The Annals of Statistics
2015, Vol. 43, No. 3, 1243–1272
DOI: 10.1214/14-AOS1308

INNOVATED INTERACTION SCREENING FOR HIGH-DIMENSIONAL NONLINEAR CLASSIFICATION[1]

BY YINGYING FAN, YINFEI KONG, DAOJI LI AND ZEMIN ZHENG

University of Southern California

This paper is concerned with the problems of interaction screening and nonlinear classification in a high-dimensional setting. We propose a two-step procedure, IIS-SQDA, where in the first step an innovated interaction screening (IIS) approach based on transforming the original p-dimensional feature vector is proposed, and in the second step a sparse quadratic discriminant analysis (SQDA) is proposed for further selecting important interactions and main effects and simultaneously conducting classification. Our IIS approach screens important interactions by examining only p features instead of all two-way interactions of order $O(p^2)$. Our theory shows that the proposed method enjoys sure screening property in interaction selection in the high-dimensional setting of p growing exponentially with the sample size. In the selection and classification step, we establish a sparse inequality on the estimated coefficient vector for QDA and prove that the classification error of our procedure can be upper-bounded by the oracle classification error plus some smaller order term. Extensive simulation studies and real data analysis show that our proposal compares favorably with existing methods in interaction selection and high-dimensional classification.

1. Introduction. Classification, aiming at identifying to which of a set of categories a new observation belongs, has been frequently encountered in various fields such as genomics, proteomics, face recognition, brain images, medicine and machine learning. In recent years, there has been a significant surge of interest in interaction selection in classification due to the importance of interactions in statistical inference and contemporary scientific discoveries. For instance, in genome-wide association studies, it has been increasingly recognized that gene–gene interactions and gene–environment interactions substantially influence the risk of developing a human disease [18]. Ignoring these interactions could potentially lead to misunderstanding about disease mechanisms as they are potential sources of the missing heritability [21].

Identification of interactions is challenging even when the number of predictors p is moderately large compared to the sample size n, as the number of all possible pairwise interaction effects is of order $O(p^2)$. This problem becomes even

Received October 2014.

[1]Supported by NSF CAREER Award DMS-11-50318 and USC Marshall Summer Research Funding.

MSC2010 subject classifications. Primary 62H30; secondary 62F05, 62J12.

Key words and phrases. Classification, dimension reduction, discriminant analysis, interaction screening, sparsity, sure screening property.

more challenging in the high-dimensional setting where p can be much larger than n. It is well known that the classical low-dimensional classification method cannot be directly used for high-dimensional classification for at least three reasons. First, many popular classifiers, such as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), are inapplicable when p exceeds n because of the singularities of the sample covariance matrices. Second, when p is large, it is commonly believed that only a subset of the p features contribute to classification. Classification using all potential features may cause difficulty in interpretation and degrade the classification performance due to the noise accumulation in estimating a large number of parameters [7]. Third, the computational cost may be extremely high when the dimensionality is ultra-high. For example, with $p = 1000$ features, the dimensionality is about half million if all possible pairwise interactions are included in classification.

In recent years, significant efforts have been made to develop effective high-dimensional classification methods. The most commonly imposed assumption is sparsity, leading to sparse classifiers. Tibshirani et al. [25] introduced the nearest shrunken centroids classifier, and Fan and Fan [7] proposed features annealed independent rules, both of which ignore correlations among features to reduce the dimensionality of parameters. Shao et al. [23] proposed and studied a sparse LDA method, which directly plugs the sparse estimates of the covariance matrix and mean vector into the linear classifier. Cai and Liu [4] introduced a direct approach to sparse LDA by estimating the product of the precision matrix and the mean difference vector of two classes, through constrained L_1 minimization. In an independent work, Mai et al. [20] also proposed a direct approach to sparse LDA, called DSDA, by reformulating the LDA problem as a penalized least squares regression. Fan et al. [9] considered HCT classifier for high-dimensional Gaussian classification with sparse precision matrix when the signals are rare and weak, and studied its optimality. A commonality of these aforementioned methods is that the underlying true classifier is assumed to be linear, and thus they belong to the class of sparse LDA methods.

A key assumption for LDA is that observations from different classes share the same correlation structure. Although this assumption can significantly reduce the number of parameters need to be estimated, it can be easily violated in real applications. In addition, linear classifiers are not capable of identifying important interaction effects between features and thus can lead to inferior feature selection and classification results, and consequently, misleading interpretations when the classification boundary is nonlinear. For instance, in a two-class Gaussian classification problem, when two classes have equal mean vectors but different covariance matrices, linear classifiers can perform no better than random guessing.

To gain some insight into the importance of interactions in classification, let us look at a simple example. Consider a two-class Gaussian classification problem

TABLE 1
The means and standard errors (in parentheses) of various performance measures for different classification methods over 100 *replications where the Bayes rule is given in* (1). *Sample size in each class is* 100, *and the number of features* p *is* 200

Measure	PLR	DSDA	IIS-SQDA	Oracle
MR (%)	49.95 (0.05)	49.87 (0.09)	26.03 (0.31)	23.99 (0.08)
FP.main	49.66 (4.93)	74.29 (6.45)	3.16 (0.82)	0 (0)
FP.inter	–	–	0.55 (0.14)	0 (0)
FN.inter	–	–	0.15 (0.05)	0 (0)

with the Bayes rule

$$Q(\mathbf{z}) = -0.3Z_{10}^2 - 0.15Z_{10}Z_{30} - 0.15Z_{10}Z_{50} - 0.3Z_{30}^2 - 0.15Z_{30}Z_{50} - 0.3Z_{50}^2 + 1.74913, \tag{1}$$

which classifies a new observation $\mathbf{z}$ to class 1 if and only if $Q(\mathbf{z}) > 0$. Thus there are no main effects, and there are three variables, Z_{10}, Z_{30} and Z_{50}, contributing to interactions. We simulated data in the same way as model 2 in Section 5.2.2, except that the mean vector in each class is zero. See Section 5.2.2 for more details. Table 1 lists the performance of different classification methods, including penalized logistic regression (PLR), DSDA, our proposal (IIS-SQDA) and the oracle procedure (Oracle). The oracle procedure uses the information of the true underlying sparse model and thus is a low-dimensional QDA. As expected, both linear classifiers, PLR and DSDA, perform no better than random guessing. Table 1 also shows the variable selection results for main effects and interactions, with FP.main standing for false positives of main effects, and FP.inter and FN.inter standing for false positives and false negatives of interaction effects, respectively. It is seen that with appropriate selection of interaction effects, the classification performance can be improved significantly.

In this paper we consider two-class classification with possibly unequal covariance matrices. Under some sparsity assumption on the main effects and interactions, we propose a two-stage classification procedure, where we first reduce the number of interactions to a moderate order by a new interaction screening approach, and then identify both important main effects and interactions using some variable selection techniques. Our interaction screening approach is motivated by a result, which will be formally demonstrated in our paper, that if an interaction term, say Z_1Z_2, appears in Bayes decision rule, then after appropriately transforming the original features, the resulting new feature $\widetilde{Z}_1$ (and $\widetilde{Z}_2$) has different variances across classes. Thus the original problem of screening $O(p^2)$ pairwise interaction effects can be recast as the problem of comparing variances of only p variables, which can be solved by some variance test procedures such as the F-test or the SIRI method proposed in [16]. The similar idea of interaction screening has

also been considered in [16] under the model setting of sliced inverse index model. Hereafter, we refer to Z_i as an interaction variable if an interaction term involving Z_i appears in Bayes rule. After obtaining interaction variables in the first step, we reconstruct interaction terms based on these screened interaction variables, and then use recent advances in variable selection literature to further select important ones from the pool of all main effects and reconstructed interactions. Under some mild conditions, we prove that with overwhelming probability, all active interaction variables will be retained using our screening procedure. For the second step of selection and classification, we first establish a sparse inequality [27], which shows the consistency of the estimated coefficient vector of QDA, then further prove that the classification error of IIS-SQDA is upper-bounded by the oracle classification error plus a smaller order term. Our numerical studies demonstrate the fine performance of the proposed method for interaction screening and high-dimensional classification.

The main contributions of this paper are as follows. First, we introduce an interaction screening approach, which has been proved to enjoy sure screening property. Second, our classification method does not rely on the linearity assumption, which makes our method more applicable in real applications. Third, our proposed classification procedure is adaptive in the sense that it automatically chooses between sparse LDA and sparse QDA. If the index set of screened interaction variables is empty in the first step, or if the index set in the first step is nonempty but none of the interaction terms is selected in the second step, then sparse LDA will be used for classification; otherwise, sparse QDA will be used for classification. Fourth, we provide theoretical justifications on the effectiveness of the proposed procedure.

The remaining part of the paper will unfold as follows. Section 2 introduces the model setting and motivation. Section 3 proposes the innovated interaction screening approach and studies its theoretical property. Section 4 considers post-screening variable selection. Section 5 presents the results of extensive simulation studies and a real data example. Section 6 concludes with some discussion. Section 7 collects all proofs for the main theorems. Additional proofs are provided in the supplementary material [10].

2. Model setting and motivation. Our interaction screening approach is motivated from the problem of two-class Gaussian classification, where the p-dimensional feature vector $\mathbf{z} = (Z_1, \ldots, Z_p)^T$ follows a mixture distribution

$$\mathbf{z} = \Delta \mathbf{z}^{(1)} + (1 - \Delta)\mathbf{z}^{(2)} \tag{2}$$

with $\mathbf{z}^{(k)}$ a Gaussian random vector with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, $k = 1, 2$, and the class label Δ following a Bernoulli distribution with probability of success π. Without loss of generality, assume that $\boldsymbol{\mu}_2 = 0$. Under this model setting, the Bayes rule admits the following form:

$$Q(\mathbf{z}) = \tfrac{1}{2}\mathbf{z}^T \boldsymbol{\Omega}\mathbf{z} + \boldsymbol{\delta}^T \mathbf{z} + \zeta, \tag{3}$$

where $\mathbf{\Omega} = \mathbf{\Sigma}_2^{-1} - \mathbf{\Sigma}_1^{-1}$, $\boldsymbol{\delta} = \mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1$ and ζ is some constant depending only on π, $\boldsymbol{\mu}_1$ and $\mathbf{\Sigma}_k$, $k = 1, 2$. A new observation $\mathbf{z}$ is classified into class 1 if and only if $Q(\mathbf{z}) > 0$.

When covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are the same, the above Bayes rule takes the linear form $Q(\mathbf{z}) = \boldsymbol{\delta}^T\mathbf{z} + \zeta$, which is frequently referred to as the Fisher's LDA and belongs to the family of linear classifiers. As discussed in the Introduction, linear classifiers may be inefficient or even fail when the true classification boundary is nonlinear. Moreover, linear classifiers are incapable of selecting important interaction terms when the covariance matrices are different across two classes. For the ease of presentation, hereafter we mean interaction in the broad sense of the term, not just the two-way interactions Z_jZ_ℓ with $j \neq \ell$, but also the quadratic terms Z_j^2. So there are $p(p+1)/2$ possible interactions in total under our definition. Throughout this paper we call Z_jZ_ℓ, $1 \leq j, \ell \leq p$ an active interaction if its coefficient is nonzero in (3), and we call Z_j an interaction variable if there exists some $\ell \in \{1, 2, \ldots, p\}$ such that Z_jZ_ℓ is an active interaction. Selecting important ones from the large number of interactions is interesting yet challenging. We next discuss our proposal for interaction screening.

From (3), one can observe that an interaction term Z_jZ_ℓ is an active interaction if and only if $\mathbf{\Omega}_{j\ell} \neq 0$. Here we use $\mathbf{A}_{j\ell}$ to denote the (j, ℓ) element of any matrix $\mathbf{A}$. This observation motivates us to select active interactions by recovering the support of $\mathbf{\Omega}$. Denote the index set of interaction variables by

$$\mathcal{I} = \{1 \leq j \leq p : Z_jZ_\ell \text{ is an active interaction for some } 1 \leq \ell \leq p\}. \tag{4}$$

In light of (3), the above set can also be written as $\mathcal{I} = \{1 \leq j \leq p : \mathbf{\Omega}_{j\ell} \neq 0 \text{ for some } 1 \leq \ell \leq p\}$. If the index set $\mathcal{I}$ can be recovered, then all active interactions can be reconstructed. For this reason, we aim at developing an effective method for screening the index set $\mathcal{I}$.

