Journal of Machine Learning Research 1 (2017) 1-48 Submitted 4/00; Published 10/00

Covariance-engaged Classification of Sets via Linear Programming

Zhao Ren ZREN@PITT.EDU
Sungkyu Jung SUNGKYU@PITT.EDU
Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260, USA

Xingye Qiao QIAO@MATH.BINGHAMTON.EDU
Department of Mathematical Sciences
Binghamton University
State University of New York
Binghamton, NY, 13902-6000, USA

Editor:

Abstract

Set classification aims to classify a set of observations as a whole, as opposed to classifying individual observations separately. To formally understand the unfamiliar concept of binary set classification, we first investigate the optimal decision rule under the normal distribution, which utilizes the empirical covariance of the set to be classified. We show that the number of observations in the set plays a critical role in bounding the Bayes risk. Under this framework, we further propose new methods of set classification. For the case where only a few parameters of the model drive the difference between two classes, we propose a computationally-efficient approach to parameter estimation using linear programming, leading to the Covariance-engaged LInear Programming Set (CLIPS) classifier. The convergence rates of estimation errors and risk of the CLIPS classifier are established to show that having multiple observations in a set leads to faster convergence rates, compared to the standard classification situation in which there is only one observation in the set. The applicable domains in which the CLIPS performs better than competitors are highlighted in a comprehensive simulation study. Finally, we illustrate the usefulness of the proposed methods in classification of real image data in histopathology.

Keywords: Bayes risk, ℓ_1-minimization, Quadratic discriminant analysis, Set classification, Sparsity

1. Introduction

Classification is a useful tool in statistical learning with applications in many important fields. A classification method aims to train a classification rule based on the training data to classify future observations. There exists a large literature for the classification problem. Some popular methods include linear discriminant analyses, quadratic discriminant analyses, logistic regressions, support vector machines, neural nets and classification trees. Traditionally, the task at hand is to classify an observation into a class label.

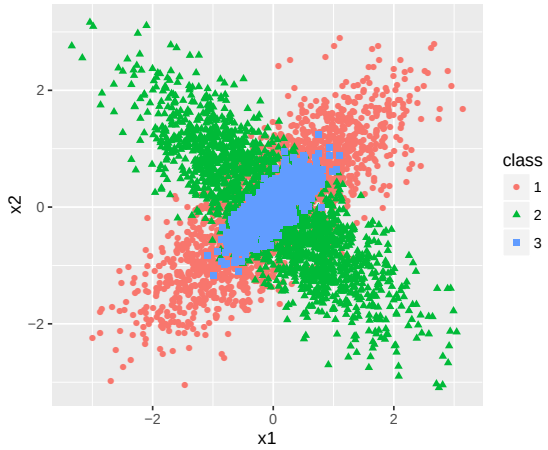

Figure 1: A 2-dimensional toy example showing classes with no difference in the mean or the marginal variance.

Advances in technology have eased the production of a large amount of data in various areas such as healthcare and manufacturing industries. Oftentimes, multiple samples collected from the same object are available. For example, it has become cheaper to obtain multiple tissue samples from a single patient in cancer prognosis (Miedema et al., 2012). A statistical task herein is to classify the whole set of observations from a single patient to normal or cancerous group. Such a problem was seen in the image-based pathology literature (Samsudin and Bradley, 2010; Wang et al., 2010; Cheplygina et al., 2015). In statistics, it was coined as *set classification* by Jung and Qiao (2014). In the machine learning literature, the term multiple-instance learning (Maron and Lozano-Pérez, 1998; Chen et al., 2006; Ali and Shah, 2010) has been used to describe a similar problem.

While conventional classification methods predict a class label for each observation, care is needed in generalizing those for set classification. In principle, more observations should ease the task at hand. Moreover, higher-order statistics such as variances and covariances can now be exploited to help classification. Our approach to set classification is to use the extra information, available to us only when there are multiple observations. To elucidate this idea, we illustrate samples from three classes in Fig. 1. All three classes have the same mean, and Classes 1 and 2 have the same marginal variances. Classifying a single observation near the mean to any of these distributions seems difficult. On the other hand, classifying several independent observations from the same class should be much easier. In particular, a set classification method needs to incorporate the difference in covariances to differentiate these classes.

In this work, we study a binary set classification framework, where a set of observations $\mathcal{X} = \{X_1, \ldots, X_M\}$ is classified to either $\mathcal{Y} = 1$ or $\mathcal{Y} = 2$. In particular, we propose set classifiers that extend quadratic discriminant analysis to the set classification setting, and are designed to work well in set-classification of high-dimensional data whose distributions are similar to those in Fig. 1.

To provide a fundamental understanding of the set classification problem, we establish the Bayesian optimal decision rule under normality and homogeneity assumptions. This Bayes rule utilizes the covariance structure of the testing set of future observations. We show in Section 2 that it becomes much easier to make accurate classification for a set when the set size, m_0, increases. In particular, we demonstrate that the Bayes risk can be reduced exponentially in the set size m_0. To the best of our knowledge, this is the first formal theoretical framework for set classification problems in the literature.

Built upon the Bayesian optimal decision rule, we propose new methods of set classification. For the situation where the dimension p of the feature vectors is much smaller than the total number of training samples, we demonstrate that a simple plug-in classifier leads to satisfactory risk bounds similar to the Bayes risk. Again, a large set size plays a key role in significantly reducing the risk. In high-dimensional situations where the number of parameters to be estimated ($\approx p^2$) is large, we make an assumption that only a few parameters drive the difference of two classes. With this sparsity assumption, we propose to estimate the parameters in the classifier via linear programming, and the resulting classifiers are called Covariance-engaged LInear Programming Set (CLIPS) classifiers. Specifically, the quadratic and linear parameters in the Bayes rule can be efficiently estimated under the sparse structure, thanks to the extra observations in the training set due to having sets of observations. Our estimation approaches are closely related to and built upon the successful estimation strategies in Cai et al. (2011) and Cai and Liu (2011). In estimation of the constant parameter, we perform a logistic regression with only one unknown, given the estimates of quadratic and linear parameters. This allows us to implement CLIPS classifier with high computation efficiency.

We provide a thorough study of theoretical properties of CLIPS classifiers and establish an oracle inequality in terms of the excess risk, in Section 4. In particular, the estimates from CLIPS are shown to be consistent and the strong signals are always selected with high probability in high dimensions. Moreover, the excess risk can be reduced by having more observations in a set.

In the conventional classification problem where $m_0 = 1$, a special case of the proposed CLIPS classifier becomes a new sparse quadratic discriminant analysis (QDA) method (cf. Fan et al., 2015, 2013; Li and Shao, 2015). As a byproduct of our theoretical study, we show that the new QDA method enjoys better theoretical properties compared to state-of-the-art sparse QDA methods such as the one recently developed by Fan et al. (2015).

The advantages of our set classifiers are further demonstrated in comprehensive simulation studies. Moreover, we provide an application to histopathology in classifying sets of nucleus images to normal and cancerous tissues in Section 5. Proofs of main results and technical lemmas can be found in the supplementary material.

2. Set Classification

We consider a binary set-classification problem. The training sample $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N$ contains N sets of observations. Each set, $\mathcal{X}_i = \{X_{i1}, X_{i2}, \ldots, X_{iM_i}\} \subset \mathbb{R}^p$, corresponds to one object, and is assumed to be from one of the two classes. The corresponding class label is denoted by $\mathcal{Y}_i \in \{1, 2\}$. The number of observations within the ith set is denoted by M_i and can be different among different sets. Given a new set of observations $(\mathcal{X}^\dagger, \mathcal{Y}^\dagger)$, the

goal of set classification is to predict $\mathcal{Y}^\dagger$ accurately based on $\mathcal{X}^\dagger$ using a classification rule $\phi(\cdot) \in \{1, 2\}$ trained on the training sample.

We assume that the sets in each class are homogeneous in the sense that all the observations in a class, regardless of the set membership, follow the same distribution independently. Specifically, we assume both the N sets $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N$ and the new set $(\mathcal{X}^\dagger, \mathcal{Y}^\dagger)$ are generated in the same way as $(\mathcal{X}, \mathcal{Y})$ independently. To describe the generating process of $(\mathcal{X}, \mathcal{Y})$, we denote the marginal class probabilities by $\pi_1 = \mathrm{pr}(\mathcal{Y} = 1)$ and $\pi_2 = \mathrm{pr}(\mathcal{Y} = 2)$, and the marginal distribution of the set size M by p_M. We assume that the random variables M and $\mathcal{Y}$ are independent. In other words, the class membership $\mathcal{Y}$ can not be predicted just based on the set size M. Conditioned on $M = m$ and $\mathcal{Y} = y$, observations $X_1, X_2, \ldots, X_M$ in the set $\mathcal{X}$ are independent and each distributed as f_y.

2.1 Covariance-engaged Set Classifiers

Suppose that there are $M^\dagger = m$ observations in the set $\mathcal{X}^\dagger = \{X_1^\dagger, \ldots, X_m^\dagger\}$ that is to be classified (called testing set), and its true class label is $\mathcal{Y}^\dagger$. The Bayes optimal decision rule classifies the set $\mathcal{X}^\dagger = \{x_1, \ldots, x_m\}$ to Class 1 if the conditional class probability of Class 1 is greater than that of Class 2, that is, $\mathrm{pr}(\mathcal{Y}^\dagger = 1 \mid M^\dagger = m,\ X_j^\dagger = x_j,\ j = 1, \ldots, m) > 1/2$. This is equivalent to $\pi_1 p_M(m) \prod_{j=1}^m f_1(x_j) > \pi_2 p_M(m) \prod_{j=1}^m f_2(x_j)$, due to Bayes theorem and the independence assumption among $\mathcal{Y}^\dagger$ and $M^\dagger$. Let us now assume that the conditional distributions are both normal, that is, $f_1 \sim N(\mu_1, \Sigma_1)$ and $f_2 \sim N(\mu_2, \Sigma_2)$. Then the Bayes optimal decision depends on the quantity

$$\begin{aligned} g(x_1, \ldots, x_m) &= \frac{1}{m} \log \left\{ \frac{\pi_1 p_M(m) \prod_{j=1}^m f_1(x_j)}{\pi_2 p_M(m) \prod_{j=1}^m f_2(x_j)} \right\} \\ &= \frac{1}{m} \log(\pi_1/\pi_2) - \frac{1}{2} \log(|\Sigma_1|/|\Sigma_2|) - \frac{1}{2} \mu_1^T \Sigma_1^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma_2^{-1} \mu_2 \\ &\quad + (\Sigma_1^{-1} \mu_1 - \Sigma_2^{-1} \mu_2)^T \bar{x} + \frac{1}{2} \bar{x}^T (\Sigma_2^{-1} - \Sigma_1^{-1}) \bar{x} + \frac{1}{2} \mathrm{tr}\{(\Sigma_2^{-1} - \Sigma_1^{-1}) S\}. \quad (1) \end{aligned}$$

Here $|\Sigma_k|$ denotes the determinant of the matrix Σ_k for $k = 1, 2$, $\bar{x} = \sum_{j=1}^m x_j/m$ and $S = \sum_{j=1}^m (x_j - \bar{x})(x_j - \bar{x})^T/m$ are the sample mean and sample covariance of the testing set. Note that the realization $\mathcal{X}^\dagger = \{x_1, x_2, \ldots, x_m\}$ implies both the number of observations m and the i.i.d. observations x_j for $j = 1, \ldots, m$. The Bayes rule can be expressed as

$$\begin{aligned} \phi_B(\mathcal{X}^\dagger) &= 2 - \mathbb{1}\{g(x_1, \ldots, x_m) > 0\}, \text{ where} \quad (2) \\ g(x_1, \ldots, x_m) &= \frac{1}{m} \log(\pi_1/\pi_2) + \beta_0 + \beta^T \bar{x} + \bar{x}^T \nabla \bar{x}/2 + \mathrm{tr}(\nabla S)/2, \end{aligned}$$

in which the constant coefficient $\beta_0 = \{-\log(|\Sigma_1|/|\Sigma_2|) - \mu_1^T \Sigma_1^{-1} \mu_1 + \mu_2^T \Sigma_2^{-1} \mu_2\}/2 \in \mathbb{R}$, the linear coefficient vector $\beta = \Sigma_1^{-1} \mu_1 - \Sigma_2^{-1} \mu_2 \in \mathbb{R}^p$ and the quadratic coefficient matrix $\nabla = \Sigma_2^{-1} - \Sigma_1^{-1} \in \mathbb{R}^{p \times p}$. The Bayes rule ϕ_B under the normal assumption in (2) uses the summary statistics m, $\bar{x}$ and S of $\mathcal{X}^\dagger$.

We refer to (2) and any estimated version of it as a covariance-engaged set classifier. In Section 3, several estimation approaches for β_0, β and ∇ will be proposed. In this section, we further discuss a rationale for considering (2).

The covariance-engaged set classifier (2) resembles the conventional QDA classifier. As a natural alternative to (2), one may consider the sample mean $\bar{x}$ as a representative of the testing set and apply QDA to $\bar{x}$ directly to make a prediction. In other words, one is about to classify this single observation $\bar{x}$ to one of the two normal distributions, that is, $f_1' \sim N(\mu_1, \Sigma_1/m)$ and $f_2' \sim N(\mu_2, \Sigma_2/m)$. This simple idea leads to

$$\begin{aligned} \phi_{B,\bar{x}}(\mathcal{X}^\dagger) &= 2 - \mathbb{1}\{g_{\mathrm{QDA}}(\bar{x}) > 0\}, \text{ where} \\ g_{\mathrm{QDA}}(\bar{x}) &= \frac{1}{m}\log(\pi_1/\pi_2) + \beta_0' + \beta^T\bar{x} + \bar{x}^T\nabla\bar{x}/2, \end{aligned} \tag{3}$$

in which $\beta_0' = \{-\frac{1}{m}\log(|\Sigma_1|/|\Sigma_2|) - \mu_1^T\Sigma_1^{-1}\mu_1 + \mu_2^T\Sigma_2^{-1}\mu_2\}/2$. One major difference between (2) and (3) is that the term $\mathrm{tr}(\nabla S)/2$ is absent from (3). Indeed, the advantage of (2) over (3) comes from the extra information in the sample covariance S of $\mathcal{X}^\dagger$. In the regular classification setting, (2) coincides with (3) since $\mathrm{tr}(\nabla S)/2$ vanishes when $\mathcal{X}^\dagger$ is a singleton.

