'Sure independence screening in generalized linear models with np-dimensionality', *The Annals of Statistics* **38**(6), 3567–3604.

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

Fan, Y., Kong, Y., Li, D. & Zheng, Z. (2015), 'Innovated interaction screening for high-dimensional nonlinear classification', *The Annals of Statistics* **43**(3), 1243–1272.

Fienberg, S. E. et al. (1970), 'An iterative procedure for estimation in contingency tables', *The Annals of Mathematical Statistics* **41**(3), 907–917.

Fisher, R. (1918), 'The correlation between relatives on the supposition of mendelian inheritance', *Trans. Royal Soc. Edin* pp. 399–433.

Haberman, S. (1974), 'The analysis of frequency data'.

Hao, N., Feng, Y. & Zhang, H. H. (2018), 'Model selection for high-dimensional quadratic regression via regularization', *Journal of the American Statistical Association* **113**(522), 615–625.

Hao, N. & Zhang, H. H. (2014), 'Interaction screening for ultrahigh-dimensional data', *Journal of the American Statistical Association* **109**(507), 1285–1301.

Hotelling, H. (1953), 'New light on the correlation coefficient and its transforms', *Journal of the Royal Statistical Society. Series B (Methodological)* **15**(2), 193–232.

Jaccard, J., Wan, C. K. & Turrisi, R. (1990), 'The detection and interpretation of interaction effects between continuous variables in multiple regression', *Multivariate behavioral research* **25**(4), 467–478.

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

Kirkwood, J. G. (1935), 'Statistical mechanics of fluid mixtures', *The Journal of Chemical Physics* **3**(5), 300–313.

Kong, Y., Li, D., Fan, Y., Lv, J. et al. (2017), 'Interaction pursuit in high-dimensional multi-response regression via distance correlation', *The Annals of Statistics* **45**(2), 897–922.

Lee, A. J. (2019), *U-statistics: Theory and Practice*, Routledge.

Lees, P., Cunningham, F. & Elliott, J. (2004), 'Principles of pharmacodynamics and their applications in veterinary pharmacology', *Journal of veterinary pharmacology and therapeutics* **27**(6), 397–414.

Li, D., Kong, Y., Fan, Y. & Lv, J. (2021), 'High-dimensional interaction detection with false sign rate control', *Journal of Business & Economic Statistics* pp. 1–12.

Li, G., Peng, H., Zhang, J., Zhu, L. et al. (2012), 'Robust rank correlation based screening', *The Annals of Statistics* **40**(3), 1846–1877.

Li, Y. & Liu, J. S. (2019), 'Robust variable and interaction selection for logistic regression and general index models', *Journal of the American Statistical Association* **114**(525), 271–286.

Lim, M. & Hastie, T. (2015), 'Learning interactions via hierarchical group-lasso regularization', *Journal of Computational and Graphical Statistics* **24**(3), 627–654.

Liu, H., Hussain, F., Tan, C. L. & Dash, M. (2002), 'Discretization: An enabling technique', *Data mining and knowledge discovery* **6**(4), 393–423.

Pan, W., Wang, X., Xiao, W. & 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. & Moore, J. H. (2001), 'Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer', *The American Journal of Human Genetics* **69**(1), 138–147.

Saldana, D. F. & Feng, Y. (2018), 'Sis: An r package for sure independence screening in ultrahigh-dimensional statistical models', *Journal of Statistical Software* **83**(2), 1–25.

Shah, R. D. (2016), 'Modelling interactions in high-dimensional data with backtracking', *Journal of Machine Learning Research* **17**(207), 1–31.

She, Y. & Tang, S. (2019), 'Iterative proportional scaling revisited: a modern optimization perspective', *Journal of Computational and Graphical Statistics* **28**(1), 48–60.

She, Y., Wang, Z. & Jiang, H. (2018), 'Group regularized estimation under structural hierarchy', *Journal of the American Statistical Association* **113**(521), 445–454.

Tang, C. Y., Fang, E. X. & Dong, Y. (2020), 'High-dimensional interactions detection with sparse principal hessian matrix.', *J. Mach. Learn. Res.* **21**, 19–1.

Thanei, G.-A., Meinshausen, N. & Shah, R. D. (2018), 'The xyz algorithm for fast interaction search in high-dimensional data', *The Journal of Machine Learning Research* **19**(1), 1343–1384.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. & Yu, W. (2010), 'Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies', *The American Journal of Human Genetics* **87**(3), 325–340.

Wang, H. (2009), 'Forward regression for ultra-high dimensional variable screening', *Journal of the American Statistical Association* **104**(488), 1512–1524.

Wang, J.-H. & Chen, Y.-H. (2018), 'Overlapping group screening for detection of gene-gene interactions: application to gene expression profiles with survival trait', *BMC bioinformatics* **19**(1), 335.

Wang, J.-H. & Chen, Y.-H. (2020), 'Interaction screening by kendall's partial correlation for ultrahigh-dimensional data with survival trait', *Bioinformatics* **36**(9), 2763–2769.