In a high-dimensional setting, to ensure the model identifiability and to enhance the model fitting accuracy and interpretability, it is commonly assumed that only a small number of interactions contribute to classification. Thus we impose the sparsity assumption that there are only a small number of active interactions. Equivalently, we can assume that $\mathbf{\Omega}$ is highly sparse with only $q = o(\min\{n, p\})$ rows (and columns, by symmetry) being nonzero, where n is the total sample size. Denote $\mathbf{\Sigma}_k^{-1}$ by $\mathbf{\Omega}_k$, $k = 1, 2$. Without loss of generality, we write $\mathbf{\Omega}$ as

$$\mathbf{\Omega} = \mathbf{\Omega}_2 - \mathbf{\Omega}_1 = \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{0} \end{pmatrix}, \tag{5}$$

where $\mathbf{B}$ is a $q \times q$ symmetric matrix with at least one nonzero element in each row. We remark that the block structure in (5) is just for the simplicity of presentation, and we do not require that the locations of nonzero rows of $\mathbf{\Omega}$ are known. In fact, we will develop a method to estimate the indices of these nonzero rows. Note that the set $\mathcal{I}$ can be further written as

$$\mathcal{I} = \{1 \leq j \leq q : \mathbf{B}_{j\ell} \neq 0 \text{ for some } 1 \leq \ell \leq q\}.$$

Thus interaction screening is equivalent to finding the indices of features related to $\mathbf{B}$.

Identifying the index set $\mathcal{I}$ is challenging when p is large. We overcome this difficulty by decomposing $\mathcal{I}$ into two subsets. Let

$$\mathcal{I}_1 = \{j \in \mathcal{I} \text{ and } \mathbf{B}_{jj} \le 0\}, \qquad \mathcal{I}_2 = \{j \in \mathcal{I} \text{ and } \mathbf{B}_{jj} > 0\}.$$

Then $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$. This allows us to estimate $\mathcal{I}$ by dealing with $\mathcal{I}_1$ and $\mathcal{I}_2$ separately.

First consider $\mathcal{I}_1$. Our main idea is to use the transformation $\tilde{\mathbf{z}} = \mathbf{\Omega}_1 \mathbf{z}$. Denote by $\tilde{\mathbf{z}}^{(k)} = \mathbf{\Omega}_1 \mathbf{z}^{(k)}$ the transformed feature vector from class k with $k = 1, 2$. Then $\operatorname{cov}(\tilde{\mathbf{z}}^{(1)}) = \mathbf{\Omega}_1$ and $\operatorname{cov}(\tilde{\mathbf{z}}^{(2)}) = \mathbf{\Omega}_1 \mathbf{\Sigma}_2 \mathbf{\Omega}_1$. It follows from linear algebra that the difference of the above two covariance matrices takes the following form:

$$\widetilde{\mathbf{\Sigma}}_1 \equiv \mathbf{\Omega}_1 \mathbf{\Sigma}_2 \mathbf{\Omega}_1 - \mathbf{\Omega}_1 = \mathbf{\Omega} \mathbf{\Sigma}_2 \mathbf{\Omega} - \mathbf{\Omega} = \begin{pmatrix} \mathbf{B}\mathbf{\Sigma}_2^{(11)}\mathbf{B} - \mathbf{B} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{0} \end{pmatrix}, \tag{6}$$

where $\mathbf{\Sigma}_2^{(11)}$ is the $q \times q$ principal submatrix of $\mathbf{\Sigma}_2$ corresponding to matrix $\mathbf{B}$. We will show that if $j \in \mathcal{I}_1$, then the jth entry in transformed feature vector $\tilde{\mathbf{z}}$ has different variances across two classes. To this end, let $\mathbf{e}_j$ be a unit vector with jth component 1 and all other components 0. Then it follows from the positive definiteness of $\mathbf{\Sigma}_2^{(11)}$ that $(\mathbf{B}\mathbf{e}_j)^T \mathbf{\Sigma}_2^{(11)} (\mathbf{B}\mathbf{e}_j)$ is positive for any $j \in \mathcal{I}_1$. Since $\mathbf{B}_{jj} \le 0$ for any $j \in \mathcal{I}_1$, the jth diagonal element of $\widetilde{\mathbf{\Sigma}}_1$ is positive by noting that

$$(\widetilde{\mathbf{\Sigma}}_1)_{jj} = (\mathbf{B}\mathbf{e}_j)^T \mathbf{\Sigma}_2^{(11)} (\mathbf{B}\mathbf{e}_j) - \mathbf{B}_{jj}. \tag{7}$$

This gives a set inclusion

$$\mathcal{I}_1 \subset \mathcal{A}_1 \equiv \{j : (\widetilde{\mathbf{\Sigma}}_1)_{jj} \neq 0\}. \tag{8}$$

Observing that $(\widetilde{\mathbf{\Sigma}}_1)_{jj}$ is the difference of between-class variances of the jth transformed variable, that is,

$$(\widetilde{\mathbf{\Sigma}}_1)_{jj} = \operatorname{var}(\mathbf{e}_j^T \tilde{\mathbf{z}}^{(2)}) - \operatorname{var}(\mathbf{e}_j^T \tilde{\mathbf{z}}^{(1)}), \tag{9}$$

the index set $\mathcal{A}_1$ can be obtained by examining which features have different variances across two classes after the transformation.

We further remark that the variance difference between $\mathbf{e}_j^T \tilde{\mathbf{z}}^{(2)}$ and $\mathbf{e}_j^T \tilde{\mathbf{z}}^{(1)}$ records the accumulated contributions of the jth feature to the interaction. To understand this, note that if $\mathbf{\Sigma}_2^{(11)}$ has the smallest eigenvalue bounded from below by a positive constant τ_1, then (7) and (9) together ensure that

$$\operatorname{var}(\mathbf{e}_j^T \tilde{\mathbf{z}}^{(2)}) - \operatorname{var}(\mathbf{e}_j^T \tilde{\mathbf{z}}^{(1)}) \ge \tau_1 \|\mathbf{B}\mathbf{e}_j\|_2^2 - \mathbf{B}_{jj},$$

where $\|\cdot\|_2$ denotes the L_2 norm of a vector. In view of (3) and (5), the jth column (and row, by symmetry) of $\mathbf{B}$ records all contributions of the jth feature to interactions. Thus the more important the jth feature is to interaction, the larger the variance difference.

Similarly, consider the transformation $\check{\mathbf{z}} = \mathbf{\Omega}_2\mathbf{z}$, and define the matrix $\widetilde{\mathbf{\Sigma}}_2 = \mathbf{\Omega}_2 - \mathbf{\Omega}_2\mathbf{\Sigma}_1\mathbf{\Omega}_2$. Then $\widetilde{\mathbf{\Sigma}}_2$ is the difference between the covariance matrices of transformed feature vectors $\check{\mathbf{z}}^{(1)} = \mathbf{\Omega}_2\mathbf{z}^{(1)}$ and $\check{\mathbf{z}}^{(2)} = \mathbf{\Omega}_2\mathbf{z}^{(2)}$. Using arguments similar to those in (8), we get another set inclusion

$$\mathcal{I}_2 \subset \mathcal{A}_2 \equiv \{j : (\widetilde{\mathbf{\Sigma}}_2)_{jj} \neq 0\}. \tag{10}$$

Similarly, the set $\mathcal{A}_2$ can be obtained by examining which features have different variances across two classes after the transformation based on $\mathbf{\Omega}_2$.

Combining (8) and (10) leads to

$$\mathcal{I} \subset \mathcal{A}_1 \cup \mathcal{A}_2. \tag{11}$$

Meanwhile, by (6) we have $\mathcal{A}_1 \subset \mathcal{I}$. Similarly, we obtain $\mathcal{A}_2 \subset \mathcal{I}$. Combining these results with the set inclusion (11) ensures that

$$\mathcal{I} = \mathcal{A}_1 \cup \mathcal{A}_2. \tag{12}$$

This motivates us to find interaction variables by testing variances of the transformed feature vectors $\widetilde{\mathbf{z}}$ and $\check{\mathbf{z}}$ across two classes. Since the transformation based on precision matrix $\mathbf{\Omega}_k$ is called innovation in the time series literature, we name our method the *innovated interaction screening* (IIS).

The innovated transform has also been explored in other papers. For example, Hall and Jin [14] proposed the innovated higher criticism based on the innovated transform on the original feature vector to detect sparse and weak signals when the noise variables are correlated, and established an upper bound to the detection boundary. In two-class Gaussian linear classification setting, Fan et al. [9] discussed in detail the advantage of innovated transform. They showed that the innovated transform is best at boosting the signal-to-noise ratio in their model setting. Detailed discussions about innovated transform in the multiple testing context can be found in [17].

3. Sampling property of innovated interaction screening.

3.1. *Technical assumptions.* We study the sampling properties of IIS procedure in this section. In our theoretical development, the Gaussian distribution assumption in Section 2 will be relaxed to sub-Gaussian, but implicitly, we still assume that our target classifier takes the form (3). A random vector $\mathbf{w} = (W_1, \ldots, W_p)^T \in \mathbb{R}^p$ is sub-Gaussian if there exist some positive constants a and b such that $P(|\mathbf{v}^T\mathbf{w}| > t) \leq a\exp(-bt^2)$ for any $t > 0$ and any vector $\mathbf{v} \in \mathbb{R}^p$ satisfying $\|\mathbf{v}\|_2 = 1$. The following conditions will be needed for our theoretical development:

CONDITION 1 (Sub-Gaussian). *Both* $\mathbf{z}^{(1)}$ *and* $\mathbf{z}^{(2)}$ *are sub-Gaussian.*

CONDITION 2 (Bounds of eigenvalues). *There exists some positive constant τ_1 and some positive sequence $\tau_{2,p}$ depending only on p such that the eigenvalues of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ satisfy*

$$\tau_1 \le \lambda_{\min}(\mathbf{\Sigma}_k) \le \lambda_{\max}(\mathbf{\Sigma}_k) \le \tau_{2,p} \qquad \text{for } k = 1, 2,$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a matrix, respectively.

CONDITION 3 (Distinguishability). *Denote by σ_j^2, $(\sigma_j^{(1)})^2$ and $(\sigma_j^{(2)})^2$ the population variances of the jth covariates in $\tilde{\mathbf{z}}$, $\tilde{\mathbf{z}}^{(1)}$ and $\tilde{\mathbf{z}}^{(2)}$, respectively. There exist some positive constants κ and c such that for any $j \in \mathcal{A}_1$ with $\mathcal{A}_1$ defined in (8), it holds that*

$$\frac{\sigma_j^2}{(\sigma_j^{(1)})^{2\pi}(\sigma_j^{(2)})^{2(1-\pi)}} \ge \exp\bigl(3cn^{-\kappa}\bigr). \tag{13}$$

Moreover, the same inequality also holds for the jth covariates in $\check{\mathbf{z}}$, $\check{\mathbf{z}}^{(1)}$ and $\check{\mathbf{z}}^{(2)}$ when $j \in \mathcal{A}_2$ with $\mathcal{A}_2$ defined in (10).

CONDITION 4 (K_p-sparsity). *For each $k = 1, 2$, the precision matrix $\mathbf{\Omega}_k$ is K_p-sparse, where a matrix is said to be K_p-sparse if each of its row has at most K_p nonzero components with K_p a positive integer depending only on p. Moreover, $\|\mathbf{\Omega}_k\|_{\max}$ is bounded from above by some positive constant independent of p, where $\|\cdot\|_{\max}$ is the elementwise infinity norm of a matrix.*

Condition 1 is used to control the tail behavior of the covariates. Gaussian distribution and distributions with bounded support are two special examples of sub-Gaussian distribution.

Condition 2 imposes conditions on the eigenvalues of the population covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. The lower bound τ_1 is a constant while the upper bound can slowly diverge to infinity with p. So the condition numbers of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ can diverge with p as well. We remark that we need a constant lower bound τ to exclude the case of perfect or nearly perfect collinearity of features at the population level. On the technical side, the constant lower bound τ_1 ensures that $\tilde{\mathbf{z}}^{(k)}$ and $\check{\mathbf{z}}^{(k)}$ are still sub-Gaussian after transformation.

Condition 3 is a signal strength condition which assumes that for any $j \in \mathcal{A}_1$, the population variances of the jth transformed feature $\widetilde{Z}_j$ are different enough across classes, by noting that

$$D_j \equiv \log\sigma_j^2 - \sum_{k=1}^{2} \pi_k \log\bigl[(\sigma_j^{(k)})^2\bigr] \ge 3cn^{-\kappa} \tag{14}$$

with $\pi_1 = \pi$ and $\pi_2 = 1 - \pi$ when $j \in \mathcal{A}_1$. Meanwhile, it is clear from the definition of $\mathcal{A}_1$ that D_j is exactly 0 when $j \in \mathcal{A}_1^c$ since the population variances of the

jth transformed covariate $\widetilde{Z}_j$ are the same across classes. The same results hold for any feature with index in $\mathcal{A}_2$, after transforming the data using $\mathbf{\Omega}_2$, based on the second part of this condition.