Given multiple observations in the testing set, another natural approach is a majority vote applied to the QDA decisions of individual observations:

$$\phi_{MV}(\mathcal{X}^\dagger) = 2 - \mathbb{1}\left\{\frac{1}{m}\sum_{j=1}^{m}\mathrm{sign}[g_{\mathrm{QDA}}(x_j)] > 0\right\}, \tag{4}$$

where $\mathrm{sign}(t) = 1, 0, -1$ for $t > 0$, $t = 0$ and $t < 0$ respectively. In contrast, since $g(\mathcal{X}^\dagger) = \frac{1}{m}\sum_{j=1}^{m} g_{\mathrm{QDA}}(x_j)$, our classifier (2) predicts the class label by a weighted vote of individual QDA decisions. In this sense, the majority voting scheme (4) can be viewed as a discretized version of (2). In Section 5, we demonstrate that our set classifier (2) performs significantly better than (4).

Remark 1 *We have assumed that M and $\mathcal{Y}$ are independent in the setting. In fact, this assumption is not essential and can be relaxed. In a more general setting, there can be two different distributions of M, $p_{M1}(m)$ and $p_{M2}(m)$ conditional on $\mathcal{Y} = 1$ and $\mathcal{Y} = 2$ respectively. Our analysis throughout the paper remains the same except that they would replace two identical factors $p_M(m)$ in the first equality of (1). If $p_{M1}(m)$ and $p_{M2}(m)$ are dramatically different, then the classification is easier as one can make decision based on the observed value of m. In this paper, we only consider the more difficult setting where $\mathcal{Y}$ and M are independent.*

2.2 Bayes Risk

We show below an advantage of having a set of observations for prediction, compared to having a single observation. For this, we suppose for now that the parameters μ_k and Σ_k, $k = 1, 2$, are known and make the following assumptions. Denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ as the greatest and smallest eigenvalues of a symmetric matrix A.

Condition 1 *The spectrum of Σ_k is bounded below and above: there exists some universal constant $C_e > 0$ such that $C_e^{-1} \le \lambda_{\min}(\Sigma_k) \le \lambda_{\max}(\Sigma_k) \le C_e$ for $k = 1, 2$.*

Condition 2 *The support of p_M is bounded between $c_m m_0$ and $C_m m_0$, where c_m and C_m are universal constants and $m_0 = \mathbb{E}(M)$. In other words, $p_M(a) = 0$ for any integer*

$a < c_m m_0$ *or* $> C_m m_0$. *The set size* m_0 *can be large or growing when a sequence of models are considered.*

Condition 3 *The prior class probability is bounded away from* 0 *and* 1*: there exists a universal constant* $0 < C_\pi < 1/2$ *such that* $C_\pi \le \pi_1, \pi_2 \le 1 - C_\pi$.

We denote $R_{Bk} = \mathrm{pr}(\phi_B(\mathcal{X}^\dagger) \neq k \mid \mathcal{Y}^\dagger = k)$ as the risk of the Bayes classifier (2) given $\mathcal{Y}^\dagger = k$. Let $\delta = \mu_2 - \mu_1$. For a matrix $B \in \mathbb{R}^{p\times p}$, we denote $\|B\|_F = (\sum_{i=1}^p \sum_{j=1}^p B_{ij}^2)^{1/2}$ as its Frobenius norm, where B_{ij} is its ijth element. For a vector $a \in \mathbb{R}^p$, we denote $\|a\| = (\sum_{i=1}^p a_i^2)^{1/2}$ as its ℓ_2 norm. The quantity $D_p = (\|\nabla\|_F^2 + \|\delta\|^2)^{1/2}$ plays an important role in deriving a convergence rate of the Bayes risk $R_B = \pi_1 R_{B1} + \pi_2 R_{B2}$. Although the Bayes risk does not have a closed form, we show that under mild assumptions, it converges to zero at a rate on the exponent.

Theorem 2 *Suppose that Conditions 1-3 hold. If* $D_p^2 m_0$ *is sufficiently large, then* $R_B \le 4\exp\left(-c' m_0 D_p^2\right)$ *for some small constant* $c' > 0$ *depending on* C_e, c_m *and* C_π *only. In particular, as* $D_p^2 m_0 \to \infty$, *we have* $R_B \to 0$.

The significance of having a set of observations is illustrated by this fundamental theorem. When $p_M(1) = 1$, which implies $M^\dagger \equiv 1$ and $m_0 = 1$, Theorem 2 provides a Bayes risk bound $R_B \le 4\exp\left(-c' D_p^2\right)$ for the theoretical QDA classifier in the regular classification setting. To guarantee a small Bayes risk for QDA, it is clear that D_p^2 must be sufficiently large. In comparison, for the set classification to be successful, we may allow D_p^2 to be very close to zero, as long as $m_0 D_p^2$ is sufficiently large. The Bayes risk of ϕ_B can be reduced exponentially in m_0 because of the extra information from the set.

Remark 3 *One can apply theoretical QDA to the sample mean* $\bar{x}$*, leading to* $\phi_{B,\bar{x}}$ *(3). A similar analysis to Theorem 2 implies that its risk is bounded by* $4\exp\left(-c\left(\|\nabla\|_F^2 + m_0\|\delta\|^2\right)\right)$ *with some universal constant* $c > 0$ *whenever* $\|\nabla\|_F^2 + m_0\|\delta\|^2$ *is sufficiently large. Compared to the result in Theorem 2, it needs a stronger assumption but has a slower rate of convergence when the mean difference* $m_0\|\delta\|^2$ *is dominated by the covariance difference* $\|\nabla\|_F^2$. *After all, this natural* $\bar{x}$*-based classification rule only relies on the first moment of the data set* $\mathcal{X}^\dagger$ *while the sufficient statistics, the first two moments, are fully used by the covariance-engaged classifier in (2).*

3. Methodologies

We now consider estimation procedures for ϕ_B based on N training sets $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N$. In Section 3.1, we first consider a moderate-dimensional setting where $p \le c_0 m_0 N$ with a sufficiently small constant $c_0 > 0$. In this case we apply a naive plug-in approach using natural estimators of the parameters π_k, μ_k and Σ_k. A direct estimation approach using linear programming, suitable for high-dimensional data, is introduced in Section 3.2. Hereafter, $p = p(N)$ and $m_0 = m_0(N)$ are considered as functions of N as N grows.

3.1 Naive Estimation Approaches

The prior class probabilities π_1 and π_2 can be consistently estimated by the class proportions in the training data, $\hat{\pi}_1 = N_1/N$ and $\hat{\pi}_2 = N_2/N$, where $N_k = \sum_{i=1}^N \mathbb{1}\{\mathcal{Y}_i = k\}$. Let $n_k = \sum_{i=1}^N M_i \mathbb{1}\{\mathcal{Y}_i = k\}$ denote the total sample size for Class $k = 1, 2$. The set membership is ignored at the training stage, due to the homogeneity assumption. Note n_k, $n_1 + n_2$ and N_k are random while N is deterministic. One can obtain consistent estimators of μ_k and Σ_k based on the training data and plug them in (2). It is natural to use the maximum likelihood estimators given n_k,

$$\hat{\mu}_k = \sum_{(i,j):\mathcal{Y}_i=k} X_{ij}/n_k \text{ and } \hat{\Sigma}_k = \sum_{(i,j):\mathcal{Y}_i=k} \{(X_{ij} - \hat{\mu}_k)(X_{ij} - \hat{\mu}_k)^T\}/n_k. \tag{5}$$

For classification of $\mathcal{X}^\dagger = \{X_1^\dagger, \ldots, X_{M^\dagger}^\dagger\}$ with $M^\dagger = m$, $X_i^\dagger = x_i$, the set classifier (2) is estimated by

$$\hat{\phi}(\mathcal{X}^\dagger) = 2 - \mathbb{1}\left\{\frac{1}{m}\log(\hat{\pi}_1/\hat{\pi}_2) + \hat{\beta}_0 + \hat{\beta}^T\bar{x} + \bar{x}^T\hat{\nabla}\bar{x}/2 + \mathrm{tr}(\hat{\nabla}S)/2 > 0\right\}, \tag{6}$$

where $\hat{\beta}_0 = -\frac{1}{2}\left\{\log(|\hat{\Sigma}_1|/|\hat{\Sigma}_2|) - \hat{\mu}_1^T\hat{\Sigma}_1^{-1}\hat{\mu}_1 + \hat{\mu}_2^T\hat{\Sigma}_2^{-1}\hat{\mu}_2\right\}$, $\hat{\beta} = \hat{\Sigma}_1^{-1}\hat{\mu}_1 - \hat{\Sigma}_2^{-1}\hat{\mu}_2$ and $\hat{\nabla} = \hat{\Sigma}_2^{-1} - \hat{\Sigma}_1^{-1}$. In (6) we have assumed $p < n_k$ so that $\hat{\Sigma}_k$ is invertible.

The generalization error of set classifier (6) is $\hat{R} = \pi_1\hat{R}_1 + \pi_2\hat{R}_2$ where $\hat{R}_k = \mathrm{pr}(\hat{\phi}(\mathcal{X}^\dagger) \neq k \mid \mathcal{Y}^\dagger = k)$. The classifier itself depends on the training data $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N$ and hence is random. In the equation above, pr is understood as the conditional probability given the training data. Theorem 4 reveals a theoretical property of $\hat{R}$ in a moderate-dimensional setting which allows p, N, m_0 to grow jointly. This includes the traditional setting in which p is fixed.

Theorem 4 *Suppose that Conditions 1-3 hold. For any fixed $L > 0$, if $D_p^2 m_0 \geq C_0$ for some sufficiently large $C_0 > 0$ and $p \leq c_0 N m_0$, $p^2/(N m_0 D_p^2) \leq c_0$, $\log p \leq c_0 N$ for some sufficiently small constant $c_0 > 0$, then with probability at least $1 - O(p^{-L})$ we have $\hat{R} \leq 4\exp\left(-c' m_0 D_p^2\right)$ for some small constant $c' > 0$ depending on C_π, c_m, L and C_e.*

In Theorem 4, large values of m_0 not only relax the assumption on D_p but also reduce the Bayes risk exponentially in m_0 with high probability. A similar result for QDA, where $M_i = M^\dagger \equiv 1$ and $m_0 = 1$, was obtained in Li and Shao (2015) under a stronger assumption $p^2/(N D_p^2) \to 0$.

For the high-dimensional data where $p = p(N) \gg N m_0$ and hence $p > n_k$ with probability 1 for $k = 1, 2$ by Condition 2, it is problematic to plug in the estimators (5) since $\hat{\Sigma}_k$ is rank deficient with probability 1. A simple remedy is to use a diagonalized or enriched version of $\hat{\Sigma}_k$, defined by $\hat{\Sigma}_{k(d)} = \mathrm{diag}\{(\hat{\sigma}_{k,ii})_{i=1,\ldots,p}\}$ or $\hat{\Sigma}_{k(e)} = \hat{\Sigma}_k + \delta I_p$, where $\delta > 0$ and I_p is a $p \times p$ identity matrix. Both $\hat{\Sigma}_{k(d)}$ and $\hat{\Sigma}_{k(e)}$ are invertible. However, to our best knowledge, no theoretical guarantee has been obtained without some structural assumptions.

3.2 A Direct Approach via Linear Programming

To have reasonable classification performance in high-dimensional data analysis, one usually has to take advantage of certain extra information of the data or model. There are often cases where only a few elements in $\nabla = \Sigma_2^{-1} - \Sigma_1^{-1}$ and $\beta = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2$ truly drive the difference between the two classes. A naive plug-in method proposed in Section 3.1 has ignored such potential structure of the data. We assume that both ∇ and β are known to be sparse such that only a few elements of those are nonzero. In light of this, the Bayes decision rule (2) implies the dimension of the problem can be significantly reduced, which makes consistency possible even in the high-dimensional setting.

We propose to directly estimate the quadratic term ∇, the linear term β and the constant β_0 coefficients respectively, taking advantage of the assumed sparsity. As the estimates are efficiently calculated by linear programming, the resulting classifiers are called Covariance-engaged Linear Programming Set (CLIPS) classifiers.

We first deal with the estimation of the quadratic term $\nabla = \Sigma_2^{-1} - \Sigma_1^{-1}$, which is the difference between the two precision matrices. We use some key techniques developed in the literature of precision matrix estimation (cf. Meinshausen and Bühlmann, 2006; Bickel and Levina, 2008; Friedman et al., 2008; Yuan, 2010; Cai et al., 2011; Ren et al., 2015). These methods estimate a single precision matrix with a common assumption that the underlying true precision matrix is sparse in some sense. For the estimation of the difference, we propose to use a two-step thresholded estimator.

As the first step, we adopt the CLIME estimator (Cai et al., 2011) to obtain initial estimators $\tilde{\Omega}_1$ and $\tilde{\Omega}_2$ of the precision matrices Σ_1^{-1} and Σ_2^{-1}. Let $\|B\|_1 = \sum_{i,j} |B_{ij}|$ and $\|B\|_\infty = \max_{i,j} |B_{ij}|$ be the vector ℓ_1 norm and vector supnorm of a $p \times p$ matrix B respectively. The CLIME estimators are defined as

$$\tilde{\Omega}_k = \underset{\Omega \in \mathbb{R}^{p\times p}}{\operatorname{argmin}} \|\Omega\|_1 \text{ subject to } \|\hat{\Sigma}_k \Omega - I\|_\infty < \lambda_{1,N},\ k = 1, 2, \tag{7}$$

for some $\lambda_{1,N} > 0$.

Having obtained $\tilde{\Omega}_1$ and $\tilde{\Omega}_2$, in the second step, we take a thresholding procedure on their difference, followed by a symmetrization to obtain our final estimator $\tilde{\nabla} = (\tilde{\nabla}_{ij})$ where

$$\tilde{\nabla}_{ij} = \min\{\breve{\nabla}_{ij}, \breve{\nabla}_{ji}\}, \breve{\nabla}_{ij} = (\tilde{\Omega}_{2,ij} - \tilde{\Omega}_{1,ij}) \mathbb{1}\Big\{\Big|\tilde{\Omega}_{2,ij} - \tilde{\Omega}_{1,ij}\Big| > \lambda'_{1,N}\Big\}, \tag{8}$$

for some thresholding level $\lambda'_{1,N} > 0$.

Although this thresholded CLIME difference estimator is obtained by first individually estimating Σ_k^{-1}, we emphasize that the estimation accuracy only depends on the sparsity of their difference ∇ rather than the sparsity of either Σ_1^{-1} or Σ_2^{-1} under a relatively mild bounded matrix ℓ_1 norm condition. It is possible that $\nabla = 0$, in which case our method has adaptively chosen to be linear. We will show in Theorem 6 in Section 4 that if the true precision matrix difference ∇ is negligible, $\tilde{\nabla} = 0$ with high probability. The computation of $\tilde{\nabla}$ (8) is fast, since the first step (CLIME) can be recast as a linear programming and the second step is a simple thresholding procedure.