Condition 4 is on the sparsity of the precision matrices, which is needed for ensuring the estimation accuracy of precision matrices. The same family of precision matrices has also been considered in [9] for high-dimensional linear classification. Condition 4 also imposes a uniform upper bound for all components of $\mathbf{\Omega}_k$. We note that we use this assumption merely to simplify the proof, and our main results will still hold with a slightly more complicated form when the upper bound diverges slowly with the number of predictors p.

3.2. *Oracle-assisted IIS.* In this subsection, we consider IIS with known precision matrices, which we call the oracle-assisted IIS. The case of unknown precision matrices will be studied in the next subsection. The results developed here are mainly of theoretical interests and will serve as a benchmark for the performance of IIS with unknown precision matrices.

As introduced in Section 2, IIS works with the transformed feature vectors $\tilde{\mathbf{z}} = \mathbf{\Omega}_1\mathbf{z}$ and $\check{\mathbf{z}} = \mathbf{\Omega}_2\mathbf{z}$ identically. For the ease of presentation we only discuss in detail IIS based on the transformation $\tilde{\mathbf{z}} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_p)^T = \mathbf{\Omega}_1\mathbf{z}$.

Suppose we observe n data points $\{(\mathbf{z}_i^T, \Delta_i), i = 1, \ldots, n\}$, n_k of which are from class k for $k = 1, 2$. Write $\widetilde{\mathbf{Z}} = \mathbf{Z}\mathbf{\Omega}_1$ as the transformed data matrix, where $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$ is the original data matrix. To test whether the jth transformed feature $\widetilde{Z}_j$ has different variances across two classes, we propose to use the following test statistic introduced in [16]:

$$\widetilde{D}_j = \log \tilde{\sigma}_j^2 - \sum_{k=1}^{2} (n_k/n) \log\big[(\tilde{\sigma}_j^{(k)})^2\big], \tag{15}$$

where $\tilde{\sigma}_j^2$ denotes the pooled sample variance estimate for $\widetilde{Z}_j$, and $(\tilde{\sigma}_j^{(k)})^2$ is the within-class sample variance estimate for $\widetilde{Z}_j$ in class k. As can be seen from (15), $\widetilde{D}_j$ is expected to be nonzero if variances of $\widetilde{Z}_j$ are different across classes. This test statistic was originally introduced in [16] in the sliced inverse index model setting for detecting important variables with pairwise or higher-order interactions among p predictors. The aforementioned paper recommends the use of $\widetilde{D}_j$ in the initial screening step of their proposed procedure, and proves the sure screening property of it under some regularity conditions.

Denote by $\tilde{\tau}_{1,p} = \min\{\pi\tau_{2,p}^{-1} + (1-\pi)\tau_1\tau_{2,p}^{-2}, 1\}$ and $\tilde{\tau}_{2,p} = \max\{\pi\tau_1^{-1} + (1-\pi)\tau_1^{-2}\tau_{2,p} + \pi(1-\pi)\tau_1^{-2}\|\boldsymbol{\mu}_1\|_2^2, \exp(1)\}$. The following proposition shows that the oracle-assisted IIS enjoys the sure screening property in interaction selection under our model setting.

PROPOSITION 1. *Assume that Conditions 1–3 hold. If* $\log p = O(n^\gamma)$ *with* $\gamma > 0$ *and* $\gamma + 2\kappa < 1$, *and* $\tilde{\tau}_{1,p}^{-2} + \log^2(\tilde{\tau}_{2,p}) = o(n^{1-2\kappa-\gamma})$, *then with probability*

at least $1-\exp\{-Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2}+\log^2(\tilde{\tau}_{2,p})]\}$ *for some positive constant* C, *it holds that*

$$\min_{j\in\mathcal{A}_1}\widetilde{D}_j \geq 2cn^{-\kappa} \quad \textit{and} \quad \max_{j\in\mathcal{A}_1^c}\widetilde{D}_j \leq cn^{-\kappa},$$

for large enough n, *where* c *is defined in Condition* 3. *The same results also hold for the sets* $\mathcal{A}_2$ *and* $\mathcal{A}_2^c$ *with the test statistics being calculated using data transformed by* $\boldsymbol{\Omega}_2$.

The assumption $\tilde{\tau}_{1,p}^{-2}+\log^2(\tilde{\tau}_{2,p})=o(n^{1-2\kappa-\gamma})$ restricts how fast the upper bound $\tau_{2,p}$ in Condition 2 can diverge with the number of predictors p. Proposition 1 entails that the oracle-assisted IIS can identify all indices in $\mathcal{A}_1\cup\mathcal{A}_2$ with overwhelming probability, by thresholding the test statistics $\widetilde{D}_j$ with threshold chosen in the interval $(cn^{-\kappa}, 2cn^{-\kappa})$. In view of (12), Proposition 1 gives the variable selection consistency of the oracle-assisted IIS; that is, the set of true interaction variables $\mathcal{I}$ can be selected with asymptotic probability one. This result holds for ultra-high dimensional p satisfying $\log p = O(n^{\gamma})$ with $0<\gamma<1-2\kappa$. The key step in proving the theorem is to analyze the deviation bound of $\widetilde{D}_j$ from its population counterpart D_j. More details can be found in the supplementary material [10].

3.3. *IIS with unknown precision matrices*. In most applications, the precision matrices $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are unknown and need to be estimated. There is a large body of literature on estimating precision matrices. See, for example, [1, 5, 12, 22, 28, 29, 31], among others. These methods share a common assumption that the underlying true precision matrix is sparse. In this paper, we focus on the family of K_p-sparse precision matrices as introduced in Condition 4. For the estimation, we use the following class of estimators.

DEFINITION 1 (Acceptable estimator). A $p\times p$ symmetric matrix $\widehat{\boldsymbol{\Omega}}$ is an acceptable estimator of the K_p-sparse population precision matrix $\boldsymbol{\Omega}$ if it satisfies the following two conditions: (1) it is independent of the test data and is K_p'-sparse with K_p' a sequence of positive integers depending only on p, and (2) it satisfies the entry-wise estimation error bound $\|\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\|_{\max}\leq C_1K_p^2\sqrt{(\log p)/n}$ with some positive constant C_1.

The same class of estimators has been introduced in and used in [9]. As discussed in [9], many existing precision matrix estimators such as CLIME [5] and Glasso [12] are acceptable under some regularity conditions. Other methods for estimating precision matrices can also yield acceptable estimators under certain conditions; see [9] for more discussions on acceptable estimators.

For each $k=1,2$, given an acceptable estimator $\widehat{\boldsymbol{\Omega}}_k$ of $\boldsymbol{\Omega}_k$, our IIS approach transforms the data matrix as $\mathbf{Z}\widehat{\boldsymbol{\Omega}}_k$. Similar to the last subsection, we only discuss

in detail IIS based on the transformation $\mathbf{Z}\widehat{\mathbf{\Omega}}_1$. Then the corresponding test statistic $\widehat{D}_j$ is

$$\widehat{D}_j = \log \hat{\sigma}_j^2 - \sum_{k=1}^{2} (n_k/n) \log\big[(\hat{\sigma}_j^{(k)})^2\big], \tag{16}$$

where $\hat{\sigma}_j^2$ is the pooled sample variance estimate for the jth feature after the transformation $\mathbf{Z}\widehat{\mathbf{\Omega}}_1$, and $(\hat{\sigma}_j^{(k)})^2$ is the class k sample variance estimate for the jth feature after the transformation for $k = 1, 2$.

With an acceptable estimate $\widehat{\mathbf{\Omega}}_1$ of $\mathbf{\Omega}_1$, the transformed data matrix $\mathbf{Z}\widehat{\mathbf{\Omega}}_1$ is expected to be close to the data matrix $\mathbf{Z}\mathbf{\Omega}_1$. Correspondingly, the test statistics $\widehat{D}_j$ are expected to be close to the test statistics $\widetilde{D}_j$ defined in (14), which ensures that the same selection consistency property discussed in Proposition 1 is inherited by using test statistics $\widehat{D}_j$. This result is formally summarized below in Theorem 1.

Define $\widehat{\mathcal{A}}_1 = \{1 \le j \le p : \widehat{D}_j > \omega_n\}$ with $\omega_n > 0$ the threshold level depending only on n. Let

$$T_{n,p} = \tilde{C}_1 \tilde{\tau}_{1,p}^{-1} \tau_{2,p} (K_p + K_p') K_p^3 \sqrt{(\log p)/n} \max\big\{(K_p + K_p') K_p \sqrt{(\log p)/n}, 1\big\},$$

where $\tilde{C}_1$ is some positive constant, and $\tilde{\tau}_{1,p}$ and $\tau_{2,p}$ are the same as in Proposition 1.

THEOREM 1. *Assume that the conditions in Proposition 1 are satisfied and that for each $k = 1, 2$, $\widehat{\mathbf{\Omega}}_k$ is an acceptable estimator of the true precision matrix $\mathbf{\Omega}_k$. In addition, assume that Condition 4 is satisfied and $T_{n,p} = o(n^{-\kappa})$. Then with probability at least $1 - \exp\{-Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2} + \log^2(\tilde{\tau}_{2,p})]\}$ for some positive constant C, it holds that*

$$\widehat{\mathcal{A}}_1 = \mathcal{A}_1 \qquad \text{with } \omega_n \in (\alpha_n - \beta_n, \alpha_n)$$

for large enough n, where $\alpha_n = 2cn^{-\kappa} - T_{n,p}$ and $\beta_n = cn^{-\kappa} - 2T_{n,p}$ with c defined in Condition 3. The same result holds for sets $\mathcal{A}_2$ and $\widehat{\mathcal{A}}_2$ with $\widehat{\mathcal{A}}_2$ defined analogously to $\widehat{\mathcal{A}}_1$ using the test statistics calculated with data transformed by $\widehat{\mathbf{\Omega}}_2$.

As shown in the proof of Theorem 1 in Section 7, it holds that

$$\min_{j \in \mathcal{A}_1} \widehat{D}_j \ge \alpha_n \quad \text{and} \quad \max_{j \in \mathcal{A}_1^c} \widehat{D}_j \le \alpha_n - \beta_n,$$

with asymptotic probability one. The term β_n measures how different the test statistics are in and outside of set $\mathcal{A}_1 \cup \mathcal{A}_2$. Thus, by thresholding the test statistics $\widehat{D}_j$ with appropriately selected threshold level, the index set $\mathcal{A}_1 \cup \mathcal{A}_2$ can be identified with asymptotic probability one, and consequently, our IIS method enjoys the variable selection consistency as described in Theorem 1. We will discuss the implementation of IIS with test statistics (16) in detail in Section 5.

Compared to Proposition 1, the lower bound of the test statistics over $\mathcal{A}_1$, which is given by α_n, is smaller than the one in Proposition 1, reflecting the sacrifice caused by estimating precision matrices. The additional assumption on $T_{n,p}$ is related to the sparsity level and estimation errors of precision matrices. Under these two assumptions, α_n and β_n are close to $2cn^{-\kappa}$ and $cn^{-\kappa}$, the bounds given in Proposition 1, respectively, implying a relatively small price paid in estimating precision matrices.

4. Post-screening variable selection. Denote by $\widehat{\mathcal{I}} = \widehat{\mathcal{A}}_1 \cup \widehat{\mathcal{A}}_2$ the index set identified by the IIS approach. Let $d = |\widehat{\mathcal{I}}|$ be its cardinality. Then the variable selection consistency of IIS guarantees that $\widehat{\mathcal{I}}$ is the true set of interaction variables $\mathcal{I}$ with asymptotic probability one. By the sparsity of $\mathbf{\Omega}$ assumed in Section 2, the cardinality d is equal to $q = o(\min\{n, p\})$ with overwhelming probability. With selected variables in $\widehat{\mathcal{I}}$, interactions can be reconstructed as $\mathcal{B} = \{Z_j Z_\ell, \text{ for all } j, \ell \in \widehat{\mathcal{I}}\}$, which indicates that IIS reduces the dimensionality of interactions from $O(p^2)$ to less than $o(\min\{n^2, p^2\})$ with overwhelming probability. Important questions are how to further select active interactions and how to conduct classification using these selected interactions.