Remark 5 *As an alternative, one can also consider a direct estimation of ∇ that does not rely on individual estimates of Σ_k^{-1}. For example, by allowing some deviations from the identity $\Sigma_1 \nabla \Sigma_2 - \Sigma_1 + \Sigma_2 = 0$, Zhao et al. (2014) proposed to minimize the vector ℓ_1 norm of ∇.*

Specifically, they proposed $\tilde{\nabla}^{ZCL} \in \operatorname{argmin}_B \|B\|_1$ *subject to* $\|\hat{\Sigma}_1 B \hat{\Sigma}_2 - \hat{\Sigma}_1 + \hat{\Sigma}_2\|_\infty \le \lambda''_{1,n}$, *where* $\lambda''_{1,n}$ *is some thresholding level. This method, however, is computationally expensive (as it has* $O(p^2)$ *number of linear constraints when casted to linear programming) and can only handle relatively small size of* p. *We chose to use (8) mainly because of fast computation.*

Next we consider the estimation of the linear coefficient vector $\beta = \beta_1 - \beta_2$, where $\beta_k = \Sigma_k^{-1}\mu_k$, $k = 1, 2$. In the literature of sparse QDA and sparse LDA, typical sparsity assumptions are placed on $\mu_1 - \mu_2$ and $\Sigma_1 - \Sigma_2$ (see Li and Shao, 2015) or placed on both β_1 and β_2 (see, for instance Cai and Liu, 2011; Fan et al., 2015). In the latter case, β is also sparse as it is the difference of two sparse vectors. For the estimation of β, we propose a new method which directly imposes sparsity on β, without specifying the sparsity for μ_k, Σ_k or β_k except for some relatively mild conditions (see Theorem 9 for details.)

The true parameter β_k satisfies $\Sigma_k \beta_k - \mu_k = 0$. However, due to the rank-deficiency of $\hat{\Sigma}_k$, there are either none or infinitely many θ_k's that satisfy an empirical equation $\hat{\Sigma}_k \theta_k - \hat{\mu}_k = 0$. Here, $\hat{\mu}_k$ and $\hat{\Sigma}_k$ are defined in (5). We relax this constraint and seek a possibly non-sparse pair (θ_1, θ_2) with the smallest ℓ_1 norm difference. We estimate the coefficients β by $\tilde{\beta} = \tilde{\beta}_1 - \tilde{\beta}_2$, where

$$(\tilde{\beta}_1, \tilde{\beta}_2) = \operatorname*{argmin}_{(\theta_1,\theta_2):\|\theta_k\|_1 \le L_1} \|\theta_1 - \theta_2\|_1 \text{ subject to } \|\hat{\Sigma}_k \theta_k - \hat{\mu}_k\|_\infty < \lambda_{2,N},\ k = 1, 2, \tag{9}$$

where L_1 is some sufficiently large constant introduced only to ease theoretical evaluations. In practice, the constraint $\|\theta_k\|_1 \le L_1$ can be removed without affecting the solution. The direct estimation approach for β above shares some similarities with that of Cai and Liu (2011), especially in the relaxed ℓ_∞ constraint. However Cai and Liu (2011) focused on a direct estimation of $\Sigma^{-1}(\mu_2 - \mu_1)$ for linear discriminant analysis in which $\Sigma = \Sigma_1 = \Sigma_2$, while we target on $\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1$ instead. Our procedure (9) can be recast as a linear programming problem (see, for example, Candes and Tao, 2007; Cai and Liu, 2011) and is computationally efficient.

Finally, we consider the estimation of the constant coefficient β_0. The conditional class probability $\eta(x_1, \ldots, x_m) = \operatorname{pr}(\mathcal{Y} = 1 \mid M = m,\ X_i = x_i,\ i = 1, \ldots, m)$ that a set belongs to Class 1 given $\mathcal{X} = \{x_1, \ldots, x_m\}$ can be evaluated by the following logit function,

$$\begin{aligned}\log\left\{\frac{\eta(x_1,\ldots,x_m)}{1-\eta(x_1,\ldots,x_m)}\right\} &= \log\frac{\pi_1}{\pi_2} + \log\left\{\frac{\prod_{i=1}^m f_1(x_i)}{\prod_{i=1}^m f_2(x_i)}\right\} \\ &= \log(\pi_1/\pi_2) + m(\beta_0 + \bar{x}^T\beta + \frac{1}{2}\bar{x}^T\nabla\bar{x} + \frac{1}{2}\operatorname{tr}(\nabla S)),\end{aligned}$$

where $\bar{x}$ and S are the sample mean and covariance of the set $\{x_1, \ldots, x_m\}$ respectively. Having obtained our estimators $\tilde{\nabla}$ and $\tilde{\beta}$ from (8) and (9), and estimated $\hat{\pi}_1$ and $\hat{\pi}_2$ by N_1/N and N_2/N from the training data, we have only a scalar β_0 undecided. We may find an estimate $\tilde{\beta}_0$ by conducting a simple logistic regression with dummy independent variable M_i and offset $\log(\hat{\pi}_1/\hat{\pi}_2) + M_i\left(\bar{X}_i^T\tilde{\beta} + \bar{X}_i^T\tilde{\nabla}\bar{X}_i/2 + \operatorname{tr}(\tilde{\nabla}S_i)/2\right)$ for the ith set of observations in the training data, where M_i, $\bar{X}_i$, and S_i are sample size, sample mean, and

sample covariance of the ith set. In particular, we solve

$$\tilde{\beta}_0 = \underset{\theta_0 \in \mathbb{R}}{\operatorname{argmin}} \ \ell(\theta_0 \mid \{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N, \tilde{\beta}, \tilde{\nabla}), \text{where the log-likelihood is} \tag{10}$$

$$\begin{aligned} &\ell(\theta_0 \mid \{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N, \tilde{\beta}, \tilde{\nabla}) \qquad (11)\\ &= \frac{1}{N}\sum_{i=1}^N \Big((\mathcal{Y}_i - 2) M_i \left(\theta_0 + \frac{\log(\hat{\pi}_1/\hat{\pi}_2)}{M_i} + \bar{X}_i^T \tilde{\beta} + \bar{X}_i^T \tilde{\nabla} \bar{X}_i/2 + \operatorname{tr}(\tilde{\nabla} S_i)/2 \right) \\ &\quad + \log\left[1 + \exp\left\{ M_i \left(\theta_0 + \frac{\log(\hat{\pi}_1/\hat{\pi}_2)}{M_i} + \bar{X}_i^T \tilde{\beta} + \bar{X}_i^T \tilde{\nabla} \bar{X}_i/2 + \operatorname{tr}(\tilde{\nabla} S_i)/2 \right) \right\} \right] \Big) \end{aligned}$$

Since there is only one independent variable in the logistic regression above, the optimization can be easily and efficiently solved.

For the purpose of evaluating theoretical properties, we apply the sample splitting technique (Wasserman and Roeder, 2009; Meinshausen and Bühlmann, 2010). Specifically, we randomly choose the first batch of $N_1/2$ and $N_2/2$ sets from two classes in the training data to obtain estimators $\tilde{\nabla}$ and $\tilde{\beta}$ using (8) and (9). Then $\tilde{\beta}_0$ is estimated based on the second batch along with $\tilde{\nabla}$ and $\tilde{\beta}$ using (10). We plug all the estimators in (8), (9) and (10) into the Bayes decision rule (2) and obtain the CLIPS classifier,

$$\tilde{\phi}(\mathcal{X}^\dagger) = 2 - \mathbb{1}\left\{ \frac{\log(\hat{\pi}_1/\hat{\pi}_2)}{m} + \tilde{\beta}_0 + \tilde{\beta}^T \bar{x} + \bar{x}^T \tilde{\nabla} \bar{x}/2 + \operatorname{tr}(\tilde{\nabla} S)/2 > 0 \right\}, \tag{12}$$

where $\bar{x}$ and S are sample mean and covariance of $\mathcal{X}^\dagger$ and $M^\dagger = m$ is its size.

4. Theoretical Properties

In this section, we derive the theoretical properties of the estimators from (8)–(10) as well as generalization errors for the CLIPS classifier (12).

To establish the statistical properties of the thresholded CLIME difference estimator $\tilde{\nabla}$ defined in (8), we assume that the true quadratic parameter $\nabla = \Sigma_2^{-1} - \Sigma_1^{-1}$ has no more than s_q nonzero entries,

$$\nabla \in \mathcal{FM}_0(s_q) = \{A = (a_{ij}) \in \mathbb{R}^{p\times p}, \text{symmetric} : \sum_{i,j=1}^p \mathbb{1}\{a_{ij} \neq 0\} \leq s_q\} \tag{13}$$

Denote supp(A) as the support of the matrix A. We summarize the estimation error and a subset selection result in the following theorem.

Theorem 6 *Suppose Conditions 1-3 hold. Moreover, assume* $\nabla \in \mathcal{FM}_0(s_q)$, $\|\Sigma_k^{-1}\|_{\ell_1} \leq C_{\ell 1}$ *with some constant* $C_{\ell 1} > 0$ *for* $k = 1, 2$ *and* $\log p \leq c_0 N$ *with some sufficiently small constant* $c_0 > 0$. *Then for any fixed* $L > 0$, *with probability at least* $1 - O(p^{-L})$, *we have that*

$$\begin{aligned} \|\tilde{\nabla} - \nabla\|_\infty &\leq 2\lambda'_{1,N}, \\ \|\tilde{\nabla} - \nabla\|_F &\leq 2\sqrt{s_q}\lambda'_{1,N}, \\ \|\tilde{\nabla} - \nabla\|_1 &\leq 2 s_q \lambda'_{1,N}, \end{aligned}$$

as long as $\lambda_{1,N} \geq CC_{\ell 1}\sqrt{\frac{\log p}{Nm_0}}$ *and* $\lambda'_{1,N} \geq 8C_{\ell 1}\lambda_{1,N}$ *in (8), where* C *depends on* L, C_e, C_π *and* c_m *only. Moreover, we have* $\mathrm{pr}(\mathrm{supp}(\tilde{\nabla}) \subset \mathrm{supp}(\nabla)) = 1 - O(p^{-L})$.

Remark 7 *The parameter space* $\mathcal{FM}_0(s_q)$ *can be easily extended into an entry-wise* ℓ_q *ball or weak* ℓ_q *ball with* $0 < q < 1$ *(Abramovich et al., 2006) and the estimation results in Theorem 6 remain valid with appropriate sparsity parameters. The subset selection result also remains true and the support of* $\tilde{\nabla}$ *contains those important signals of* ∇ *above the noise level* $\sqrt{(\log p)/Nm_0}$. *To simplify the analysis, we only consider* ℓ_0 *balls in this work.*

Remark 8 *Theorem 6 implies that both the error bounds of estimating* ∇ *under vector* ℓ_1 *norm and Frobenius norm rely on the sparsity* s_q *imposed on* ∇ *rather than those imposed on* Σ_2^{-1} *or* Σ_1^{-1}. *Therefore, even if both* Σ_2^{-1} *and* Σ_1^{-1} *are relatively dense, we still have an accurate estimate of* ∇ *as long as* ∇ *is very sparse and* $C_{\ell 1}$ *is not large.*

The proof of Theorem 6, provided in the supplementary material, partially follows from Cai et al. (2011).

Next we assume $\beta = \beta_1 - \beta_2$ is sparse in the sense that it belongs to the s_l-sparse ball,

$$\beta \in \mathcal{F}_0(s_l) = \{\alpha = (a_j) \in \mathbb{R}^p : \sum_{j=1}^{p} \mathbb{1}\{\alpha_j \neq 0\} \leq s_l\}. \tag{14}$$

Theorem 9 gives the rates of convergence of the linear coefficient estimator $\tilde{\beta}$ in (9) under the ℓ_1 and ℓ_2 norms. Both depend on the sparsity of β only rather than that of β_1 or β_2.

Theorem 9 *Suppose Conditions 1-3 hold. Moreover, assume that* $\beta \in \mathcal{F}_0(s_l)$, $\log p \leq c_0 N$, $\|\beta_k\|_1 \leq C_\beta$ *and* $\|\mu_k\| \leq C_\mu$ *with some constants* $C_\beta, C_\mu > 0$ *for* $k = 1, 2$ *and some sufficiently small constant* $c_0 > 0$. *Then for any fixed* $L > 0$, *with probability at least* $1 - O(p^{-L})$, *we have that*

$$\begin{aligned} \|\tilde{\beta} - \beta\|_1 &\leq C'' C_{\ell 1} s_l \lambda_{2,N}, \\ \|\tilde{\beta} - \beta\| &\leq C'' C_{\ell 1} \sqrt{s_l} \lambda_{2,N}, \end{aligned}$$

as long as $\lambda_{2,N} \geq C'\sqrt{\frac{\log p}{Nm_0}}$ *in (9), where* $\max\{\|\Sigma_1^{-1}\|_{\ell_1}, \|\Sigma_2^{-1}\|_{\ell_1}\} \leq C_{\ell 1}$ *and* C'', C' *depend on* $L, C_e, c_m, C_\pi, C_\beta$ *and* C_μ *only.*

Remark 10 *The parameter space* $\mathcal{F}_0(s)$ *can be easily extended into an* ℓ_q *ball or weak* ℓ_q *ball with* $0 < q < 1$ *as well and the results in Theorem 9 remain valid with appropriate sparsity parameters. We only focus on* $\mathcal{F}_0(s)$ *in this paper to ease the analysis.*

Lastly, we derive the rate of convergence for estimating the constant coefficient β_0. Since $\tilde{\beta}_0$ is obtained by maximizing the log-likelihood function after plugging $\tilde{\beta}$ and $\tilde{\nabla}$ in (10), the behavior of our estimator $\tilde{\beta}_0$ critically depends on the accuracy for estimating β and ∇. Theorem 11 provides the result for $\tilde{\beta}_0$ based on certain general initial estimators $\tilde{\beta}$ and $\tilde{\nabla}$ with the following mild condition.

Condition 4 *The expectation of the conditional variance of class label $\mathcal{Y}$ given $\mathcal{X}$ is bounded below, that is, $\mathbb{E}\left(\mathrm{Var}(\mathcal{Y} \mid \mathcal{X})\right) > C_{\log} > 0$, where $C_{\log}$ is some universal constant.*

Theorem 11 *Suppose Conditions 1-4 hold, $\log p \le c_0 N$ with some sufficiently small constant $c_0 > 0$ and $\|\mu_k\| \le C_\mu$ with some constant $C_\mu > 0$ for $k = 1, 2$. Besides, we have some initial estimators $\tilde{\beta}$, $\tilde{\nabla}$, $\hat{\pi}_1$ and $\hat{\pi}_2$ such that $m_0(1 + \sqrt{(\log p)/m_0})(\|\tilde{\beta} - \beta\| + \|\tilde{\nabla} - \nabla\|_1) + \max_{k=1,2} |\pi_k - \hat{\pi}_k| \le C_p$ for some sufficiently small constant $C_p > 0$ with probability at least $1 - O(p^{-L})$. Then, with probability at least $1 - O(p^{-L})$, we have*

$$\left|\tilde{\beta}_0 - \beta_0\right| \le C_\delta \left((\|\tilde{\beta} - \beta\| + \|\tilde{\nabla} - \nabla\|_1)(1 + \sqrt{\frac{\log p}{m_0}}) + \max_{k=1,2} |\pi_k - \hat{\pi}_k|/m_0 + \sqrt{\frac{\log p}{N}} \right),$$

where constant C_δ depends on $L, C_e, C_\pi, C_{\log}, C_\mu, C_m$ and c_m.