In the classification literature, variable selection techniques have been frequently used to construct high-dimensional classifiers, for example, the penalized logistic regression [13, 32], the LPD rule [4], and the DSDA approach [20], among many others. In this paper, we use the idea of penalized logistic regression to further select important main effects and interactions. Before going into details, we first introduce some nation. For a feature vector $\mathbf{z} = (Z_1, \ldots, Z_p)^T$, let $\mathbf{x} = (1, Z_1, \ldots, Z_p, Z_1^2, Z_1 Z_2, \ldots, Z_{p-1} Z_p, Z_p^2)^T$ be the $\tilde{p}$-dimensional full augmented feature vector with $\tilde{p} = (p+1)(p+2)/2$. Assume that the conditional probability of success $\pi(\mathbf{x}) = P(\Delta = 1|\mathbf{x}) = P(\Delta = 1|\mathbf{z})$ is linked to the feature vector $\mathbf{x}$ by the following logistic regression model:

$$\operatorname{logit}\big(\pi(\mathbf{x})\big) = \log \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \mathbf{x}^T \boldsymbol{\theta}, \tag{17}$$

where $\boldsymbol{\theta}$ is the regression coefficient vector. Based on (17), a new observation $\mathbf{z}$ is classified into class 1 if and only if $\mathbf{x}^T \boldsymbol{\theta} > 0$. We remark that if both $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ are Gaussian distributed, the decision rule derived from the logistic regression model (17) is identical to the Bayes rule (3), which is our main reason of using penalized logistic regression for selecting important main effects and interactions.

Write $\mathscr{T} \subset \{1, \ldots, \tilde{p}\}$, the set of indices formed by the intercept, all main effects $Z_1, \ldots, Z_p$ and interactions $Z_k Z_\ell$ with $k, \ell \in \widehat{\mathcal{I}}$. If there is no interaction screening and $\widehat{\mathcal{I}} = \{1, \ldots, p\}$, then $\mathscr{T} = \{1, \ldots, \tilde{p}\}$, meaning that all pairwise interactions are used in post-screening variable selection step.

Denote by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{\tilde{p}})$ the full augmented design matrix with $\mathbf{x}_i$ the full augmented feature vector for the ith observation $\mathbf{z}_i$. In order to estimate the regression coefficient vector $\boldsymbol{\theta}$, we consider the reduced feature space

spanned by the $1 + p + d(d+1)/2$ columns of $\mathbf{X}$ with indices in $\mathscr{T}$ and estimate $\boldsymbol{\theta}$ by solving the following regularization problem:

$$\widehat{\boldsymbol{\theta}} = \mathop{\arg\min}_{\boldsymbol{\theta}\in\mathbb{R}^{\tilde{p}}, \boldsymbol{\theta}_{\mathscr{T}^c}=\mathbf{0}} \left\{ n^{-1} \sum_{i=1}^{n} \ell(\mathbf{x}_i^T \boldsymbol{\theta}, \Delta_i) + \operatorname{pen}(\boldsymbol{\theta}) \right\}, \tag{18}$$

where $\mathscr{T}^c$ is the complement of $\mathscr{T}$, $\ell(\mathbf{x}^T\boldsymbol{\theta}, \Delta) = -\Delta(\mathbf{x}^T\boldsymbol{\theta}) + \log[1 + \exp(\mathbf{x}^T\boldsymbol{\theta})]$ is the logistic loss function and $\operatorname{pen}(\boldsymbol{\theta})$ is some penalty function on the parameter vector $\boldsymbol{\theta}$. Various penalty functions have been proposed in the literature for high-dimensional variable selection; see, for example, Lasso [24], SCAD [8], SICA [19] and MCP [30], among many others. See also [11] for the asymptotic equivalence of various regularization methods. Due to the existence of interactions, the design matrix $\mathbf{X}$ can have highly correlated columns. To overcome the difficulty caused by potential high collinearity, in our application we propose to use the elastic net penalty [33], which takes the form $\operatorname{pen}(\boldsymbol{\theta}) = \lambda_1\|\boldsymbol{\theta}\|_1 + \lambda_2\|\boldsymbol{\theta}\|_2^2$ with λ_1 and λ_2, two nonnegative regularization parameters. Similar types of penalty functions have also been used and studied in [3] and [15]. Note that solving the regularization problem (18) in the reduced parameter space $\mathscr{T}$ is computationally more efficient than solving it in the original $\tilde{p}$-dimensional parameter space.

Generally speaking, the post-screening variable selection is able to reduce the number of false positive interactions. Thus, only when there are interactions surviving both the screening step and variable selection step, sparse QDA will be used for classification; otherwise, sparse LDA will be used for classification. In this sense, our approach is adaptive and automatically chooses between sparse LDA and sparse QDA.

4.1. *Oracle inequalities.* Denote by $S = \operatorname{supp}(\boldsymbol{\theta}_0)$ the support of the true regression coefficient vector $\boldsymbol{\theta}_0$ and S^c its complement. Let $s = |S|$ be the cardinality of the set S. For any $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{\tilde{p}})^T \in \mathbb{R}^{\tilde{p}}$, we use $\boldsymbol{\delta}_S$ to denote the subvector formed by the components δ_j with $j \in S$. The following conditions are needed for establishing the oracle inequalities for $\widehat{\boldsymbol{\theta}}$ defined in (18):

CONDITION 5. *There exists some positive constant* $0 < \pi_{\min} < 1/2$ *such that* $\pi_{\min} < P(\Delta = 1|\mathbf{z}) < 1 - \pi_{\min}$ *for all* $\mathbf{z}$.

CONDITION 6. *There exists some constant* $\phi > 0$ *such that*

$$\boldsymbol{\delta}^T \tilde{\boldsymbol{\Sigma}} \boldsymbol{\delta} \geq \phi^2 \boldsymbol{\delta}_S^T \boldsymbol{\delta}_S \tag{19}$$

for any $\boldsymbol{\delta} \in \mathbb{R}^{\tilde{p}}$ *satisfying* $\|\boldsymbol{\delta}_{S^c}\|_1 \leq 4(s^{1/2} + \lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2$, *where* $\tilde{\boldsymbol{\Sigma}} = E(\mathbf{x}^T\mathbf{x})$.

Condition 5 is a mild condition which is commonly imposed in logistic regression and ensures that the conditional variance of the response variable is uniformly

bounded away from zero. Condition 6 is inspired by the restricted eigenvalue (RE) assumptions in [2], where it was introduced for establishing the oracle inequalities for the lasso estimator [24] and the Dantzig selector [6]. The set on which (19) holds in Condition 6 also involves $\lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2$, which is needed to deal with the L_2 term in the elastic net penalty [33]. A similar condition has been used in [15] for studying the oracle inequalities for the smooth-Lasso and other $\ell_1+\ell_2$ methods in ultrahigh-dimensional linear regression models with deterministic design and no interactions. In our setting, the logistic regression model with random design and the existence of interactions add extra technical difficulties in establishing the oracle inequalities.

THEOREM 2. *Assume that all conditions in Theorem 1 and Conditions 5–6 are satisfied. Moreover, assume that* $\lambda_1 \geq c_0\sqrt{\log(p)/n}$ *with some positive constant* c_0, $5s^{1/2}+4\lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2 = O(n^{\xi/2})$, *and* $\log(p) = o(n^{1/2-2\xi})$ *with some constant* $0 \leq \xi < 1/4$. *Then with probability at least* $1-\exp\{-Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2}+\log^2(\tilde{\tau}_{2,p})]\} - O(p^{-c_1})$, *it holds simultaneously that*

$$\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1 \leq 32\widetilde{C}^{-1}\phi^{-2}\big(\lambda_1 s^{1/2}+\lambda_2\|\boldsymbol{\theta}_0\|_2\big)^2/\lambda_1,$$
$$n^{-1/2}\big\|\mathbf{X}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\big\|_2 \leq 4\widetilde{C}^{-1}\phi^{-1}\big(\lambda_1 s^{1/2}+\lambda_2\|\boldsymbol{\theta}_0\|_2\big),$$

where $\widetilde{C}$ *is some positive constant. Moreover, the same results hold with probability at least* $1-O(p^{-c_1})$ *for the regularized estimator* $\widehat{\boldsymbol{\theta}}$ *without the interaction screening step, that is, without the constraint* $\boldsymbol{\theta}_{\mathcal{J}^c}=\mathbf{0}$ *in* (18).

Theorem 2 presents the oracle inequalities for the regularized estimator $\widehat{\boldsymbol{\theta}}$ defined in (18). It extends the oracle inequalities in Theorem 1 of [15] from the linear model with deterministic design and no interactions to the logistic regression model with random design and interactions. Dealing with interactions and large random design matrix needs more delicate analysis. It is worth pointing out that the results in Theorem 2 also apply to the regularized estimator with $d=p$, that is, the case without interaction screening.

4.2. *Oracle inequality for misclassification rate.* Recall that based on the logistic regression model (17), the oracle classifier classifies a new observation $\mathbf{z}$ to class 1 if and only if $\mathbf{x}^T\boldsymbol{\theta}_0>0$, where $\mathbf{x}$ is the $\tilde{p}$-dimensional augmented feature vector corresponding to $\mathbf{z}$. Thus the oracle misclassification rate is

$$R = \pi R(2|1)+(1-\pi)R(1|2),$$

where $R(i|j)$ is the probability that a new observation from class j is misclassified to class i based on the oracle classifier. As discussed in the last subsection, the oracle classifier $\mathbf{x}^T\boldsymbol{\theta}_0$ is the Bayes rule if the feature vectors $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ from classes 1 and 2 are both Gaussian.

Correspondingly, given the sample $\{(\mathbf{z}_i^T, \Delta_i)\}_{i=1}^n$, the misclassification rate of the plug-in classifier $\mathbf{x}^T\widehat{\boldsymbol{\theta}}$ with $\widehat{\boldsymbol{\theta}}$ defined in (18) takes the following form:

$$R_n = \pi R_n(2|1) + (1-\pi)R_n(1|2),$$

where $R_n(i|j)$ is the probability that a new observation from class j is misclassified to class i by the plug-in classifier.

We introduce some notation before stating our theoretical result on misclassification rate. Denote by $F_1(x)$ and $F_2(x)$ the cumulative distribution functions of the oracle classifier $\mathbf{x}^T\boldsymbol{\theta}_0$ under classes 1 and 2, respectively. Let

$$r_n = \max\Big\{\sup_{x\in[-\epsilon_0,\epsilon_0]}|F_1'(x)|, \sup_{x\in[-\epsilon_0,\epsilon_0]}|F_2'(x)|\Big\},$$

where ϵ_0 is a small positive constant, and $F_1'(x)$ and $F_2'(x)$ are the first-order derivatives of $F_1(x)$ and $F_2(x)$, respectively.

CONDITION 7. *Let* $\Lambda_n = \log(p)(\lambda_1 s^{1/2} + \lambda_2\|\boldsymbol{\theta}_0\|_2)^2/\lambda_1$. *It holds that* $\Lambda_n = o(1)$ *and* $r_n\Lambda_n = o(1)$.

THEOREM 3. *Assume that all conditions in Theorem 2 and Condition 7 are satisfied. Then with probability at least* $1 - \exp\{-Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2} + \log^2(\tilde{\tau}_{2,p})]\} - O(p^{-c_1})$, *we have*

$$0 \le R_n \le R + O\big(p^{-c_1}\big) + O(r_n\Lambda_n) \tag{20}$$

for all sufficiently large n, *where* c_1 *is some positive constant. Moreover, the same inequality holds with probability at least* $1 - O(p^{-c_1})$ *for the plug-in classifier based on the regularization estimator* $\widehat{\boldsymbol{\theta}}$ *without interaction screening*.

Theorem 3 ensures that with overwhelming probability, the misclassification rate of the plug-in classifier is at most $O(p^{-c_1}) + O(r_n\Lambda_n)$ worse than that of the oracle classifier. If r_n is upper-bounded by some constant, $\lambda_1 = O(\sqrt{(\log p)/n})$, and $\lambda_2\|\boldsymbol{\theta}_0\|_2 = O(s^{1/2}\lambda_1)$, then (20) becomes $0 \le R_n \le R + O(p^{-c_1}) + O(s(\log p)^{3/2}n^{-1/2})$. In the setting of two-class Gaussian classification, the misclassification rate R_n can also be lower bounded by R, by noting that the oracle classifier $\mathbf{x}^T\boldsymbol{\theta}_0$ is the Bayes rule. Thus the plug-in classifier is consistent. This result is formally summarized in the following corollary.

COROLLARY 1. *Assume that both* $\mathbf{z}^{(1)}$ *and* $\mathbf{z}^{(2)}$ *are Gaussian distributed. Then under the same conditions as in Theorem 3, with probability at least* $1 - \exp\{-Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2} + \log^2(\tilde{\tau}_{2,p})]\} - O(p^{-c_1})$, *it holds that*

$$R \le R_n \le R + O\big(p^{-c_1}\big) + O(r_n\Lambda_n).$$

5. Numerical studies.

5.1. *Implementation*. We apply the SIRI method in [16] to implement IIS in our proposal. See Section 5.2 in [16] for more details on how to choose thresholds in SIRI. The R code for SIRI is available at http://www.people.fas.harvard.edu/~junliu/SIRI/.

It is worth mentioning that as recommended in [16], SIRI is implemented as an iterative stepwise procedure. That is, the next active interaction variable is chosen based on the current set of interaction variables rather than using a one-time hard-thresholding to select all interaction variables. The iterative stepwise procedure is more stable in practice. Jiang and Liu [16] proved the nice property of SIRI method in selecting interaction variables in the sliced inverse index model setting. We remark that the same theoretical results hold under our model setting as long as an extra condition similar to the stepwise detectable condition in [16] is imposed on the population variances. Since the proofs are very similar to the ones in [16], to save space, we do not formally state the results here. Instead, we refer the readers to [16] for more details.

In the second stage of our proposal, we employ the R package glmnet for variable selection. An refitting step after selection is added when calculating classification error. For the ease of presentation, our two-stage procedure is referred to as IIS-SQDA.

For comparisons, we also include LDA, QDA, penalized logistic regression (PLR), DSDA and the oracle procedure (Oracle). The LDA and QDA methods are implemented by directly plugging in the sample estimates of the unknown parameters. The oracle procedure uses the information of the true underlying sparse model and is thus a low-dimensional QDA. For PLR, we consider two different versions, PLR and PLR2, where the former uses main effects only and the latter includes additionally all possible pairwise interactions. For fair comparison, an refitting step is also conducted for PLR and PLR2, as we do for IIS-SQDA.

5.2. *Simulation studies*. We conducted two simulation studies to evalaute the performace of IIS-SQDA. The class 1 distribution is chosen to be $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ with $\boldsymbol{\mu}_1 = \boldsymbol{\Sigma}_1\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Omega}_1^{-1}$, and the class 2 distribution is chosen to be $N(\mathbf{0}, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\Sigma}_2 = \boldsymbol{\Omega}_2^{-1}$, where $\boldsymbol{\Omega}_1$, $\boldsymbol{\Omega}_2$ and $\boldsymbol{\delta}$ will be specified later.

5.2.1. *Study* 1. We demonstrate the performance of the oracle-assisted IIS approach and examine the resulting classification and variable selection performance. The results presented here can be used as a benchmark for evaluating the performance of IIS with unknown precision matrices. We consider the following setting for $\boldsymbol{\delta}$ and precision matrices $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$:

Model 1: $(\boldsymbol{\Omega}_1)_{ij} = 0.5^{|i-j|}$, $\boldsymbol{\Omega}_2 = \boldsymbol{\Omega}_1 + \boldsymbol{\Omega}$ where $\boldsymbol{\Omega}$ is a symmetric and sparse matrix with $\boldsymbol{\Omega}_{5,5} = \boldsymbol{\Omega}_{25,25} = \boldsymbol{\Omega}_{45,45} = -0.29$ and $\boldsymbol{\Omega}_{5,25} = \boldsymbol{\Omega}_{5,45} = \boldsymbol{\Omega}_{25,45} =$

-0.15. The other 3 nonzero entries in the lower triangle of $\mathbf{\Omega}$ are determined by symmetry. $\boldsymbol{\delta} = (0.6, 0.8, 0, \ldots, 0)^T$. The dimension p is 2000.

Thus there are two main effects and six interaction terms under our broad definition of interaction in the Bayes rule (3).

We use two performance measures, false positive (FP), and false negative (FN), to evaluate the screening performance of IIS. FP is defined as the number of irrelevant interaction variables falsely kept while FN is defined as the number of true interaction variables falsely excluded by IIS. An effective variable screening procedure is expected to have the value of FP reasonably small and the value of FN close to zero. The former implies that the variable screening procedure can effectively reduce the dimensionality whereas the latter implies that the sure screening property holds. The means and standard errors (in parentheses) of FP and FN for interaction variables based on 100 replications are 0.63 (0.08) and 0.14 (0.03), respectively, in the screening step. This demonstrates the fine performance of our IIS approach in selecting interaction variables.

We further investigate the classification and variable selection performance of our proposal. Five performance measures are employed to summarize the results. The first measure is the misclassification rate (MR), which is calculated as the proportion of observations in an independently simulated test set of size 10,000 being allocated to the incorrect class. The second and third are FP.main and FP.inter, which represent the numbers of irrelevant main effects and irrelevant interaction effects falsely included in the classification rule, respectively. The fourth and fifth are FN.main and FN.inter, which represent the numbers of relevant main effects and relevant interaction effects falsely excluded in the classification rule, respectively. Note that the definitions of FP.inter and FN.inter here are different from those screening performance measures FP and FN, which are defined earlier. In fact, FP.inter and FN.inter are defined with respect to the number of interaction effects whereas screening performance measures FP and FN are defined with respect to the number of interaction variables.

The variable selection and classification results for different methods are reported in Table 2. PLR2 is not computationally efficient in this case due to the huge number of two-way interactions. The conventional LDA and QDA are not applicable as $n_1 = n_2 = 100 < p$. So we only compare the variable selection and classification performance of our proposal, IIS-SQDA, with DSDA, PLR and the Oracle. It is made clear that IIS-SQDA has better classification performance than PLR and DSDA.

5.2.2. *Study* 2. In this study, we evaluate the performance of the IIS approach with the estimated precision matrices and examine the resulting classification and variable selection performance. We consider the following four different model settings for precision matrices:

TABLE 2
The means and standard errors (in parentheses) of various performance measures by different classification methods for study 1 *based on* 100 *replications*

Measure	PLR	DSDA	IIS-SQDA	Oracle
MR (%)	40.59 (0.40)	38.04 (0.35)	15.09 (0.40)	12.07 (0.06)
FP.main	25.69 (5.45)	22.80 (5.10)	2.63 (0.66)	0 (0)
FP.inter	–	–	0.62 (0.12)	0 (0)
FN.main	1.80 (0.04)	0.82 (0.05)	1.22 (0.05)	0 (0)
FN.inter	–	–	0.47 (0.10)	0 (0)

Model 2: $\mathbf{\Omega}_1 = I_p$, $\mathbf{\Omega}_2 = \mathbf{\Omega}_1 + \mathbf{\Omega}$, where $\mathbf{\Omega}$ is a symmetric and sparse matrix with $\mathbf{\Omega}_{10,10} = \mathbf{\Omega}_{30,30} = \mathbf{\Omega}_{50,50} = -0.6$ and $\mathbf{\Omega}_{10,30} = \mathbf{\Omega}_{10,50} = \mathbf{\Omega}_{30,50} = -0.15$. The other 3 nonzero entries in the lower triangle of $\mathbf{\Omega}$ are determined by symmetry. $\boldsymbol{\delta} = (0.6, 0.8, 0, \ldots, 0)^T$.

Model 3: $\mathbf{\Omega}_1$ is a band matrix with $(\mathbf{\Omega}_1)_{ii} = 1$ for $i = 1, \ldots, p$ and $(\mathbf{\Omega}_1)_{ij} = 0.3$ for $|i - j| = 1$. $\mathbf{\Omega}_2 = \mathbf{\Omega}_1 + \mathbf{\Omega}$ where $\mathbf{\Omega}$ is a symmetric and sparse matrix with $\mathbf{\Omega}_{10,10} = -0.3785$, $\mathbf{\Omega}_{10,30} = 0.0616$, $\mathbf{\Omega}_{10,50} = 0.2037$, $\mathbf{\Omega}_{30,30} = -0.5482$, $\mathbf{\Omega}_{30,50} = 0.0286$ and $\mathbf{\Omega}_{50,50} = -0.4614$. The other 3 nonzero entries in the lower triangle of $\mathbf{\Omega}$ are determined by symmetry. $\boldsymbol{\delta} = (0.6, 0.8, 0, \ldots, 0)^T$.

Model 4: Similar to model 1 in the last subsection, except for the dimension p.

Model 5: $\mathbf{\Omega}_1$ is a block diagonal matrix comprised of equal blocks $\mathbf{A}$, where $\mathbf{A}$ is a 2-by-2 matrix with diagonal elements equal to 1 and off-diagonal elements equal to 0.4. $\mathbf{\Omega}_2 = \mathbf{\Omega}_1 + \mathbf{\Omega}$ where $\mathbf{\Omega}$ is a symmetric and sparse matrix with $\mathbf{\Omega}_{3,3} = \mathbf{\Omega}_{6,6} = \mathbf{\Omega}_{9,9} = \mathbf{\Omega}_{12,12} = -0.2$, $\mathbf{\Omega}_{3,6} = \mathbf{\Omega}_{9,12} = 0.4$ and $\mathbf{\Omega}_{3,9} = \mathbf{\Omega}_{3,12} = \mathbf{\Omega}_{6,9} = \mathbf{\Omega}_{6,12} = -0.4$. The other 6 nonzero entries in the lower triangle of $\mathbf{\Omega}$ are determined by symmetry. The nonzero elements of $\boldsymbol{\delta}$ are located at coordinates 3, 6, 9 and 12. The corresponding values for these nonzero elements are simulated from a uniform distribution over $[0.3, 0.7]$ and remain unchanged during simulations.

For each model, we consider three different dimentionalities, $p = 50$, $p = 200$ and $p = 500$. There are two main effects and six interaction terms (including quadratic terms) in the Bayes rules for models 2–4, four main effects and ten interaction terms in the Bayes rules for model 5. In models 2–4 no interaction variables are main effect variables whereas in model 5 all interaction variables are also main effect variables.

We use the same measures as in study 1 to examine the variable screening performance of the IIS approach and the variable selection and classification performance of IIS-SQDA. The means and standard errors (in parentheses) of FP and FN for these models based on 100 replications are reported in Table 3, which shows the effectiveness of our interaction screening approach. For comparison purposes,

TABLE 3
Interaction screening results for models 2–5. *The numbers reported are the means and standard errors* (*in parentheses*) *of FP and FN based on* 100 *replications*

		IIS with true Ω_1 and Ω_2		IIS with estimated Ω_1 and Ω_2	
p	Model	FP	FN	FP	FN
50	Model 2	0.45 (0.08)	0.02 (0.01)	1.57 (0.15)	0.01 (0.01)
	Model 3	0.86 (0.09)	0.48 (0.06)	1.93 (0.15)	0.15 (0.04)
	Model 4	1.68 (0.13)	0.09 (0.03)	1.04 (0.11)	0.01 (0.01)
	Model 5	1.79 (0.16)	0.02 (0.02)	1.54 (0.13)	0.01 (0.01)
200	Model 2	0.43 (0.08)	0.04 (0.02)	1.16 (0.13)	0.02 (0.01)
	Model 3	0.74 (0.09)	0.48 (0.05)	1.03 (0.14)	0.15 (0.04)
	Model 4	1.52 (0.12)	0.08 (0.03)	0.44 (0.07)	0.03 (0.02)
	Model 5	1.10 (0.12)	0.36 (0.08)	0.90 (0.10)	0.04 (0.02)
500	Model 2	0.42 (0.07)	0.11 (0.03)	0.68 (0.09)	0.01 (0.01)
	Model 3	0.53 (0.06)	0.73 (0.07)	0.65 (0.09)	0.21 (0.04)
	Model 4	1.25 (0.12)	0.09 (0.03)	0.43 (0.07)	0.03 (0.02)
	Model 5	0.85 (0.10)	0.42 (0.09)	0.59 (0.09)	0.03 (0.02)

we also include in Table 3 the screening results by oracle-assisted IIS. It is interesting to observe that the IIS with estimated precision matrices gives smaller FNs than and comparable FPs to the IIS with true precision matrices.

Tables 4–7 summarize the variable selection and classification results based on 100 replications. We observe the following:

(1) IIS-SQDA exhibits the best performance in terms of MR and interaction selection across all settings.

(2) PLR2 also has good classification accuracy in low-dimensional situations ($p = 50$), but it has inferior interaction selection results than IIS-SQDA in all settings.

(3) All linear classifiers have poor performance when the true classification boundary is nonlinear.

(4) Comparing QDA with LDA shows that including all possible interactions may not necessarily improve the classification performance. This is not surprising because QDA has many more parameters to estimate than LDA, while the sample size is very limited. Thus interaction selection is very important, even with moderate dimensionality.

(5) Comparing the results of QDA with those of PLR2 or IIS-SQDA, we observe that the classification performance can be improved substantially by using interaction screening and selection. Particularly, in most cases, the improvement becomes more significant as the dimensionality increases.