Remark 12 *Condition 4 is determined by our data generating process stated in Section 2.1. It is satisfied when the classification problem is non-trivial. For example, it is valid if $\mathrm{pr}\{C' < \mathrm{pr}(\mathcal{Y} = 1 \mid \mathcal{X}) < 1 - C'\} > C$ with some constants C and $C' \in (0, 1)$. As a matter of fact, Condition 4 is weaker than the typical assumption: $C_{\log} < \mathrm{pr}(\mathcal{Y} = 1 \mid \mathcal{X}) < 1 - C_{\log}$ with probability 1 for $\mathcal{X}$, which is often seen in the literature of logistic regression. See, for example, Fan and Lv (2013) and Fan et al. (2015).*

Theorems 6–11 demonstrate the estimation accuracy for the quadratic, linear and constant coefficients in our CLIPS classifier (12) respectively. We conclude this section by establishing an oracle inequality for its generalization error via providing a rate of convergence of the excess risk. To this end, we define the generalization error of CLIPS classifier as $\tilde{R} = \pi_1 \tilde{R}_1 + \pi_2 \tilde{R}_2$, where $\tilde{R}_k = \mathrm{pr}(\tilde{\phi}(\mathcal{X}^\dagger) \neq k \mid \mathcal{Y}^\dagger = k)$ is the probability that a new set observation from Class k is misclassified by the CLIPS classifier $\tilde{\phi}(\mathcal{X}^\dagger)$. Again pr is the conditional probability given the training data $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N$ which $\tilde{\phi}(\mathcal{X}^\dagger)$ depends on.

We introduce some notation related to the Bayes decision rule in (2). Recall that given $M^\dagger = m$, the Bayes decision rule $\phi_B(\mathcal{X}^\dagger)$ solely depends on the sign of the function $g(\mathcal{X}^\dagger) = \frac{1}{m}\log(\pi_1/\pi_2) + \beta_0 + \beta^T \bar{x} + \bar{x}^T \nabla \bar{x}/2 + \mathrm{tr}(\nabla S)/2$. We define by $F_{k,m}$ the conditional cumulative distribution function of the oracle statistic $g(\mathcal{X}^\dagger)$ given that $M^\dagger = m$ and $\mathcal{Y}^\dagger = k$. The upper bound of the first derivatives of $F_{1,m}$ and $F_{2,m}$ for all possible m near 0 is denoted by d_N,

$$d_N = \max_{m \in [c_m m_0, C_m m_0],\ k=1,2} \left\{ \sup_{t \in [-\delta_0, \delta_0]} \left| F'_{k,m}(t) \right| \right\},$$

where δ_0 is any sufficiently small constant. The value of d_N is determined by the generating process and is usually small whenever the Bayes rule performs reasonably well. According to Theorems 6–11, with probability at least $1 - O(p^{-L})$, our estimators satisfy that

$$\Xi_N := (1 + \sqrt{\frac{\log p}{m_0}}) \left(\|\tilde{\beta} - \beta\| + \|\tilde{\nabla} - \nabla\|_1 \right) + \max_{k=1,2} \frac{|\hat{\pi}_k - \pi_k|}{m_0} + \left|\tilde{\beta}_0 - \beta_0\right| = O(\xi_N),$$

where $\xi_N := (1 + \sqrt{(\log p)/m_0})(s_q \lambda'_{1,N} + C_{\ell 1}\sqrt{s_l}\lambda_{2,N}) + \sqrt{(\log p)/N}$. It turns out the quantity $\xi_N d_N$ is the key to obtain the oracle inequality, as in the following condition.

Condition 5 *Suppose* $\xi_N d_N \le c_0$ *and* $m_0(1+\sqrt{(\log p)/m_0})(s_q\lambda'_{1,N} + C_{\ell 1}\sqrt{s_l}\lambda_{2,N}) \le c_0$ *with some sufficiently small constant* $c_0 > 0$.

Theorem 13 below reveals the oracle property of CLIPS classifier and provides a rate of convergence of the excess risk, that is, the generalization error of CLIPS classifier less the Bayes risk R_B defined in Section 2.2.

Theorem 13 *Suppose that the assumptions of Theorems 6 and 9 hold and that Condition 5 also holds. Then with probability at least* $1 - O(p^{-L})$*, we have the oracle inequality*

$$\tilde{R} \le R_B + C_g(\xi_N d_N + p^{-L}),$$

where constant C_g *depends on* $L, C_e, C_\pi, C_{\log}, C_\beta, C_m, c_m$ *and* C_μ *only. In particular, we have* $\tilde{R}$ *converges to the Bayes risk* R_B *in probability as* N *goes to infinity.*

Theorem 13 implies that with high probability, the generalization error of CLIPS classifier is close to the Bayes risk with rate of convergence no slower than $\xi_N d_N$. In particular, whenever the the quantities d_N and $C_{\ell 1}$ are bounded by some universal constant, the thresholding levels $\lambda'_{1,N} = O(\sqrt{\log p/(m_0 N)})$ and $\lambda_{2,N} = O(\sqrt{\log p/(m_0 N)})$ yield the rate of convergence $\xi_N d_N$ in the order of

$$(1+\sqrt{(\log p)/m_0})\sqrt{\log p/(m_0 N)}(s_q + \sqrt{s_l}) + \sqrt{\log p/N}. \tag{15}$$

The advantage of having large m_0 can be understood by investigating (15) as a function of m_0. For this we assume $\log p = o(N)$ and $(\log p)/s^2 = o(1)$, where $s = s_q + \sqrt{s_l}$. Then the leading term of (15) is

$$\begin{cases} \dfrac{\log p}{\sqrt{N}}\sqrt{\dfrac{s^2}{m_0}}, & \text{if } m_0 \le \log p; \\ \sqrt{\dfrac{\log p}{N}}\sqrt{\dfrac{s^2}{m_0}}, & \text{if } \log p \le m_0 \le s^2; \\ \sqrt{\dfrac{\log p}{N}}, & \text{if } s^2 \le m_0. \end{cases}$$

As m_0 increases, the error decreases at the order of $\sqrt{m_0}$, up to certain point. When m is large enough so that $m_0 \ge \log p$, then there is an additional gain at the order of $\sqrt{\log p}$. However, when m_0 is too large that it exceeds s^2, then the last term $\sqrt{(\log p)/N}$, resulting essentially from estimating prior class probabilities π_1 and π_2, dominates. This may be improved with additional assumptions on the parameters. A similar phase transition phenomenon is also observed for the case $s^2/\log p = O(1)$.

To further emphasize the advantage of having sets of observations, we compare a general case $m_0 = m^*$ where $\log p \le m^* \le s^2$ with the special case that $m_0 = 1$, i.e., the regular QDA situation. Then the quantity ξ_N with m^* has a faster decay rate with a factor of order $\sqrt{m^* \log p}$ compared to the $m_0 = 1$ case, thanks to the extra observations within each set. It is worthwhile to point out that even in the special QDA situation where $m_0 = 1$, we find that our rate of convergence $\sqrt{(\log p)^2/N}(s_q + \sqrt{s_l})$ in Theorem 13 is still faster than the one $(\log p)^{3/2}/N^{1/2}(s_q + s_l)$ derived in the oracle inequality of Fan et al. (2015) under similar assumptions.

5. Numerical Studies

In this section we compare various versions of covariance-engaged set classifiers with other set classifiers adapted from traditional methods. In addition to the CLIPS classifier, we use the diagonalized and enriched versions of $\hat{\Sigma}_k$ respectively (labeled as Plugin(d) and Plugin(e)) introduced at the end of Section 3.1, and plug them in the Bayes rule (2), as done in (6). For comparisons, we also supply the estimated β_0, β and ∇ from the CLIPS procedure to a QDA classifier which is applied to all the observations in a testing set, followed by a majority voting scheme (labeled as QDA-MV). Lastly, we calculate the sample mean and variance of each variable in an observation set to form a new feature vector as done in Miedema et al. (2012); then support vector machine (SVM; Cortes and Vapnik, 1995) and distance weighted discrimination (DWD; Marron et al., 2007) are applied to the features to make predictions (labeled as SVM and DWD respectively). We use R library `clime` to calculate the CLIME estimates, R library `e1071` to calculate the SVM classifier, and R library `sdwd` (Wang and Zou, 2016) to calculate the DWD classifier.

5.1 Simulations

Three scenarios are considered for simulations. In each scenario, we consider a binary setting with $N = 7$ sets in a class, and $M = 10$ observations from normal distribution in each set.

Scenario 1 We set the precision matrix for Class 1 to be $\Sigma_1^{-1} = (1+\sqrt{p})I_p$. For Class 2, we set $\Sigma_2^{-1} = \Sigma_1^{-1} + \tilde{\nabla}$, where $\tilde{\nabla}$ is a $p \times p$ symmetric matrix with 10 elements randomly selected from the upper-triangular part whose values are ζ and other elements being zeros. For the mean vectors, we set $\mu_1 = \Sigma_1(u, u, 0, \ldots, 0)^T$ and $\mu_2 = (0, \ldots, 0)^T$. Note that this makes the true value of $\beta = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2 = (u, u, 0, \ldots, 0)^T$, that is, only the first two covariates have linear impacts on the discriminant function if $u \neq 0$. In this scenario, the true difference in the precision matrices has some sparse and large non-zero entries, whose magnitude is controlled by ζ. Note that while the diagonals of the precision matrices are the same, the diagonals of the covariance matrices are different between the two classes.

Scenario 2 We set the covariance matrices for both classes to be the identity matrix, except that for Class 1 the leading 5 by 5 submatrix of Σ_1 has its off-diagonal elements set to ρ. The rest of the setting is the same as in Scenario 1. In this scenario, both the difference in the covariance and the difference in the precision matrix are confined in the leading 5 by 5 submatrix, so that the majority of matrix entries are the same between the two classes. The level of difference is controlled by ρ: when $\rho = 0$, the two classes have the same covariance matrix.

Scenario 3 We set the precision matrix Σ_1 for Class 1 to be a Toeplitz matrix whose first row is $(1 - \rho^2)^{-1}(\rho^0, \rho^1, \rho^2, \ldots, \rho^{p-1})$. The covariance for Class 2, Σ_2, is a diagonal matrix with the same diagonals as those of Σ_1. It can be shown that the precision matrix for Class 1 is a band matrix with degree 1, that is, a matrix whose nonzero entries are confined to the main diagonal and one more diagonal on both sides. Since the precision matrix for Class 2 is a diagonal matrix, the difference between the

precision matrix has up to $p+2(p-1)$ nonzero entries. The magnitude of the difference is controlled by the parameter ρ. The rest of the setting is the same as in Scenario 1.

We consider different comparisons where we vary the magnitude of the difference in the precision matrices (ζ or ρ), the magnitude of the difference in mean vectors (u), or the dimensionality (p), when the other parameters are fixed.

Comparison 1 (varying ζ or ρ) We vary ζ or ρ but fix $p = 100$ and $u = 0$, which means that the mean vectors have no discriminant power since the true value of β is a zero vector. It shows the performance with different potentials in the covariance structure.

Comparison 2 (varying u) We vary u while fixing $p = 100$ and $\zeta = 0.55$ in Scenario 1 or $\rho = 0.5$ and 0.3 in Scenarios 2 and 3. This case illustrates the potentials of the mean difference when there is some useful discriminative power in the covariance matrices.

Comparison 3 (varying p) We let $p = 80, 100, 120, 140, 160$ while fixing ζ or ρ in the same way as in Comparison 2 and fixing $u = 0.05$, 0.025 and 0.025 in Scenarios 1, 2 and 3 respectively.

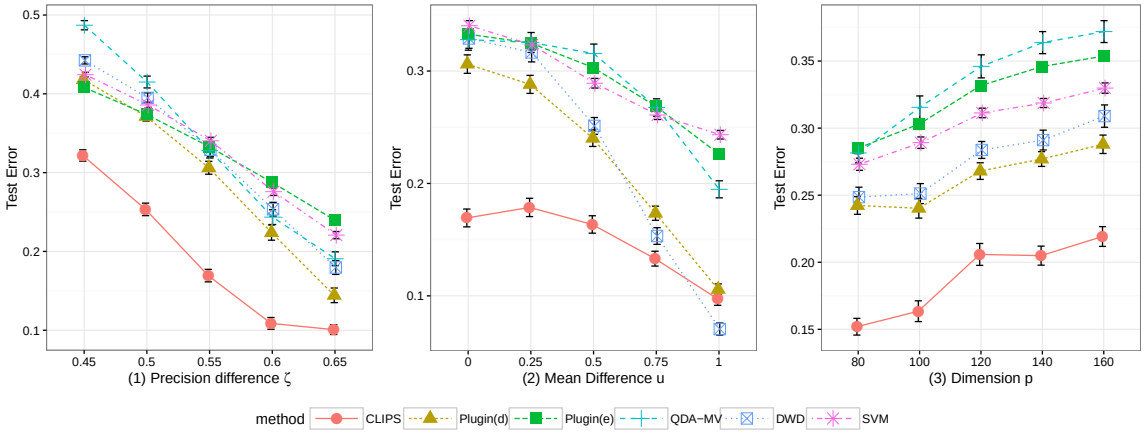

Figure 2: Set classification for Scenario 1. The three panels are corresponding to varying ζ, varying u and varying p respectively. The CLIPS classifier performs very well when the effect of covariance dominates that of the mean difference.

Figure 2 shows the performance for Scenario 1. In the left panel, as ζ increases, the difference between the true precision matrices increases. The proposed CLIPS classifier performs the best among all methods under consideration. It may be surprising that the Plugin(d) method, which does not consider the off-diagonal elements in the sample covariance, can work reasonably well in this setting where the major mode of variation is in the off-diagonal of the precision matrices. However, since large values in the off-diagonal of the precision matrix can lead to large values of some diagonal entries of the covariance matrix, the good performance of Plugin(d) has some partial justification.