TABLE 4
The means and standard errors (in parentheses) of various performance measures by different classification methods for model 2 *based on* 100 *replications*

p	Method	MR (%)	FP.main	FP.inter	FN.main	FN.inter
50	LDA	37.91 (0.13)	48 (0)	–	0 (0)	–
	QDA	39.89 (0.11)	48 (0)	1269 (0)	0 (0)	0 (0)
	PLR	32.83 (0.23)	2.37 (0.49)	–	1.09 (0.03)	–
	DSDA	32.70 (0.18)	4.61 (0.74)	–	0.10 (0.03)	–
	PLR2	22.56 (0.33)	0.13 (0.05)	3.17 (0.70)	0.35 (0.05)	0.75 (0.09)
	IIS-SQDA	21.78 (0.22)	3.67 (0.67)	1.32 (0.23)	0.08 (0.03)	0.09 (0.04)
	Oracle	19.86 (0.08)	0 (0)	0 (0)	0 (0)	0 (0)
200	PLR	33.64 (0.31)	4.29 (1.34)	–	1.09 (0.03)	–
	DSDA	33.33 (0.26)	10.83 (2.25)	–	0.18 (0.04)	–
	PLR2	24.65 (0.51)	0.11 (0.05)	7.71 (2.27)	0.42 (0.06)	0.93 (0.09)
	IIS-SQDA	22.14 (0.30)	4.48 (0.91)	0.54 (0.11)	0.09 (0.03)	0.15 (0.05)
	Oracle	19.66 (0.06)	0 (0)	0 (0)	0 (0)	0 (0)
500	PLR	34.59 (0.39)	6.00 (1.46)	–	1.12 (0.03)	–
	DSDA	33.87 (0.28)	14.76 (3.10)	–	0.17 (0.04)	–
	PLR2	26.83 (0.58)	0.07 (0.04)	8.95 (2.02)	0.56 (0.06)	1.53 (0.11)
	IIS-SQDA	22.09 (0.30)	3.25 (1.02)	0.25 (0.08)	0.25 (0.05)	0.69 (0.09)
	Oracle	19.65 (0.06)	0 (0)	0 (0)	0 (0)	0 (0)

Another phenomenon we observed in simulation is that when the number of predictors p is as high as $p = 500$, PLR2 requires a huge memory space that it easily causes memory outflow in a regular office PC with 8 GB memory.

In addition, note that the misclassification rates of all methods in model 5 are significantly higher than that of the Oracle classifier. We emphasize that it is due to the small true coefficients in the Bayes rule and the relatively complex true model. In fact, the setting of model 5 is so challenging that all other methods have close to or over 40% MR when $p = 200$ or 500.

5.3. *Real data analysis*. We apply the same classification methods as in Section 5.2 to the breast cancer data, originally studied in [26]. The purpose of the study is to classify female breast cancer patients according to relapse and nonrelapse clinical outcomes using gene expression data. The total sample size is 78 with 44 patients in the good prognosis group and 34 patients in the poor prognosis group. There are some missing values with one patient in the poor prognosis group so it was removed from study here. Thus $n_1 = 44$ and $n_2 = 33$. Our study uses the $p = 231$ genes reported in [26].

We randomly split the 77 samples into a training set and a test set such that the training set consists of 26 samples from the good prognosis group and 19 samples from the poor prognosis group. Correspondingly, the test set has 18 samples from the good prognosis group and 14 samples from the poor prognosis group. For each

TABLE 5

The means and standard errors (in parentheses) of various performance measures by different classification methods for model 3 *based on* 100 *replications*

p	Method	MR (%)	FP.main	FP.inter	FN.main	FN.inter
50	LDA	39.43 (0.15)	48 (0)	–	0 (0)	–
	QDA	43.47 (0.10)	48 (0)	1269 (0)	0 (0)	0 (0)
	PLR	36.12 (0.26)	5.95 (0.93)	–	1.21 (0.04)	–
	DSDA	35.05 (0.22)	8.81 (1.06)	–	0.07 (0.03)	–
	PLR2	30.15 (0.44)	0.51 (0.14)	11.26 (2.78)	0.60 (0.05)	2.62 (0.09)
	IIS-SQDA	27.56 (0.27)	5.60 (0.82)	2.16 (0.32)	0.19 (0.04)	2.05 (0.09)
	Oracle	24.13 (0.07)	0 (0)	0 (0)	0 (0)	0 (0)
200	PLR	37.62 (0.34)	7.82 (1.87)	–	1.47 (0.05)	–
	DSDA	36.34 (0.30)	15.06 (3.37)	–	0.36 (0.05)	–
	PLR2	32.55 (0.53)	0.25 (0.06)	17.44 (3.63)	0.90 (0.05)	2.72 (0.08)
	IIS-SQDA	26.94 (0.31)	6.43 (1.24)	0.78 (0.17)	0.42 (0.05)	2.22 (0.08)
	Oracle	22.99 (0.07)	0 (0)	0 (0)	0 (0)	0 (0)
500	PLR	38.82 (0.33)	9.31 (1.99)	–	1.58 (0.05)	–
	DSDA	37.10 (0.29)	16.06 (3.02)	–	0.42 (0.05)	–
	PLR2	35.45 (0.64)	0.34 (0.09)	55.69 (12.67)	0.99 (0.05)	3.05 (0.10)
	IIS-SQDA	26.78 (0.31)	3.22 (1.09)	0.23 (0.05)	0.98 (0.02)	2.65 (0.09)
	Oracle	23.00 (0.08)	0 (0)	0 (0)	0 (0)	0 (0)

TABLE 6

The means and standard errors (in parentheses) of various performance measures by different classification methods for model 4 *based on* 100 *replications*

p	Method	MR (%)	FP.main	FP.inter	FN.main	FN.inter
50	LDA	38.84 (0.16)	48 (0)	–	0 (0)	–
	QDA	31.10 (0.16)	48 (0)	1269 (0)	0 (0)	0 (0)
	PLR	36.06 (0.24)	5.89 (0.78)	–	1.39 (0.05)	–
	DSDA	35.36 (0.21)	10.41 (1.18)	–	0.24 (0.04)	–
	PLR2	16.55 (0.40)	0.40 (0.08)	22.80 (1.72)	1.08 (0.06)	0.33 (0.06)
	IIS-SQDA	15.49 (0.33)	9.51 (1.34)	2.91 (0.38)	0.39 (0.05)	0.04 (0.03)
	Oracle	12.14 (0.06)	0 (0)	0 (0)	0 (0)	0 (0)
200	PLR	38.01 (0.30)	9.86 (2.04)	–	1.64 (0.05)	–
	DSDA	36.39 (0.25)	13.98 (2.18)	–	0.46 (0.05)	–
	PLR2	16.79 (0.48)	0.09 (0.03)	19.99 (1.76)	1.40 (0.05)	0.48 (0.08)
	IIS-SQDA	13.98 (0.28)	2.30 (0.72)	0.26 (0.09)	0.98 (0.05)	0.10 (0.05)
	Oracle	12.12 (0.07)	0 (0)	0 (0)	0 (0)	0 (0)
500	PLR	39.51 (0.35)	12.98 (2.13)	–	1.72 (0.05)	–
	DSDA	37.90 (0.29)	24.04 (3.94)	–	0.53 (0.05)	–
	PLR2	16.38 (0.52)	0.06 (0.02)	16.79 (1.36)	1.43 (0.05)	0.74 (0.10)
	IIS-SQDA	14.10 (0.28)	2.11 (0.57)	0.16 (0.07)	1.07 (0.05)	0.12 (0.06)
	Oracle	12.11 (0.06)	0 (0)	0 (0)	0 (0)	0 (0)

TABLE 7
The means and standard errors (in parentheses) of various performance measures by different classification methods for model 5 *based on* 100 *replications*

p	Method	MR (%)	FP.main	FP.inter	FN.main	FN.inter
50	LDA	43.18 (0.14)	46 (0)	–	0 (0)	–
	QDA	41.69 (0.12)	46 (0)	1265 (0)	0 (0)	0 (0)
	PLR	40.16 (0.26)	4.77 (0.73)	–	1.93 (0.10)	–
	DSDA	38.89 (0.26)	7.98 (1.22)	–	1.12 (0.10)	–
	PLR2	34.55 (0.39)	1.06 (0.22)	19.51 (3.53)	2.14 (0.11)	4.16 (0.13)
	IIS-SQDA	27.68 (0.23)	7.64 (0.86)	2.11 (0.28)	0.90 (0.09)	2.61 (0.18)
	Oracle	22.30 (0.10)	0 (0)	0 (0)	0 (0)	0 (0)
200	PLR	42.15 (0.32)	18.18 (3.10)	–	2.22 (0.12)	–
	DSDA	39.22 (0.32)	16.23 (3.82)	–	1.36 (0.11)	–
	PLR2	41.50 (0.38)	0.34 (0.08)	72.73 (10.99)	2.66 (0.10)	5.24 (0.14)
	IIS-SQDA	30.04 (0.32)	11.29 (1.83)	0.91 (0.18)	1.52 (0.10)	4.08 (0.17)
	Oracle	22.24 (0.08)	0 (0)	0 (0)	0 (0)	0 (0)
500	PLR	43.83 (0.32)	29.19 (4.70)	–	2.36 (0.13)	–
	DSDA	40.03 (0.32)	20.54 (4.30)	–	1.58 (0.10)	–
	PLR2	44.92 (0.32)	0.77 (0.13)	123.39 (15.77)	2.97 (0.09)	7.19 (0.15)
	IIS-SQDA	32.84 (0.32)	19.59 (3.32)	0.57 (0.10)	1.61 (0.12)	4.61 (0.18)
	Oracle	22.12 (0.07)	0 (0)	0 (0)	0 (0)	0 (0)

split, we applied four different methods, PLR, PLR2, DSDA and IIS-SQDA to the training data and then calculated the classification error using the test data. The tuning parameters were selected using the cross-validation. We repeated the random splitting for 100 times. The means and standard errors of classification errors and model sizes for different classification methods are summarized in Table 8. The average number of genes contributing to the selected interactions over 100 random splittings were 22.96 and 2.86 for PLR2 and IIS-SQDA, respectively. We can observe that our proposed procedure has the best classification performance.

TABLE 8
Misclassification rate and model size on the breast cancer data in [26] *over* 100 *random splits. Standard errors are in the parentheses*

		Model size		
Method	MR (%)	Main	Interaction	All
DSDA	23.62 (0.74)	37.38 (1.57)	–	37.38 (1.57)
PLR	21.72 (0.78)	45.04 (1.35)	–	45.04 (1.35)
PLR2	40.47 (0.61)	14.87 (1.81)	19.95 (3.28)	34.82 (4.77)
IIS-SQDA	19.97 (0.77)	47.77 (1.16)	3.03 (0.32)	50.80 (1.31)

6. Discussion. We have proposed a new two-stage procedure, IIS-SQDA, for two-class classification with possibly unequal covariance matrices in the high-dimensional setting. The proposed procedure first selects interaction variables and reconstructs interactions using these retained variables and then achieves main effects and interactions selection through regularization. The fine performance of IIS-SQDA has been demonstrated through theoretical study and numerical analyses.

For future study, it would be interesting to extend the proposed procedure to multi-class classification problems. In addition, IIS transforms the data using the CLIME estimates of the precision matrices $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$, which can be slow to calculate when the number of predictors p is very large. One possible solution is to first reduce the dimensionality using some screening method and then apply our IIS-SQDA for interaction screening and classification. We are also in the process of developing a scalable version of the IIS which significantly improves the computational efficiency.

7. Proofs of main theorems. In this section, we list main lemmas and present the proofs for main theorems. The secondary lemmas and additional technical proofs for all lemmas are provided in the supplementary material [10].

7.1. *Lemmas*. We introduce the following lemmas which are used in the proofs of Theorems 1–3.

LEMMA 1. *Under model setting* (2) *and the conditions in Theorem* 1, *for sufficiently large* n, *with probability at least* $1 - p\exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa})$, *it holds that*

$$\max_{1\le j\le p} |\hat{\sigma}_j^2/\tilde{\sigma}_j^2 - 1| \le T_{n,p}/6$$

for some positive constant C, *where* $T_{n,p}$ *is the same as in Theorem* 1.

LEMMA 2. *Under Condition* 5, *we have*

$$\widetilde{C}n^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \text{pen}(\widehat{\boldsymbol{\theta}}) \le \|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty\|\boldsymbol{\delta}\|_1 + \text{pen}(\boldsymbol{\theta}_0), \tag{21}$$

where $\widetilde{C}$ *is some positive constant depending on the positive constant* $\pi_{\min}$ *in Condition* 5, $\boldsymbol{\delta} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ *is the estimation error for the regularized estimator* $\widehat{\boldsymbol{\theta}}$ *defined in* (18) *and* $\boldsymbol{\varepsilon} = \mathbf{y} - E(\mathbf{y}|\mathbf{X})$ *with* $\mathbf{y} = (\Delta_1, \ldots, \Delta_n)^T$.