In the middle panel of Figure 2, the mean difference starts to increase. While every method more or less gets some improvement, the DWD method has gained the most (it is

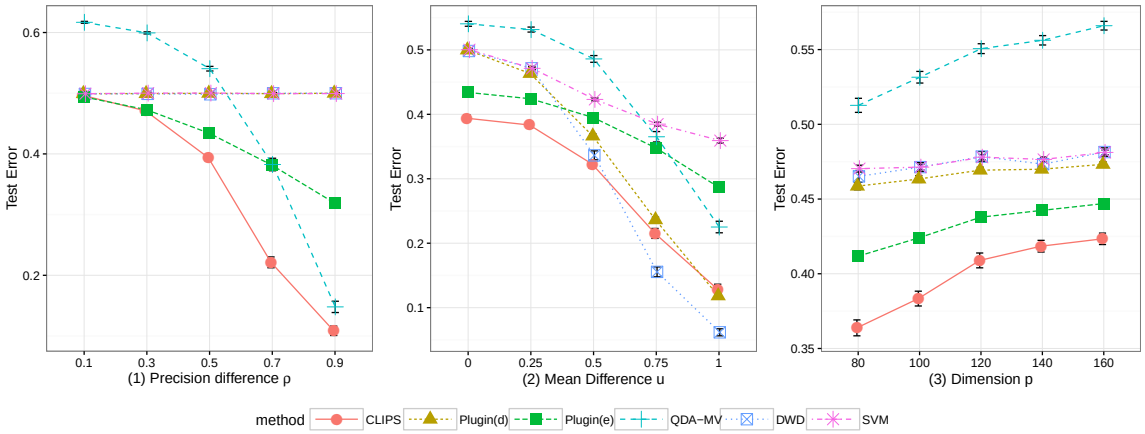

Figure 3: Set classification for Scenario 2. The three panels are corresponding to varying ρ, varying u and varying p respectively. The classifiers that do not engage covariance perform poorly when there is no mean difference signal.

even the best performing classifier when the mean difference u is as large as 1.) This may be due to the fact that the mean difference on which DWD relies, instead of the difference in the precision matrix, is sufficiently large to secure a good performance in separating sets between two classes.

Figure 3 shows the results for Scenario 2. In contrast to Scenario 1, there is no difference in the diagonals of the covariances between the two classes (the precision matrices are still different). When there is no mean difference (see the left panel), it is clear that DWD, SVM and the Plugin(d) method fail for obvious reasons (note that the Plugin(d) method does not read the off-diagonal of the sample covariances and hence both classes have the same

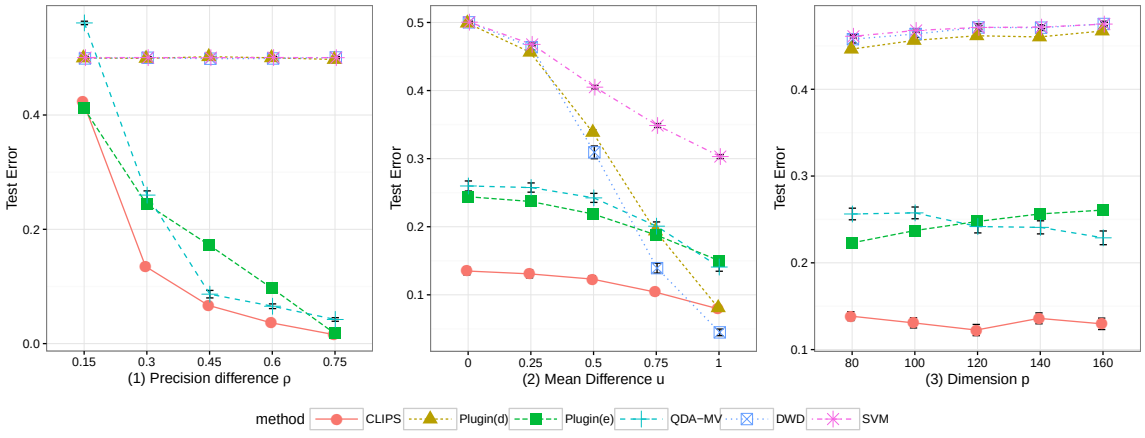

Figure 4: Set classification for Scenario 3. The three panels are corresponding to varying ρ, varying u and varying p respectively. As in Scenario 2, the classifiers that do not engage covariance perform poorly when there is no mean difference signal.

precision matrices from its viewpoint.) As a matter of fact, all these methods perform as badly as random-guess. The CLIPS classifier always performs the best in this scenario in the left panel. Similar to the case in Scenario 1, as the mean difference increases (see the middle panel), the DWD method starts to get some improvement.

The results for Scenario 3 (Figure 4) are similar to Scenario 2, except that, this time the advantage of two covariance-engaged set classification methods, CLIPS and Plugin(e), seems to be more obvious when the mean difference is 0 (see left panel). Moreover, the QDA-MV method also enjoys some good performance, although not as good as the CLIPS classifier.

In all three scenarios, it seems that the test classification error is linearly increasing in the dimension p, except for Scenario 3 in which the signal level depends on p too (greater dimensions lead to greater signals.)

5.2 Data Example

One of the common procedures used to diagnose hepatoblastoma (a rare malignant liver cancer) is biopsy. A sample tissue of a tumor is removed and examined under a microscope. A tissue sample contains a number of nuclei, a subset of which is then processed to obtain segmented images of nuclei. The data we analyzed contain 5 sets of nuclei from normal liver tissues and 5 sets of nuclei from cancerous tissues. Each set contains 50 images. The data set is publicly available (http://www.andrew.cmu.edu/user/gustavor/software.html) and was introduced in Wang et al. (2011, 2010).

We tested the performance of the proposed method on the liver cell nuclei image data set. First, the dimension was reduced from 36,864 to 30 using principal component analysis. Then, among the 50 images of each set, 16 images are retained as training set, 16 are tuning set and another 16 are test set. In other words, for each of the training, tuning, and testing data sets, there are 10 sets of images, five from each class, with 16 images in each set.

Table 1 summarizes the comparison between the methods under consideration. All three covariance-engaged set classifiers (CLIPS, Plugin(d) and Plugin(e)), along with the QDA-MV method, perform better than methods which do not take the covariance matrices much into account, such as DWD and SVM (note that they do look into the diagonal of the covariance matrix.)

To get some insights to the reason that covariance-engaged set classifiers work and traditional methods fail, we visualize the data set in Figure 5. Subfigure (1) shows the scatter plot of the first two principal components of all the elementary observations (ignoring

Method	number of misclassified sets	standard error
CLIPS	0.01/10	0.0104
Plugin(d)	0.74/10	0.0450
Plugin(e)	0.97/10	0.0178
QDA-MV	0.08/10	0.0284
DWD	3.24/10	0.1164
SVM	3.13/10	0.1130

Table 1: Classification performance for the liver cell nucleus image data.

the set memberships) in the data sets, in which different colors (blue versus violet) depict the two different classes. Observations in the same set are shown in the same symbol. The first strong impression is that there is no mean difference between the two classes on the observation level. In contrast, it seems that it is the second moment such as the variance that distinguishes the two classes.

One may argue that DWD and SVM should theoretically work here because they work on the augmented space where the mean and variance of each variable are calculated for each observation set, leading to a $2p$-dimensional feature vector for each set. However, Subfigures (2)–(4) invalidate this argument. We plot the augmented training data in the space formed by the first two principal components (Subfigure (2)). The augmented test data are shown in the same space in Subfigure (3) with a zoomed-in version in Subfigure (4). Note that the scales for Subfigures (2) and (3) are the same. These figures show that

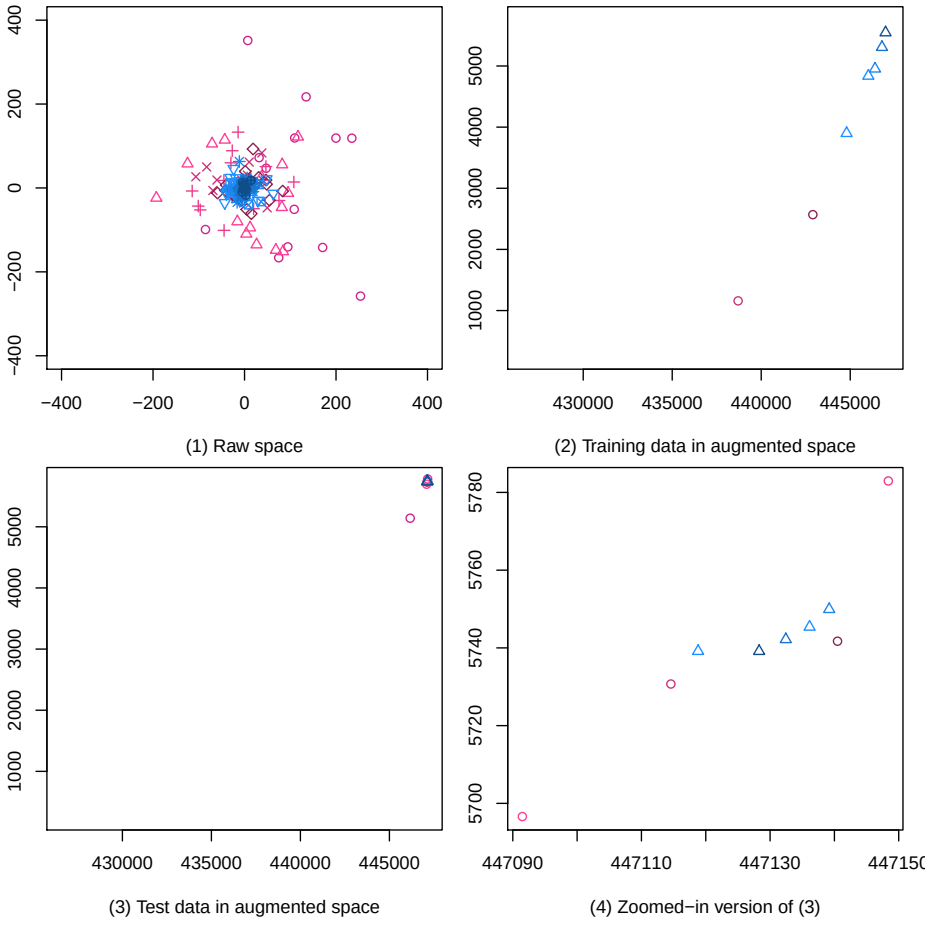

Figure 5: PCA scatter plots for the liver cell nucleus image data. Both classes are shown in different colors. (1): the elementary observations in the raw space; different sets are shown in different symbols. (2) and (3): the augmented space seen by the DWD and SVM methods. (4) is a zoomed-in version of (3). It is shown that traditional multivariate methods have a fundamental difficulty for this data set.

there are more than just the marginal mean and variance that are useful here, and our covariance-engaged set classification methods have used the information in the right way.

Acknowledgments

Jung's research is partially supported by a *National Science Foundation* grant (DMS-1307178) and Qiao's research is partially supported by a collaboration grant from *Simons Foundation* (award number 246649).

Appendix 1: Proofs of Main Results

Proof of Theorem 2

Proof We only prove that $R_{B1} \to 0$ and the proof of $R_{B2} \to 0$ is similar. In addition note that

$$\begin{aligned} R_{Bk} &= \text{pr}(\phi_B(\mathcal{X}^\dagger) \neq k \mid \mathcal{Y}^\dagger = k) \\ &= \sum_{m=c_m m_0}^{C_m m_0} \text{pr}(\phi_B(\mathcal{X}^\dagger) \neq k \mid \mathcal{Y}^\dagger = k, M^\dagger = m) \cdot p_M(m) \\ &:= \sum_{m=c_m m_0}^{C_m m_0} R_{Bk,m} \cdot p_M(m), \end{aligned}$$

where the last equality is due to independence of $\mathcal{Y}^\dagger$ and $M^\dagger$, and Condition 2. Hence it is sufficient for us to focus on any fixed $m \in [c_m m_0, C_m m_0]$.

Given that the set is from Class 1, we have $X_i^\dagger \sim N(\mu_1, \Sigma_1), i = 1, \ldots, m$. The Bayes decision rule classifies the set to Class 2, i.e., $\phi_B(\mathcal{X}^\dagger) = 2$ in (2) if $g(X_1^\dagger, \ldots, X_m^\dagger) < 0$, which is equivalent to

$$\sum_{i=1}^{m} \left(X_i^\dagger - \mu_1\right)^T \nabla \left(X_i^\dagger - \mu_1\right) - 2m\delta^T \Sigma_2^{-1}(\bar{X} - \mu_1) + m\delta^T \Sigma_2^{-1}\delta - m \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 2\log\left(\frac{\pi_1}{\pi_2}\right) < 0, \tag{16}$$

where $\bar{X} = \sum_{i=1}^m X_i^\dagger / m$ is the sample mean.

Define $V = \Sigma_1^{1/2}\Sigma_2^{-1}\Sigma_1^{1/2} - I$ where I is the identity matrix. We set $Z_i = \Sigma_1^{-1/2}(X_i^\dagger - \mu_1) \sim N(0, I)$, $A_{m,p} = \sum_{i=1}^m Z_i^T V Z_i - 2m\delta^T \Sigma_2^{-1}\Sigma_1^{1/2}\bar{Z}$ with $\bar{Z} = \sum_{i=1}^m Z_i/m$. Then the Bayes risk $R_{B1,m}$ can be written as, following from (16),

$$R_{B1,m} = \text{pr}\left(A_{m,p} - \mathbb{E}A_{m,p} < -\alpha\right),$$

where $\alpha = m\text{tr}(V) + m\delta^T\Sigma_2^{-1}\delta - m\log\{|\Sigma_1|/|\Sigma_2|\} + 2\log(\pi_1/\pi_2)$ since $\mathbb{E}A_{m,p} = m\text{tr}(V)$. The strategy to bound $R_{B1,m}$ is to show that $|A_{m,p} - \mathbb{E}A_{m,p}|$ concentrates on $\sqrt{m}D_p$ but $\alpha > 0$ diverges at a faster rate of mD_p^2.