LEMMA 3. *Assume that Condition* 1 *holds. If* $\log(p) = o(n)$, *then with probability* $1 - O(p^{-\tilde{c}_1})$, *we have* $\|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty \le 2^{-1}c_0\sqrt{\log(p)/n}$, *where* c_0 *is some positive constant and* $\boldsymbol{\varepsilon} = \mathbf{y} - E(\mathbf{y}|\mathbf{X})$ *with* $\mathbf{y} = (\Delta_1, \ldots, \Delta_n)^T$.

LEMMA 4. *Assume that Conditions 1 and 6 hold. If* $5s^{1/2} + 4\lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2 = O(n^{\xi/2})$ *and* $\log(p) = o(n^{1/2-2\xi})$ *with constant* $0 \le \xi < 1/4$, *then when* n *is sufficiently large, with probability at least* $1 - O(p^{-\tilde{c}_2})$, *where* $\tilde{c}_2$ *is some positive constant, it holds that*

$$n^{-1/2}\|\mathbf{X}\boldsymbol{\delta}\|_2 \ge (\phi/2)\|\boldsymbol{\delta}_S\|_2$$

for any $\boldsymbol{\delta} \in \mathbb{R}^{\tilde{p}}$ *satisfying* $\|\boldsymbol{\delta}_{S^c}\|_1 \le 4(s^{1/2} + \lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2$.

LEMMA 5. *Assume that* $\mathbf{w} = (W_1, \ldots, W_p)^T \in \mathbb{R}^p$ *is sub-Gaussian. Then for any positive constant* c_1, *there exists some positive constant* C_2 *such that*

$$P\Big\{\max_{1\le j\le p}|W_j| > C_2\sqrt{\log(p)}\Big\} = O\big(p^{-c_1}\big).$$

7.2. *Proof of Theorem 1.* Since we have the inequality

$$|\widehat{D}_j - D_j| \le |\widehat{D}_j - \widetilde{D}_j| + |\widetilde{D}_j - D_j|, \tag{22}$$

the key of the proof is to show that with overwhelming probability, $\widehat{D}_j$ and $\widetilde{D}_j$ are uniformly close as $n \to \infty$. Then together with Proposition 1, we can prove the desired result in Theorem 1. The same notation C will be used to denote a generic constant without loss of generality.

We proceed to prove that $\widehat{D}_j$ and $\widetilde{D}_j$ are uniformly close. By definitions of $\widehat{D}_j$ and $\widetilde{D}_j$, along with the fact that $|n_k/n| \le 1$ for $k = 1$ and 2, we decompose the difference between $\widehat{D}_j$ and $\widetilde{D}_j$ as

$$\begin{aligned}&\max_{1\le j\le p}|\widehat{D}_j - \widetilde{D}_j| \\ &\quad\le \max_{1\le j\le p}\big|\log\hat{\sigma}_j^2 - \log\tilde{\sigma}_j^2\big| + \sum_{k=1}^{2}\max_{1\le j\le p}\big|\log\big[\big(\tilde{\sigma}_j^{(k)}\big)^2\big] - \log\big[\big(\hat{\sigma}_j^{(k)}\big)^2\big]\big|.\end{aligned} \tag{23}$$

The following argument is conditioning on the event, denoted by $\mathcal{E}_1$, such that the results hold in Lemma 1. Then $\hat{\sigma}_j^2$ and $\tilde{\sigma}_j^2$ are uniformly close. Since $x_n^{-1}\log(1 + x_n) \to 1$ as $x_n \to 0$, it follows that

$$\log\big(\hat{\sigma}_j^2/\tilde{\sigma}_j^2\big)/\big(\hat{\sigma}_j^2/\tilde{\sigma}_j^2 - 1\big) \to 1$$

uniformly for all j as $n \to \infty$. Thus, with a sufficiently large n uniformly over j, we have

$$\begin{aligned}&P\Big(\max_{1\le j\le p}\big|\log\hat{\sigma}_j^2 - \log\tilde{\sigma}_j^2\big| > T_{n,p}/3|\mathcal{E}_1\Big) \\ &\quad\le P\Big(\max_{1\le j\le p}\big|\hat{\sigma}_j^2/\tilde{\sigma}_j^2 - 1\big| > T_{n,p}/6|\mathcal{E}_1\Big) \le p\exp\big(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa}\big).\end{aligned} \tag{24}$$

By a similar argument, we can derive for $k = 1, 2$,

$$P\Big(\max_{1\le j\le p}\big|\log\big[(\hat{\sigma}_j^{(k)})^2\big] - \log\big[(\tilde{\sigma}_j^{(k)})^2\big]\big| > T_{n,p}/3|\mathcal{E}_1\Big) \le p\exp\big(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa}\big).$$

In view of (23), we get

$$\begin{aligned}
&P\Big(\max_{1\le j\le p}|\widehat{D}_j - \widetilde{D}_j| > T_{n,p}|\mathcal{E}_1\Big)\\
&\quad\le P\Big(\max_{1\le j\le p}\big|\log\hat{\sigma}_j^2 - \log\tilde{\sigma}_j^2\big| > T_{n,p}/3|\mathcal{E}_1\Big)\\
&\qquad+\sum_{k=1}^{2} P\Big(\max_{1\le j\le p}\big|\log\big[(\hat{\sigma}_j^{(k)})^2\big] - \log\big[(\tilde{\sigma}_j^{(k)})^2\big]\big| > T_{n,p}/3|\mathcal{E}_1\Big)\\
&\quad\le p\exp\big(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa}\big).
\end{aligned}$$

By Lemma 1, $P(\mathcal{E}_1^c) \le p\exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa})$. It follows that

$$\begin{aligned}
P\Big(\max_{1\le j\le p}|\widehat{D}_j - \widetilde{D}_j| > T_{n,p}\Big) &\le P\Big(\max_{1\le j\le p}|\widehat{D}_j - \widetilde{D}_j| > T_{n,p}|\mathcal{E}_1\Big) + P\big(\mathcal{E}_1^c\big)\\
&\le p\exp\big(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa}\big).
\end{aligned} \tag{25}$$

Therefore, for any p satisfying $\log p = O(n^\gamma)$ with $0 < \gamma < 1 - 2\kappa$ and $\tilde{\tau}_{1,p}^{-2} = o(n^{1-2\kappa-\gamma})$, we get that for large enough n,

$$P\Big(\max_{1\le j\le p}|\widehat{D}_j - \widetilde{D}_j| > T_{n,p}\Big) \le \exp\big(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa}\big).$$

Since the same conditions hold for the matrices $\mathbf{\Sigma}_2$ and $\mathbf{\Omega}_2$, using similar arguments we can prove that the same results hold for the covariates in $\mathcal{A}_2$ with the test statistics calculated using the transformed data $\mathbf{Z}\widehat{\mathbf{\Omega}}_2$. This completes the proof of Theorem 1.

7.3. *Proof of Theorem 2.* By Theorem 1, it is sufficient to show the second part of Theorem 2. The main idea of the proof is to first define an event which holds high probability and then analyze the behavior of the regularized estimator $\widehat{\boldsymbol{\theta}}$ conditional on that event.

Define $\boldsymbol{\varepsilon} = \mathbf{y} - E(\mathbf{y}|\mathbf{X})$ with $\mathbf{y} = (\Delta_1, \ldots, \Delta_n)^T$. Since $\log(p) = o(n^{1/2-2\xi})$, it follows from Condition 1 and Lemma 3 that for any $\lambda_1 \ge c_0\sqrt{\log(p)/n}$,

$$P\{\|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty > 2^{-1}\lambda_1\} \le P\{\|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty > 2^{-1}c_0\sqrt{\log(p)/n}\} = O\big(p^{-\tilde{c}_1}\big),$$

where $\tilde{c}_1$ is some positive constant. Meanwhile, from Lemma 4, under the assumptions that $5s^{1/2} + 4\lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2 = O(n^{\xi/2})$ and $\log(p) = o(n^{1/2-2\xi})$, we have $n^{-1/2}\|\mathbf{X}\boldsymbol{\delta}\|_2 \ge (\phi/2)\boldsymbol{\delta}_S^T\boldsymbol{\delta}_S$ for any $\boldsymbol{\delta} \in \mathbb{R}^{\tilde{p}}$ satisfying $\|\boldsymbol{\delta}_{S^c}\|_1 \le 4(s^{1/2} + \lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2$ when n is sufficiently large, with probability at least $1 -$

$O(p^{-\tilde{c}_2})$. Combining these two results we obtain that with probability at least $1 - O(p^{-\tilde{c}_1}) - O(p^{-\tilde{c}_2}) = 1 - O(p^{-c_1})$ with $c_1 = \min\{\tilde{c}_1, \tilde{c}_2\}$, it holds simultaneously that

$$\|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty \le 2^{-1}\lambda_1, \tag{26}$$

$$n^{-1/2}\|\mathbf{X}\boldsymbol{\delta}\|_2 \ge (\phi/2)\boldsymbol{\delta}_S^T\boldsymbol{\delta}_S, \tag{27}$$

for any $\lambda_1 \ge c_0\sqrt{\log(p)/n}$ and $\boldsymbol{\delta} \in \mathbb{R}^{\tilde{p}}$ satisfying $\|\boldsymbol{\delta}_{S^c}\|_1 \le 4(s^{1/2} + \lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2$ when n is sufficiently large. From now on, we condition on the event that inequalities (26) and (27) hold.

It follows from Condition 5 and Lemma 2 that

$$\widetilde{C}n^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \lambda_1\|\widehat{\boldsymbol{\theta}}\|_1 + \lambda_2\|\widehat{\boldsymbol{\theta}}\|_2^2 \le \|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty\|\boldsymbol{\delta}\|_1 + \lambda_1\|\boldsymbol{\theta}_0\|_1 + \lambda_2\|\boldsymbol{\theta}_0\|_2^2,$$

where $\widetilde{C}$ is some positive constant. Thus, by inequality (26), we have

$$\widetilde{C}n^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \lambda_1\|\widehat{\boldsymbol{\theta}}\|_1 + \lambda_2\|\widehat{\boldsymbol{\theta}}\|_2^2 \le 2^{-1}\lambda_1\|\boldsymbol{\delta}\|_1 + \lambda_1\|\boldsymbol{\theta}_0\|_1 + \lambda_2\|\boldsymbol{\theta}_0\|_2^2.$$

Recall that $\boldsymbol{\delta} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$. Adding $2^{-1}\|\boldsymbol{\delta}\|_1 - 2\lambda_2\boldsymbol{\theta}_0^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ to both sides of the above inequality and rearranging terms yield

$$\begin{aligned}&\widetilde{C}n^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + 2^{-1}\lambda_1\|\boldsymbol{\delta}\|_1 + \lambda_2\|\boldsymbol{\delta}\|_2^2\\ &\quad\le \lambda_1(\|\boldsymbol{\theta}_0\|_1 - \|\widehat{\boldsymbol{\theta}}\|_1 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1) - 2\lambda_2\boldsymbol{\theta}_0^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).\end{aligned} \tag{28}$$

Note that $\|\boldsymbol{\theta}_0\|_1 - \|\widehat{\boldsymbol{\theta}}\|_1 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 = \|\boldsymbol{\theta}_{0,S}\|_1 - \|\widehat{\boldsymbol{\theta}}_S\|_1 + \|\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}\|_1$ since $|\boldsymbol{\theta}_{0,j}| - |\widehat{\boldsymbol{\theta}}_j| + |\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_{0,j}| = 0$ for all $j \in S^c$. By the triangle inequality and the Cauchy–Schwarz inequality, we have

$$\|\boldsymbol{\theta}_0\|_1 - \|\widehat{\boldsymbol{\theta}}\|_1 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 \le 2\|\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}\|_1 \le 2s^{1/2}\|\boldsymbol{\delta}_S\|_2. \tag{29}$$

Note that $|\boldsymbol{\theta}_0^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)| = |\boldsymbol{\theta}_{0,S}^T(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S})|$. An application of the Cauchy–Schwarz inequality gives

$$|\boldsymbol{\theta}_0^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)| \le \|\boldsymbol{\theta}_{0,S}\|_2\|\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}\|_2 = \|\boldsymbol{\theta}_0\|_2\|\boldsymbol{\delta}_S\|_2. \tag{30}$$