We first give an upper bound of the magnitude of $A_{m,p} - \mathbb{E}A_{m,p}$. Write the eigen-decomposition of V as $U\Lambda U^T$ and the diagonal matrix $\Lambda = \text{diag}(\lambda_j)$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Moreover, set $\tilde{Z}_i = U^T Z_i \sim N(0, I)$ with $\tilde{Z}_{i,j}$ its jth entry. Note that

$$A_{m,p} - \mathbb{E}A_{m,p} = \sum_{i=1}^{m}\sum_{j=1}^{p} \lambda_j(\tilde{Z}_{i,j}^2 - 1) - 2m\delta^T\Sigma_2^{-1}\Sigma_1^{1/2}\bar{Z}.$$

The tail probability of normal distribution implies

$$\text{pr}(|2m\delta^T\Sigma_2^{-1}\Sigma_1^{1/2}\bar{Z}| > t) \leq 2\exp\left\{-\frac{1}{2}\left(\frac{t}{2\sqrt{m}\|\delta^T\Sigma_2^{-1}\Sigma_1^{1/2}\|}\right)^2\right\} \leq 2\exp\left(-\frac{C_e^{-3}t^2}{8m\|\delta\|^2}\right), \tag{17}$$

where the last inequality is due to Condition 1. Since $\tilde{Z}_{i,j}^2 - 1$ is sub-exponential, Bernstein's inequality (e.g. Vershynin, 2012, Proposition 5.16) implies that there exists some universal constant $c_1 > 0$ such that

$$\mathrm{pr}(|\sum_{i=1}^{m}\sum_{j=1}^{p}\lambda_j(\tilde{Z}_{i,j}^2 - 1)| > t) \leq 2\exp\left(-c_1\min(\frac{t^2}{m\|\Lambda\|_F^2}, \frac{t}{\max\{|\lambda_1|, |\lambda_p|\}})\right). \tag{18}$$

Now we focus on the lower bound of α. First of all, notice that $m\delta^T\Sigma_2^{-1}\delta \geq mC_e^{-1}\|\delta\|^2$ by Condition 1. Moreover, there exists some constant $c_2 > 0$ depending on C_e only such that

$$\begin{aligned} m\mathrm{tr}(V) - m\log\{|\Sigma_1|/|\Sigma_2|\} &= m\left(\mathrm{tr}(V) - \log|I+V|\right) \\ &= m\sum_{j=1}^{p}(\lambda_j - \log(1+\lambda_j)) \geq c_2 m\|\Lambda\|_F^2 \end{aligned} \tag{19}$$

where the last inequality follows from that $\lambda_j + 1 \in [C_e^{-2}, C_e^2]$ according to Condition 1. Note that $\|\Lambda\|_F = \|V\|_F = \|\Sigma_1^{1/2}\nabla\Sigma_1^{1/2}\|_F$ and $C_e^{-1} \leq \|V\|_F/\|\nabla\|_F \leq C_e$ according to Condition 1. Therefore by combining the above two results we conclude $\alpha \geq c_3 m D_p^2 + 2\log(\pi_1/\pi_2)$ with $c_3 = \min(c_2 C_e^{-2}, C_e^{-1}) > 0$.

Note that by Conditions 1 and 3, λ_1 in equation (18) and $2\log(\pi_1/\pi_2)$ in the expression of α are bounded. When mD_p^2 is large enough, we can pick $t = cmD_p^2$ for small enough $c > 0$ in equations (17) and (18) such that $A_{m,p} - \mathbb{E}A_{m,p} > -\alpha$ with probability at least $1 - 4\exp\left(-c'mD_p^2\right)$. Therefore we complete our proof by seeing that for each fixed m, $R_{B1,m} \leq 4\exp\left(-c'mD_p^2\right)$ for some small constant $c' > 0$, together with the fact $m \in [c_m m_0, C_m m_0]$ from Condition 2. ■

Proof of Theorem 4

Proof We only prove that $\hat{R}_1 \to 0$ with high probability and $\hat{R}_2 \to 0$ can be shown by symmetry. The strategy of the proof is similar to that for Theorem 2. We further focus on each fixed $m \in [c_m m_0, C_m m_0]$ since

$$\begin{aligned} \hat{R}_k &= \sum_{m=c_m m_0}^{C_m m_0} \mathrm{pr}(\hat{\phi}(\mathcal{X}^\dagger) \neq k \mid \mathcal{Y}^\dagger = k, M^\dagger = m) \cdot p_M(m) \\ &:= \sum_{m=c_m m_0}^{C_m m_0} \hat{R}_{k,m} \cdot p_M(m). \end{aligned} \tag{20}$$

The quadratic set classifier classifies the set to 2, that is, $\hat{\phi}(\mathcal{X}^\dagger) = 2$ in (6) if

$$\sum_{i=1}^{m}\left(X_i^\dagger - \hat{\mu}_1\right)^T\hat{\nabla}\left(X_i^\dagger - \hat{\mu}_1\right) - 2m\hat{\delta}^T\hat{\Sigma}_2^{-1}(\bar{X} - \hat{\mu}_1) + m\hat{\delta}^T\hat{\Sigma}_2^{-1}\hat{\delta} - m\log\left(\frac{\left|\hat{\Sigma}_1\right|}{\left|\hat{\Sigma}_2\right|}\right) + 2\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) < 0,$$

where $\hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1$ and $\bar{X} = \sum_{i=1}^m X_i^\dagger/m$. Define

$$\hat{A}_{m,p} = \sum_{i=1}^m \left(X_i^\dagger - \hat{\mu}_1\right)^T \hat{\nabla} \left(X_i^\dagger - \hat{\mu}_1\right) - 2m\hat{\delta}^T\hat{\Sigma}_2^{-1}(\bar{X} - \hat{\mu}_1) := \hat{A}_{1,m,p} + \hat{A}_{2,m,p}.$$

Then the generalization error $\hat{R}_{1,m}$, which is a random variable as a function of $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N$, can be written as

$$\hat{R}_1 = \hat{R}_1((\mathcal{X}, \mathcal{Y})) = \mathrm{pr}\left(\hat{A}_{m,p} - \mathbb{E}\hat{A}_{m,p} < -\hat{\alpha}\right), \tag{21}$$

where pr and $\mathbb{E}$ are understood as the conditional expectation given $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N$ and

$$\hat{\alpha} = \mathbb{E}(\hat{A}_{1,m,p} + \hat{A}_{2,m,p}) + m\hat{\delta}^T\hat{\Sigma}_2^{-1}\hat{\delta} - m\log\left(\left|\hat{\Sigma}_1\right| / \left|\hat{\Sigma}_2\right|\right) + 2\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right).$$

The following lemma facilitates our analysis.

Lemma 14 *For any fixed $L > 0$, under the assumptions $p \le c_0 N m_0$ and $\log p \le c_0 N$ with sufficiently small $c_0 > 0$, we have that (i) $C'^{-1} \le \lambda_{\min}(\hat{\Sigma}_k) \le \lambda_{\max}(\hat{\Sigma}_k) \le C'$; (ii) $\|\mu_k - \hat{\mu}_k\| \le C'\sqrt{\frac{p}{Nm_0}}$; (iii) $\|\Sigma_k - \hat{\Sigma}_k\|_F \le C'\sqrt{\frac{p^2}{Nm_0}}$ and (iv) $|\pi_k - \hat{\pi}_k| \le C'\sqrt{\frac{\log p}{N}}$, $k = 1, 2$ with probability at least $1 - O(p^{-L})$, where positive constant C' depend on C_e, c_m, L and C_π only.*

From now on, we condition on the event $\mathcal{E}$ in which results (i)-(iv) of Lemma 14 hold for training data $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N$. All positive constants used hereafter only depend on C_e and c_0. Clearly, since $p^2/(Nm_0D_p^2)$ is sufficiently small, Lemma 14 (ii) and (iii) imply that

$$\hat{D}_p = \left(\|\hat{\nabla}\|_F^2 + \|\hat{\delta}\|^2\right)^{1/2} \asymp D_p. \tag{22}$$

We show the concentration radius of $\hat{A}_{m,p} - \mathbb{E}\hat{A}_{m,p}$ is much smaller than $\hat{\alpha}$ under our assumptions.

First of all, we analyze the left side $\hat{A}_{m,p} - \mathbb{E}\hat{A}_{m,p} = \Sigma_{k=1}^2(\hat{A}_{k,m,p} - \mathbb{E}\hat{A}_{k,m,p})$. Note that $\hat{A}_{2,m,p} - \mathbb{E}\hat{A}_{2,m,p} = -2\sum_{i=1}^m \hat{\delta}^T\hat{\Sigma}_2^{-1}\Sigma_1^{1/2}Z_i$, where $Z_i = \Sigma_1^{-1/2}(X_i^\dagger - \mu_1) \overset{\text{i.i.d}}{\sim} N(0, I)$. Note Lemma 14 implies the spectral norm $\left\|\hat{\Sigma}_2^{-1}\Sigma_1^{1/2}\right\|_{\ell_2} \le C'C_e^{1/2}$. The tail probability of normal distribution implies (similarly as in equation (17)) there exists some constant $C_1 > 0$ such that,

$$\mathrm{pr}(|\hat{A}_{2,m,p} - \mathbb{E}\hat{A}_{2,m,p}| > t) \le 2\exp\left(-\frac{C_1 t^2}{m\|\hat{\delta}\|^2}\right). \tag{23}$$

Besides, $\hat{A}_{1,m,p} - \mathbb{E}\hat{A}_{1,m,p} = W_1 + W_2$, where

$$W_1 \;:\; = \mathrm{tr}[\hat{\nabla}(\sum_{i=1}^m \left(X_i^\dagger - \mu_1\right)\left(X_i^\dagger - \mu_1\right)^T)] - \mathrm{tr}[\hat{\nabla}m\Sigma_1],$$

$$W_2 := 2\left(\mu_1 - \hat{\mu}_1\right)^T\hat{\nabla}\Sigma_1^{1/2}\sum_{i=1}^m Z_i.$$

Set $\hat{V} = \Sigma_1^{1/2}\hat{\nabla}\Sigma_1^{1/2}$ and its eigen-values $\{\hat{\lambda}_j\}_{j=1}^p$. By a similar argument using Bernstein's inequality like (18), we have that there exists some constant $c_1 > 0$ such that

$$\mathrm{pr}(|W_1| > t) \leq 2\exp\left(-c_1\min(\frac{t^2}{m\|\hat{V}\|_F^2}, \frac{t}{\max\{\left|\hat{\lambda}_1\right|, \left|\hat{\lambda}_p\right|\}})\right). \tag{24}$$

To control W_2, we apply again the tail probability of normal distribution to obtain that for some constants C_2, $C_3 > 0$,

$$\mathrm{pr}(|W_2| > t) \leq 2\exp\left(-\frac{C_2t^2}{m\|\hat{\nabla}\|_{\ell_2}^2 \cdot \|\mu_1 - \hat{\mu}_1\|^2}\right) \leq 2\exp\left(-\frac{C_3t^2}{m\|\hat{\nabla}\|_F^2}\right), \tag{25}$$

since $\|\mu_1 - \hat{\mu}_1\| \leq C'\sqrt{\frac{p}{Nm_0}} \leq C'c_0^{1/2}$ by Lemma 14. Therefore equations (23)-(25), together with (22), imply that for some $C_4 > 0$,

$$\mathrm{pr}(|\hat{A}_{m,p} - \mathbb{E}\hat{A}_{m,p}| > t) \leq 6\exp\left(-\frac{C_4t^2}{mD_p^2}\right). \tag{26}$$

Now we lower bound the right side $\hat{\alpha}$. This term can be decomposed into six terms.

$$\begin{aligned}\hat{\alpha} \;=\; & m\hat{\delta}^T\hat{\Sigma}_2^{-1}\hat{\delta} + \left[m\mathrm{tr}(\hat{\nabla}\hat{\Sigma}_1) - m\log\left(\left|\hat{\Sigma}_1\right| / \left|\hat{\Sigma}_2\right|\right)\right] + 2\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) + \\ & m\mathrm{tr}(\hat{\nabla}(\Sigma_1 - \hat{\Sigma}_1)) - 2m\hat{\delta}^T\hat{\Sigma}_2^{-1}(\mu_1 - \hat{\mu}_1) + m(\mu_1 - \hat{\mu}_1)\hat{\nabla}(\mu_1 - \hat{\mu}_1)^T.\end{aligned}$$

These terms have the following bounds respectively with some constant $C_5, C_6, C_7, C_8, C_9 > 0$,

$$\begin{aligned}m\hat{\delta}^T\hat{\Sigma}_2^{-1}\hat{\delta} &\geq C_5m\|\hat{\delta}\|^2, &(27)\\ m\mathrm{tr}(\hat{\nabla}\hat{\Sigma}_1) - m\log\left(\left|\hat{\Sigma}_1\right| / \left|\hat{\Sigma}_2\right|\right) &\geq C_5m\|\hat{\nabla}\|_F^2, &(28)\\ \left|m\mathrm{tr}(\hat{\nabla}(\Sigma_1 - \hat{\Sigma}_1))\right| &\leq C_6m\|\hat{\nabla}\|_F\|\Sigma_1 - \hat{\Sigma}_1\|_F \leq C_7m\|\hat{\nabla}\|_F(p^2/Nm_0)^{1/2}, &(29)\\ \left|2m\hat{\delta}^T\hat{\Sigma}_2^{-1}(\mu_1 - \hat{\mu}_1)\right| &\leq C_6m\|\hat{\delta}\|\|\mu_1 - \hat{\mu}_1\| \leq C_7m\|\hat{\delta}\|(p/Nm_0)^{1/2}, &(30)\\ |2\log(\hat{\pi}_1/\hat{\pi}_2)| &\leq C_6, &(31)\\ \left|m(\mu_1 - \hat{\mu}_1)\hat{\nabla}(\mu_1 - \hat{\mu}_1)^T\right| &\leq C_8m\|\hat{\nabla}\|_{\ell_2}\|\mu_1 - \hat{\mu}_1\|^2 \leq C_9m(p/Nm_0). &(32)\end{aligned}$$

Equations (27) and (28) are due to (i) of Lemma 14. In particular, (28) follows from a similar argument as (19). Equations (29) and (30) follow from (iii) and (ii) of Lemma 14 respectively while equation (31) is due to (iv) of Lemma 14 and Condition 3. Equation (32) follows from (i) and (ii) of Lemma 14. Furthermore, notice that $p^2/(Nm_0D_p^2)$ is sufficiently small and $m_0D_p^2$ is sufficiently large, equations (27)-(32) as well as (22) yield that $\hat{\alpha} \geq C_{10}mD_p^2$ for some small constant $C_{10} > 0$.

Finally, the lower bound of $\hat{\alpha}$ and concentration of $\hat{A}_{m,p} - \mathbb{E}\hat{A}_{m,p}$ in (26) with $t = c''mD_p^2$ for small enough $c'' > 0$, together with the assumption D_p^2m is sufficiently large, imply that

the generalization error of the quadratic set classification rule $\hat{R}_{1,m} \leq 2\exp\left(-c'mD_p^2\right)$ for each $m \in [c_m m_0, C_m m_0]$ on the event $\mathcal{E}$. Hence we complete our proof by applying Lemma 14 and equation (20), that is, $\hat{R} \leq 4\exp\left(-c'm_0 D_p^2\right)$ with probability at least $1 - O(p^{-L})$. ■

Proof of Theorem 6

Proof First we show that Σ_k^{-1} is feasible for the optimization problem (7), that is $\|\hat{\Sigma}_k\Sigma_k^{-1} - I\|_\infty < \lambda_{1,N}$. It suffices to show that $\|\hat{\Sigma}_k - \Sigma_k\|_\infty < C_{\ell 1}^{-1}\lambda_{1,N}$ because $\|\hat{\Sigma}_k\Sigma_k^{-1} - I\|_\infty \leq \|\hat{\Sigma}_k - \Sigma_k\|_\infty \|\Sigma_k^{-1}\|_{\ell_1} \leq \|\hat{\Sigma}_k - \Sigma_k\|_\infty C_{\ell 1}$. The following lemma establishes this result, given our choice of $\lambda_{1,N}$.