Combining these three results in (28)–(30) yields

$$\widetilde{C}n^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + 2^{-1}\lambda_1\|\boldsymbol{\delta}\|_1 + \lambda_2\|\boldsymbol{\delta}\|_2^2 \le 2(\lambda_1 s^{1/2} + \lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2, \tag{31}$$

which, together with the fact that $\|\boldsymbol{\delta}_{S^c}\|_1 \le \|\boldsymbol{\delta}\|_1$, implies a basic constraint

$$\|\boldsymbol{\delta}_{S^c}\|_1 \le 4(s^{1/2} + \lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2.$$

Thus, by inequality (27), we have $n^{-1/2}\|\mathbf{X}\boldsymbol{\delta}\|_2 \ge (\phi/2)\boldsymbol{\delta}_S^T\boldsymbol{\delta}_S$. This, together with (31), gives

$$4^{-1}\widetilde{C}\phi^2\|\boldsymbol{\delta}_S\|_2^2 \le \widetilde{C}n^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \le 2(\lambda_1 s^{1/2} + \lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2.$$

Solving this inequality yields $\|\boldsymbol{\delta}_S\|_2 \le 8\widetilde{C}^{-1}\phi^{-2}(\lambda_1 s^{1/2} + \lambda_2\|\boldsymbol{\theta}_0\|_2)$. Combining this with (31) entails that

$$\widetilde{C}n^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + 2^{-1}\lambda_1\|\boldsymbol{\delta}\|_1 + \lambda_2\|\boldsymbol{\delta}\|_2^2 \le 16\widetilde{C}^{-1}\phi^{-2}\big(\lambda_1 s^{1/2} + \lambda_2\|\boldsymbol{\theta}_0\|_2\big)^2$$

holds with probability at least $1 - O(p^{-c_1})$. Thus from the above inequality we have

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 = \|\boldsymbol{\delta}\|_1 \le 32\widetilde{C}^{-1}\phi^{-2}\big(\lambda_1 s^{1/2} + \lambda_2\|\boldsymbol{\theta}_0\|_2\big)^2/\lambda_1,$$
$$n^{-1/2}\big\|\mathbf{X}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big\|_2 = n^{-1/2}\|\mathbf{X}\boldsymbol{\delta}\|_2 \le 4\widetilde{C}^{-1}\phi^{-1}\big(\lambda_1 s^{1/2} + \lambda_2\|\boldsymbol{\theta}_0\|_2\big),$$

hold simultaneously with probability at least $1 - O(p^{-c_1})$. This completes the proof of Theorem 2.

7.4. *Proof of Theorem* 3. Recall that $\mathbf{z} = (Z_1, \ldots, Z_p)^T = \Delta\mathbf{z}^{(1)} + (1-\Delta)\mathbf{z}^{(2)}$ and $\mathbf{x} = (1, Z_1, \ldots, Z_p, Z_1^2, Z_1Z_2, \ldots, Z_{p-1}Z_p, Z_p^2)^T$. Define an event

$$\mathcal{E}_2 = \big\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 \le 32\widetilde{C}^{-1}\phi^{-2}\big(\lambda_1 s^{1/2} + \lambda_2\|\boldsymbol{\theta}_0\|_2\big)^2/\lambda_1\big\},$$

where positive constant $\widetilde{C}$ is given in Theorem 2. From Theorem 2, we have $P(\mathcal{E}_2^c) \le O(p^{-c_1})$. By Lemma 5, under Condition 1, there exists a positive constant C_2 such that

$$P\Big\{\max_{1\le j\le p}|Z_j^{(k)}| > C_2\sqrt{\log(p)}\Big\} \le O\big(p^{-c_1}\big) \tag{32}$$

for $k = 1, 2$, where $(Z_1^{(k)}, \ldots, Z_p^{(k)})^T = \mathbf{z}^{(k)}$. Define an event $\mathcal{E}_3 = \{\|\mathbf{z}\|_\infty \le C_2\sqrt{\log(p)}\}$. Then $P(\mathcal{E}_3^c) \le O(p^{-c_1})$. An application of the Bonferroni inequality gives

$$P\big(\mathcal{E}_2^c \cup \mathcal{E}_3^c\big) \le O\big(p^{-c_1}\big) + O\big(p^{-c_1}\big) = O\big(p^{-c_1}\big). \tag{33}$$

Denote by $\mathcal{C}_1$ the event $\{\mathbf{z}$ from class 1$\}$. Note that on the event $\mathcal{E}_3$, we have $\|\mathbf{x}\|_\infty \le C_2\log(p)$ where we use a generic constant C_2 to simplify notation. Using the property of conditional probability gives

$$\begin{aligned} R_n(2|1) &= P\big(\mathbf{x}^T\widehat{\boldsymbol{\theta}} \le 0|\mathcal{C}_1\big) = P\big(\mathbf{x}^T\boldsymbol{\theta}_0 \le \mathbf{x}^T(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})|\mathcal{C}_1\big) \\ &\le P\big(\mathbf{x}^T\boldsymbol{\theta}_0 \le \mathbf{x}^T(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})|\mathcal{C}_1, \mathcal{E}_2 \cap \mathcal{E}_3\big) + P\big(\mathcal{E}_2^c \cup \mathcal{E}_3^c\big). \end{aligned} \tag{34}$$

Note that conditioning on the event $\mathcal{E}_2 \cap \mathcal{E}_3$, $\mathbf{x}^T(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})$ can be bounded as

$$\begin{aligned} \big|\mathbf{x}^T(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})\big| &\le \|\mathbf{x}\|_\infty\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 \\ &\le 32C^{-1}C_2\phi^{-2}\log(p)\big(\lambda_1 s^{1/2} + \lambda_2\|\boldsymbol{\theta}_0\|_2\big)^2/\lambda_1. \end{aligned}$$

Then $|\mathbf{x}^T(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})| \le C_3\Lambda_n$ with positive constant $C_3 = 32C^{-1}C_2\phi^{-2}$. Thus we have

$$\begin{aligned} &P\big(\mathbf{x}^T\boldsymbol{\theta}_0 \le \mathbf{x}^T(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})|\mathcal{C}_1, \mathcal{E}_2 \cap \mathcal{E}_3\big) \\ &\quad \le P\big(\mathbf{x}^T\boldsymbol{\theta}_0 \le C_3\Lambda_n|\mathcal{C}_1, \mathcal{E}_2 \cap \mathcal{E}_3\big) \\ &\quad = P\big(\mathbf{x}^T\boldsymbol{\theta}_0 \le C_3\Lambda_n|\mathcal{C}_1, \mathcal{E}_3\big) = \frac{P(\mathbf{x}^T\boldsymbol{\theta}_0 \le C_3\Lambda_n, \mathcal{E}_3|\mathcal{C}_1)}{P(\mathcal{E}_3|\mathcal{C}_1)} \\ &\quad \le \frac{P(\mathbf{x}^T\boldsymbol{\theta}_0 \le C_3\Lambda_n|\mathcal{C}_1)}{P(\mathcal{E}_3|\mathcal{C}_1)} = \frac{F_1(C_3\Lambda_n)}{P(\mathcal{E}_3|\mathcal{C}_1)}, \end{aligned}$$

where $F_1(\cdot)$ is the cumulative distribution function of $\mathbf{x}^T\boldsymbol{\theta}_0|\mathcal{C}_1$. This inequality, together with (34), entails

$$R_n(2|1) \le \frac{F_1(C_3\Lambda_n)}{P(\mathcal{E}_3|\mathcal{C}_1)} + P\big(\mathcal{E}_2^c \cup \mathcal{E}_3^c\big).$$

By the definition of $F_1(\cdot)$, we have $R(2|1) = F_1(0)$. Thus

$$\begin{aligned} &R_n(2|1) - R(2|1) \\ &\quad \le \frac{F_1(C_3\Lambda_n) - F_1(0)}{P(\mathcal{E}_3|\mathcal{C}_1)} + \left[\frac{1}{P(\mathcal{E}_3|\mathcal{C}_1)} - 1\right]F_1(0) + P\big(\mathcal{E}_2^c \cup \mathcal{E}_3^c\big). \end{aligned} \tag{35}$$

From Condition 7, we have $0 \le \Lambda_n < \epsilon_0$ when n is sufficiently large and $C_3\Lambda_n = o(1)$. It follows that $F_1(C_3\Lambda_n) - F_1(0) = F'(\Lambda_n^*)\Lambda_n \le C_3 r_n \Lambda_n$ where Λ_n^* is between 0 and $C_3\Lambda_n$. In view of (32), we have

$$P(\mathcal{E}_3|\mathcal{C}_1) = P\Big\{\max_{1\le j\le p}|Z_j^{(1)}| \le C_2\sqrt{\log(p)}\Big\} = O\big(p^{-c_1}\big).$$

Combining this with (33) and (35) entails

$$\begin{aligned} R_n(2|1) - R(2|1) &\le \frac{C_3 r_n \Lambda_n}{1 - O(p^{-c_1})} + \left|\frac{1}{1 - O(p^{-c_1})} - 1\right| + O\big(p^{-c_1}\big) + O\big(p^{-c_1}\big) \\ &= O(r_n\Lambda_n) + O\big(p^{-c_1}\big). \end{aligned}$$

Similarly, we can show that $R_n(1|2) \le R(1|2) + O(r_n\Lambda_n) + O(p^{-c_1})$. Combining these two results completes the proof of Theorem 3.

Acknowledgments. The authors sincerely thank the Co-Editor, Associate Editor and two referees for their valuable comments that helped improve the article substantially.

SUPPLEMENTARY MATERIAL

Supplement to "Innovated interaction screening for high-dimensional nonlinear classification" (DOI: 10.1214/14-AOS1308SUPP; .pdf). We provide additional lemmas and technical proofs.

REFERENCES

[1] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[2] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[3] BUNEA, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat.* **2** 1153–1194. MR2461898

[4] CAI, T. and LIU, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* **106** 1566–1577. MR2896857

[5] CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

[6] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[7] FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009

[8] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[9] FAN, Y., JIN, J. and YAO, Z. (2013). Optimal classification in sparse Gaussian graphic model. *Ann. Statist.* **41** 2537–2571. MR3161437

[10] FAN, Y., KONG, Y., LI, D. and ZHENG, Z. (2015). Supplement to "Innovated interaction screening for high-dimensional nonlinear classification." DOI:10.1214/14-AOS1308SUPP.

[11] FAN, Y. and LV, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *J. Amer. Statist. Assoc.* **108** 1044–1061. MR3174683

[12] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[13] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

[14] HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357

[15] HEBIRI, M. and VAN DE GEER, S. (2011). The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electron. J. Stat.* **5** 1184–1226. MR2842904

[16] JIANG, B. and LIU, J. S. (2014). Variable selection for general index models via sliced inverse regression. *Ann. Statist.* **42** 1751–1786. MR3262467

[17] JIN, J. (2012). Comment: "Estimating false discovery proportion under arbitrary covariance dependence" [MR3010887]. *J. Amer. Statist. Assoc.* **107** 1042–1045. MR3010891

[18] KOOPERBERG, C., LEBLANC, M., DAI, J. Y. and RAJAPAKSE, I. (2009). Structures and assumptions: Strategies to harness gene × gene and gene × environment interactions in GWAS. *Statist. Sci.* **24** 472–488. MR2779338

[19] LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567

[20] MAI, Q., ZOU, H. and YUAN, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42. MR2899661

[21] PAN, W., BASU, S. and SHEN, X. (2011). Adaptive tests for detecting gene–gene and gene–environment interactions. *Human Heredity* **72** 98–109.

[22] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

[23] SHAO, J., WANG, Y., DENG, X. and WANG, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* **39** 1241–1265. MR2816353

[24] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

[25] TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99** 6567–6572.

[26] VAN'T VEER, L. J., DAI, H., VAN DE VIJVER, M. J., HE, Y. D., HART, A. A., MAO, M., PETERSE, H. L., VAN DER KOOY, K., MARTON, M. J., WITTEVEEN, A. T., SCHREIBER, G. J., KERKHOVEN, R. M., ROBERTS, C., LINSLEY, P. S., BERNARDS, R. and FRIEND, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415** 530–536.

[27] VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809

[28] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

[29] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

[30] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[31] ZHANG, T. and ZOU, H. (2014). Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika* **101** 103–120. MR3180660

[32] ZHU, J. and HASTIE, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5** 427–443.

[33] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

Y. FAN
D. LI
DATA SCIENCES AND OPERATIONS DEPARTMENT
MARSHALL SCHOOL OF BUSINESS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089
USA
E-MAIL: fanyingy@marshall.usc.edu
daojili@marshall.usc.edu

Y. KONG
DEPARTMENT OF PREVENTIVE MEDICINE
KECK SCHOOL OF MEDICINE
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089
USA
E-MAIL: yinfeiko@usc.edu

Z. ZHENG
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089
USA
E-MAIL: zeminzhe@usc.edu