Lemma 15 *Under the assumption $\log p \leq c_0 N$ with some sufficiently small $c_0 > 0$, we have that (i) $\|\hat{\mu}_k - \mu_k\|_\infty \leq C\sqrt{(\log p)/(Nm_0)}$ and (ii) $\|\hat{\Sigma}_k - \Sigma_k\|_\infty \leq C\sqrt{(\log p)/(Nm_0)}$, $k = 1, 2$ with probability at least $1 - O(p^{-L})$, where positive constant C depends on C_e, c_m, C_π and L only.*

From now on, we condition on the event in which both results in Lemma 15 hold. We next control the supnorm bound $\left\|\Sigma_k^{-1} - \tilde{\Omega}_k\right\|_\infty$. Since both Σ_k^{-1} and $\tilde{\Omega}_k$ are feasible for (7), we have $\|\hat{\Sigma}_k(\Sigma_k^{-1} - \tilde{\Omega}_k)\|_\infty = \|\hat{\Sigma}_k\Sigma_k^{-1} - I - (\hat{\Sigma}_k\tilde{\Omega}_k - I)\|_\infty \leq 2\lambda_{1,N}$. Moreover,

$$\begin{aligned}
\|\Sigma_k(\Sigma_k^{-1} - \tilde{\Omega}_k)\|_\infty &\leq \|(\hat{\Sigma}_k - \Sigma_k)(\Sigma_k^{-1} - \tilde{\Omega}_k)\|_\infty + \|\hat{\Sigma}_k(\Sigma_k^{-1} - \tilde{\Omega}_k)\|_\infty \\
&\leq \|\Sigma_k^{-1} - \tilde{\Omega}_k\|_{\ell_1}\|\hat{\Sigma}_k - \Sigma_k\|_\infty + 2\lambda_{1,N} \\
&\leq \left(\left\|\Sigma_k^{-1}\right\|_{\ell_1} + \|\tilde{\Omega}_k\|_{\ell_1}\right) C_{\ell 1}^{-1}\lambda_{1,N} + 2\lambda_{1,N} \\
&\leq 2C_{\ell 1}C_{\ell 1}^{-1}\lambda_{1,N} + 2\lambda_{1,N} = 4\lambda_{1,N},
\end{aligned}$$

where we have used the fact $\tilde{\Omega}_k$ is the solution of CLIME which implies for each $j = 1, \ldots, p$, $\|(\tilde{\Omega}_k)_j\|_1 \leq \left\|(\Sigma_k^{-1})_j\right\|_1$ and hence $\|\tilde{\Omega}_k\|_{\ell_1} \leq \left\|\Sigma_k^{-1}\right\|_{\ell_1}$, where $(\tilde{\Omega}_k)_j$ and $(\Sigma_k^{-1})_j$ denote the jth column of $\tilde{\Omega}_k$ and Σ_k^{-1} respectively. We conclude with $\|\Sigma_k^{-1} - \tilde{\Omega}_k\|_\infty \leq \|\Sigma_k^{-1}\|_{\ell_1}\|\Sigma_k(\Sigma_k^{-1} - \tilde{\Omega}_k)\|_\infty \leq 4M_0\lambda_{1,N}$.

Based on the supnorm bound obtained above, we have

$$\|(\tilde{\Omega}_2 - \tilde{\Omega}_1) - \nabla\|_\infty \leq \|\Sigma_1^{-1} - \tilde{\Omega}_1\|_\infty + \|\Sigma_2^{-1} - \tilde{\Omega}_2\|_\infty \leq 8C_{\ell 1}\lambda_{1,N}. \tag{33}$$

Recall that $\mathrm{supp}(\nabla)$ is the support of the matrix ∇. The thresholding step (8), together with (33), guarantees that $\tilde{\nabla}_{ij} = 0$ for any $(i, j) \notin \mathrm{supp}(\nabla)$, noting that $\lambda'_{1,N} \geq 8C_{\ell 1}\lambda_{1,N}$. Therefore we have shown the subset selection result, that is, $\mathrm{pr}(\mathrm{supp}(\tilde{\nabla}) \subset \mathrm{supp}(\nabla)) = 1 - O(p^{-L})$. Moreover, we have that $\|\tilde{\nabla} - \nabla\|_\infty \leq 8C_{\ell 1}\lambda_{1,N} + \lambda'_{1,N} \leq 2\lambda'_{1,N}$. In the end, we complete the proof by noting that the Frobenius norm bound and vector ℓ_1 norm bound are the consequences of supnorm bound and subset selection result, that is, $\mathrm{pr}(\|\tilde{\nabla} - \nabla\|_F \leq 2\lambda'_{1,N}\sqrt{s_q}) = 1 - O(p^{-L})$ and $\mathrm{pr}(\|\tilde{\nabla} - \nabla\|_1 \leq 2\lambda'_{1,N}s_q) = 1 - O(p^{-L})$. ■

Proof of Theorem 9

Proof We first show that $(\beta_1, \beta_2) = (\Sigma_1^{-1}\mu_1, \Sigma_2^{-1}\mu_2)$ is feasible in (9) with the constant L_1 set as C_β. Note since $\|\beta_k\|_1 \le C_\beta$, The pair (β_1, β_2) satisfies the ℓ_1 norm constraint. This fact, together with the following lemma, implies that (β_1, β_2) is feasible with probability at least $1 - O(p^{-L})$ and hence $\|\hat{\beta}\|_1 \le \|\beta\|_1$.

Lemma 16 *Under the assumption* $\log p \le c_0 N$ *with some sufficiently small constant* $c_0 > 0$*, we have that* $\mathrm{pr}(\|\hat{\Sigma}_k\beta_k - \hat{\mu}_k\|_\infty \ge C\sqrt{\frac{\log p}{Nm_0}}) \le C'p^{-L}$*,* $k = 1, 2$*, where* $C' > 0$ *is some universal constant and constant* $C > 0$ *depends on* $C_e, c_m, C_\pi, C_\beta, C_\mu$ *and* L *only.*

Next we show that $\|\tilde{\beta} - \beta\|_\infty \le 6C_{\ell 1}\lambda_{2,N}$. Notice that for $k = 1, 2$, there exists some constant $C > 0$ such that with probability at least $1 - O(p^{-L})$,

$$\begin{aligned}
\|\Sigma_k\left(\tilde{\beta}_k - \beta_k\right)\|_\infty &\le \|\hat{\Sigma}_k\left(\tilde{\beta}_k - \beta_k\right)\|_\infty + \|\left(\Sigma_k - \hat{\Sigma}_k\right)\left(\tilde{\beta}_k - \beta_k\right)\|_\infty \\
&\le \|\hat{\Sigma}_k\beta_k - \hat{\mu}_k\|_\infty + \|\hat{\Sigma}_k\tilde{\beta}_k - \hat{\mu}_k\|_\infty + \|\Sigma_k - \hat{\Sigma}_k\|_\infty\left(\|\beta_k\|_1 + \|\tilde{\beta}_k\|_1\right) \\
&\le 2\lambda_{2,N} + 2C_\beta C\sqrt{\frac{\log p}{Nm_0}} \le 3\lambda_{2,N},
\end{aligned}$$

where we have used assumption on $\|\beta_k\|_1$, constraints on estimators, the choice of our $\lambda_{2,N}$ and the result (ii) of Lemma 15. Therefore we further have,

$$\|\tilde{\beta} - \beta\|_\infty \le \sum_{k=1}^{2}\|\tilde{\beta}_k - \beta_k\|_\infty \le \sum_{k=1}^{2}\|\Sigma_k^{-1}\|_{\ell_1}\|\Sigma_k\left(\tilde{\beta}_k - \beta_k\right)\|_\infty \le 6C_{\ell 1}\lambda_{2,N}. \tag{34}$$

In the end, we condition on the event in which both (34) and the fact that (β_1, β_2) is feasible hold. The arguments above imply this event holds with probability at least $1 - O(p^{-L})$. We are ready to prove the rates of convergence of $\tilde{\beta}$ under ℓ_1 and ℓ_2 norm losses. Denote the support of β by T. Set $t = 6C_{\ell 1}\lambda_{2,N}$ and the thresholded version of $\tilde{\beta}$ as $\tilde{\beta}^{thr} = (\tilde{\beta}_j^{thr})$, where $\tilde{\beta}_j^{thr} = \tilde{\beta}_j\mathbb{1}\left\{|\tilde{\beta}_j| \ge 2t\right\}$. Since $\beta = \beta_1 - \beta_2$ is feasible, we have that $\|\beta\|_1 \ge \|\tilde{\beta}\|_1 = \|\tilde{\beta}^{thr}\|_1 + \|\tilde{\beta} - \tilde{\beta}^{thr}\|_1 \ge \|\tilde{\beta} - \tilde{\beta}^{thr}\|_1 + \|\beta\|_1 - \|\tilde{\beta}^{thr} - \beta\|_1$. Therefore we obtain that $\|\tilde{\beta} - \tilde{\beta}^{thr}\|_1 \le \|\tilde{\beta}^{thr} - \beta\|_1$, which further implies that $\|\tilde{\beta} - \beta\|_1 \le 2\|\tilde{\beta}^{thr} - \beta\|_1$. To show the bound of $\|\tilde{\beta} - \beta\|_1$, it suffices to bound $\|\tilde{\beta}^{thr} - \beta\|_1$. Indeed, we bound its ℓ_2 norm as an intermediate step,

$$\begin{aligned}
\|\tilde{\beta}^{thr} - \beta\|^2 &= \|\left(\tilde{\beta}^{thr} - \beta\right)_T\|^2 \\
&= \sum_{j\in T}\left(\tilde{\beta}_j^{thr} - \beta_j\right)^2\mathbb{1}\left\{\tilde{\beta}_j^{thr} = 0\right\} + \sum_{j\in T}\left(\tilde{\beta}_j - \beta_j\right)^2\mathbb{1}\left\{\tilde{\beta}_j^{thr} \ne 0\right\} \\
&\le \sum_{j\in T}\beta_j^2\mathbb{1}\{\beta_j \le 3t\} + s_l t^2 \le 10 s_l t^2,
\end{aligned} \tag{35}$$

where we have used supnorm bound (34) in the first and third equations and the fact $|T| \le s_l$ due to $\beta \in \mathcal{F}_0(s_l)$ in the third and fourth equations. Consequently,

$$\|\tilde{\beta}^{thr} - \beta\|_1 = \|\left(\tilde{\beta}^{thr} - \beta\right)_T\|_1 \le \sqrt{s_l}\|\tilde{\beta}^{thr} - \beta\| = \sqrt{10 s_l}t,$$

which completes our first desired result $\|\tilde{\beta}-\beta\|_1 \leq 2\sqrt{10}s_l t = 12\sqrt{10}C_{\ell 1}s_l\lambda_{2,N}$.

To show the bound of $\|\tilde{\beta}-\beta\| \leq \|\tilde{\beta}^{thr}-\beta\| + \|\tilde{\beta}-\tilde{\beta}^{thr}\|$, it suffices to bound $\|\tilde{\beta}-\tilde{\beta}^{thr}\|$ given (35). To this end, we note $\|\beta\|_1 \geq \|\tilde{\beta}\|_1$ implies that $\|\tilde{\beta}_{T^c}\|_1 \leq \|\tilde{\beta}-\beta\|_1 \leq 2\sqrt{10}s_l t$. Moreover,

$$\begin{aligned}
\|\tilde{\beta}-\tilde{\beta}^{thr}\|^2 &= \|\left(\tilde{\beta}^{thr}-\tilde{\beta}\right)_T\|^2 + \|\left(\tilde{\beta}^{thr}-\tilde{\beta}\right)_{T^c}\|^2 \\
&\leq 4t^2 s_l + \sum_{j\in T^c}\tilde{\beta}_j^2 \mathbb{1}\Big\{|\tilde{\beta}_j| < 2t\Big\} \\
&\leq 4t^2 s_l + \|\tilde{\beta}_{T^c}\|_1 \max_{j\in T^c}\{|\tilde{\beta}_j|\mathbb{1}\Big\{|\tilde{\beta}_j| < 2t\Big\}\} \leq (4+4\sqrt{10})t^2 s_l, \qquad (36)
\end{aligned}$$

where the first inequality follows from $|\tilde{\beta}_j^{thr}-\tilde{\beta}_j| < 2t$ and $|T| \leq s_l$, and the second one is due to Hölder's inequality. Therefore combining (35) and (36), we obtained the second desired result $\|\tilde{\beta}-\beta\| \leq \sqrt{s_l}t(\sqrt{10}+(4+4\sqrt{10})^{1/2})$. ■

Proof of Theorem 11

Proof Since we use sample splitting technique, estimators $\tilde{\beta}$ and $\tilde{\nabla}$ are independent with the second batch of the training data used in (10). We assume fixed $\tilde{\beta}$ and $\tilde{\nabla}$, which satisfy our assumptions throughout the analysis. With a slight abuse of notation, we still use N to denote the number of sample sets, although only half of the sample sets are applied to count n_k and $\hat{\pi}_k$, $k=1,2$.

Recall that $\bar{X}_i$ and S_i are the sample mean and variance of the ith set of observations. Define $\tilde{Z}_i = \log(\hat{\pi}_1/\hat{\pi}_2)/M_i + \bar{X}_i^T\tilde{\beta} + \bar{X}_i^T\tilde{\nabla}\bar{X}_i/2 + \mathrm{tr}(\tilde{\nabla}S_i)/2$, which is used to approximate $Z_i = \log(\pi_1/\pi_2)/M_i + \bar{X}_i^T\beta + \bar{X}_i^T\nabla\bar{X}_i/2 + \mathrm{tr}(\nabla S_i)/2$. To facilitate analysis, we denote $\ell(\theta_0|\{(\mathcal{X}_i,\mathcal{Y}_i)\}_{i=1}^N,\tilde{\beta},\tilde{\nabla})$ as $\ell(\theta_0)$ for short. Rewrite our estimator in the following way,

$$\begin{aligned}
\tilde{\beta}_0 &= \operatorname*{argmin}_{\theta_0\in\mathbb{R}}\ell(\theta_0), \text{where} \\
\ell(\theta_0) &= \frac{1}{N}\Sigma_{i=1}^N[\log\Big(1+\exp(M_i(\theta_0+\tilde{Z}_i))\Big) - (2-\mathcal{Y}_i)M_i\Big(\theta_0+\tilde{Z}_i\Big)].
\end{aligned}$$

We start our analysis by conditioning on $\{\mathcal{X}_i\}_{i=1}^N$. Define $\ell_0(\theta_0,\tilde{Z}) = \mathbb{E}(\ell(\theta_0)|\{\mathcal{X}_i\}_{i=1}^N)$ where the expectation is understood as the conditional expectation given $\{\mathcal{X}_i\}_{i=1}^N$. Note that the function $\ell_0(\theta_0,\tilde{Z})$ depends on $\theta_0,\{M_i\}_{i=1}^N$ and $\{\tilde{Z}_i\}_{i=1}^N$ only. Then the difference $\ell(\theta_0)-\ell_0(\theta_0,\tilde{Z}) = \frac{1}{N}\Sigma_{i=1}^N(\mathcal{Y}_i-\mathbb{E}(\mathcal{Y}_i|\mathcal{X}_i))M_i(\theta_0+\tilde{Z}_i) := E_{\theta_0}$. Recall β_0 is the true constant coefficient. Since $\tilde{\beta}_0$ is the minimizer, we have $\ell(\tilde{\beta}_0) \leq \ell(\beta_0)$, i.e.,

$$\begin{aligned}
\ell_0(\tilde{\beta}_0,\tilde{Z}) &\leq \ell_0(\beta_0,\tilde{Z}) + E_{\beta_0} - E_{\tilde{\beta}_0} \\
&\leq \ell_0(\beta_0,\tilde{Z}) + m_0 R_1\left|\tilde{\beta}_0-\beta_0\right|. \qquad (37)
\end{aligned}$$

In the end, we need to bound the term $R_1 = |\frac{1}{Nm_0}\Sigma_{i=1}^N(\mathcal{Y}_i-\mathbb{E}(\mathcal{Y}_i|\mathcal{X}_i))M_i|$. By applying Hoeffding's inequality (e.g. Vershynin, 2012, Proposition 5.10), we obtain $R_1 \leq C_r\sqrt{(\log p)/N}$ with probability at least $1-O(p^{-L})$, where constant C_r depends on L and

C_m only, noting that $M_i \le C_m m_0$ by Condition 2. This probabilistic statement on bounding R_1 is valid conditioning on any realization of $\{\mathcal{X}_i\}_{i=1}^N$ and thus is also valid unconditionally.

Next we apply the Taylor expansion to the function $\ell_0(\theta_0, \tilde{Z})$ to analyze our estimator. Here due to misspecified values $\tilde{Z}_i$, we need a refined version of Taylor expansion (Bach et al., 2010, Proposition 1).

Lemma 17 (Bach et al. (2010)) *Let $g(t) : \mathbb{R} \to \mathbb{R}$ be a convex three times differentiable function such that it satisfies for all $t \in \mathbb{R}$, $|g'''(t)| \le L g''(t)$ for some $L > 0$. Then we have for any t and $v \in \mathbb{R}$,*

$$g(t+v) \ge g(t) + v g'(t) + \frac{g''(t)}{L^2}(e^{-L|v|} + L|v| - 1).$$

It is not hard to see that the third derivative of $\ell_0(\theta_0, \tilde{Z})$ w.r.t. θ_0 is bounded by its second derivative up to a multiplicative factor $\max_i M_i$, i.e.,

$$\max_{\theta_0} \left| \ell_0'''(\theta_0, \tilde{Z}) / \ell_0''(\theta_0, \tilde{Z}) \right| \le \max_i M_i,$$

where hereafter $\ell_0'(\cdot,\cdot)$, $\ell_0''(\cdot,\cdot)$ and $\ell_0'''(\cdot,\cdot)$ are defined as the first, second and third derivative of $\ell_0(\cdot,\cdot)$ w.r.t. the first argument respectively. Applying Lemma 17 to $\ell_0(\theta_0, \tilde{Z})$ at point β_0 and by Condition 2, we obtain that

$$\ell_0(\tilde{\beta}_0, \tilde{Z}) - \ell_0(\beta_0, \tilde{Z}) \ge \ell_0'(\beta_0, \tilde{Z})(\tilde{\beta}_0 - \beta_0) + \frac{\ell_0''(\beta_0, \tilde{Z})}{C_m^2 m_0^2}(e^{-C_m m_0 |\tilde{\beta}_0 - \beta_0|} + C_m m_0 \left|\tilde{\beta}_0 - \beta_0\right| - 1). \tag{38}$$

Note that with misspecified values $\tilde{Z}_i$, in general $\ell_0'(\beta_0, \tilde{Z}) \neq 0$. To finish our proof, we need an upper bound for $\ell_0'(\beta_0, \tilde{Z})$ and a lower bound for $\ell_0''(\beta_0, \tilde{Z})$ with misspecified values $\tilde{Z}_i$. Thus the term $|\tilde{Z}_i - Z_i|$ critically determines the estimation accuracy. The following bound of $|\tilde{Z}_i - Z_i|$ is helpful for our later analysis.

Lemma 18 *Under the assumptions of Theorem 11, there exists some constant $C_z > 0$ depending on c_m, C_m, C_π, C_μ and C_e such that with probability at least $1 - O(p^{-L})$ we have uniformly for all $i = 1, \ldots, N$*

$$\begin{aligned} \left|\tilde{Z}_i - Z_i\right| &\le \frac{1}{M_i}\left|\log\left(\frac{\hat{\pi}_1\pi_2}{\hat{\pi}_2\pi_1}\right)\right| + \left|\bar{X}_i^T\left(\tilde{\beta} - \beta\right)\right| + \frac{1}{M_i}\left|\sum_{j=1}^{M_i} X_{ij}^T\left(\tilde{\nabla} - \nabla\right) X_{ij}/2\right| \\ &\le C_z\left((1 + \sqrt{\frac{\log p}{m_0}})\left(\|\tilde{\beta} - \beta\| + \|\tilde{\nabla} - \nabla\|_1\right) + \max_{k=1,2}|\pi_k - \hat{\pi}_k|/m_0\right). \end{aligned} \tag{39}$$

Indeed, the conclusion (39) is valid with the same probability $1 - O(p^{-L})$ conditioning on any realization of $\{\mathcal{Y}_i\}_{i=1}^N$ and $\{M_i\}_{i=1}^N$.

Lemma 18 and our assumption imply that with probability at least $1 - O(p^{-L})$ we have $m_0 \max_i |\tilde{Z}_i - Z_i| := R_2$ is sufficiently small.

Note that the expectation of the score function $\ell_0'(\beta_0, Z) = 0$ where $\ell_0'(\beta_0, Z)$ is obtained by replacing $\tilde{Z}_i$ by Z_i in $\ell_0'(\beta_0, \tilde{Z})$, $i = 1, \ldots, N$. We are ready to bound the magnitude of $\ell_0'(\beta_0, \tilde{Z})$,

$$\begin{aligned}\left|\ell_0'(\beta_0, \tilde{Z})\right| &= \left|\frac{1}{N}\Sigma_{i=1}^N \left(\frac{M_i \exp(M_i(\beta_0 + \tilde{Z}_i))}{1 + \exp(M_i(\beta_0 + \tilde{Z}_i))} - \frac{M_i \exp(M_i(\beta_0 + Z_i))}{1 + \exp(M_i(\beta_0 + Z_i))}\right)\right| \\ &\leq \frac{1}{N}\Sigma_{i=1}^N M_i^2 \left|\tilde{Z}_i - Z_i\right| \\ &\leq C_m^2 m_0 R_2, \end{aligned} \tag{40}$$

where the first inequality follows from that the derivative of $\frac{\exp(M_i(\beta_0+\tilde{Z}_i))}{1+\exp(M_i(\beta_0+\tilde{Z}_i))}$ w.r.t. $\tilde{Z}_i$ is always bounded by M_i and the second inequality is due to Condition 2, $M_i \leq C_m m_0$ and definition of R_2.

Moreover, by Condition 4, we have that the expectation of the i.i.d. bounded random variable $\mathrm{Var}(\mathcal{Y}_i \mid \mathcal{X}_i) = \frac{\exp(M_i(\beta_0+Z_i))}{(1+\exp(M_i(\beta_0+Z_i)))^2}$, $i = 1, \ldots, N$, is bounded away from $C_{\log}$. We apply Hoeffding's inequality and the fact $\log p \leq c_0 N$ to obtain that with probability at least $1 - O(p^{-L})$, we have

$$\frac{1}{N}\Sigma_{i=1}^N M_i^2 \left(\frac{\exp\left(M_i(\beta_0 + Z_i)\right)}{1 + \exp\left(M_i(\beta_0 + Z_i)\right)}\right)\left(\frac{1}{1 + \exp\left(M_i(\beta_0 + Z_i)\right)}\right) \geq C'_{low} m_0^2,$$

where the positive constant $C'_{low} > 0$ depends on $C_{\log}$ and L. Since $m_0 \max_i |\tilde{Z}_i - Z_i| := R_2$ is sufficiently small with probability at least $1 - O(p^{-L})$, the union bound argument further implies that

$$\begin{aligned}\ell_0''(\beta_0, \tilde{Z}) &= \frac{1}{N}\Sigma_{i=1}^N M_i^2 \left(\frac{\exp\left(M_i(\beta_0 + \tilde{Z}_i)\right)}{1 + \exp\left(M_i(\beta_0 + \tilde{Z}_i)\right)}\right)\left(\frac{1}{1 + \exp\left(M_i(\beta_0 + \tilde{Z}_i)\right)}\right) \\ &\geq C_{low} m_0^2, \end{aligned} \tag{41}$$

with probability at least $1 - O(p^{-L})$ for some positive constant $C_{low} > 0$.

In the end, plugging (37), (40) and (41) into (38) and applying the union bound argument, we obtain that with probability $1 - O(p^{-L})$,

$$C_{low} C_m^{-2}(e^{-C_m m_0|\tilde{\beta}_0 - \beta_0|} + C_m m_0 \left|\tilde{\beta}_0 - \beta_0\right| - 1) \leq m_0 \left(C_m^2 R_2 + R_1\right)\left|\tilde{\beta}_0 - \beta_0\right|. \tag{42}$$

We apply the following fact

$$e^{-2\gamma/(1-\gamma)} + (1-\gamma)\frac{2\gamma}{1-\gamma} - 1 \geq 0 \text{ for } \gamma \in (0, 1),$$

to (42) and obtain that

$$C_m m_0 \left|\tilde{\beta}_0 - \beta_0\right| \leq \frac{2C_m(C_m^2 R_2 + R_1)/C_{\log}}{1 - C_m(C_m^2 R_2 + R_1)/C_{\log}}.$$

Since $C_m^2 R_2 + R_1$ are sufficiently small, we have that $C_m(C_m^2 R_2 + R_1)/C_{\log} < 1/2$ which implies $C_m m_0 |\tilde{\beta}_0 - \beta_0| < 2$. This fact itself further implies that $(e^{-C_m m_0 |\tilde{\beta}_0 - \beta_0|} + C_m m_0 |\tilde{\beta}_0 - \beta_0| - 1) \geq (C_m m_0 |\tilde{\beta}_0 - \beta_0|)^2/2$. Consequently, (42) implies that

$$\left| \tilde{\beta}_0 - \beta_0 \right| \leq 2 C_{low}^{-1} m_0^{-1} \left(C_m^2 R_2 + R_1 \right),$$

which further completes our proof, together with Lemma 18 (bound of R_2) and the bound of R_1,

$$\left| \tilde{\beta}_0 - \beta_0 \right| \leq C_\delta \left((\|\tilde{\beta} - \beta\| + \|\tilde{\nabla} - \nabla\|_1)(1 + \sqrt{\frac{\log p}{m_0}}) + \max_{k=1,2} |\pi_k - \hat{\pi}_k|/m_0 + \sqrt{\frac{\log p}{N}} \right),$$

where the constant $C_\delta = 2C_{low}^{-1}(C_m^2 C_z + C_r)$. ■

Proof of Theorem 13

Proof Recall that for each $k = 1, 2$, the corresponding Bayes risk and generalization error of CLIPS classifier can be decomposed as

$$\begin{aligned} R_{Bk} &= \sum_{m=c_m m_0}^{C_m m_0} \mathrm{pr}(\phi_B(\mathcal{X}^\dagger) \neq k \mid \mathcal{Y}^\dagger = k, M^\dagger = m) p_M(m) := \sum_m R_{Bk,m} p_M(m), \\ \tilde{R}_k &= \sum_{m=c_m m_0}^{C_m m_0} \mathrm{pr}(\tilde{\phi}(\mathcal{X}^\dagger) \neq k \mid \mathcal{Y}^\dagger = k, M^\dagger = m) p_M(m) := \sum_m \tilde{R}_{k,m} p_M(m). \end{aligned}$$

Therefore, it is sufficient to bound the difference $\tilde{R}_{k,m} - R_{Bk,m}$ for each fixed $k = 1, 2$ and fixed $m \in [c_m m_0, C_m m_0]$.

Recall that $\Xi_N = (1 + \sqrt{\frac{\log p}{m_0}})(\|\tilde{\beta} - \beta\| + \|\tilde{\nabla} - \nabla\|_1) + \max_{k=1,2} |\hat{\pi}_k - \pi_k|/m_0 + |\tilde{\beta}_0 - \beta_0|$. Define the event $\mathcal{E}_0 = \{\Xi_N \leq C_\Xi \xi_N\}$, where $\xi_N = (1 + \sqrt{\frac{\log p}{m_0}})(s_q \lambda'_{1,N} + C_{\ell 1}\sqrt{s_l}\lambda_{2,N}) + \sqrt{\frac{\log p}{N}}$, the constant $C_\Xi = 2(2 + C'')(C_\delta + 1)$ and other constants C'', C_δ can be tracked back from Theorems 6-11. We first show that our estimators satisfy that $\mathrm{pr}(\mathcal{E}_0) = 1 - O(p^{-L})$ by Theorems 6-11. Indeed, Theorems 6-9 provides bounds of $(\|\tilde{\beta} - \beta\| + \|\tilde{\nabla} - \nabla\|_1)$. The estimation error of $\max_{k=1,2} |\hat{\pi}_k - \pi_k|/m_0$ follows from Lemma 14 and is negligible compared to the first term in ξ_N. Assuming these bounds hold, the second part of Condition 5 implies that the assumption in Theorem 11 is satisfied with the initial estimators being our quadratic and linear estimators. Thus Theorem 11 further implies the upper bound for $|\tilde{\beta}_0 - \beta_0|$. Hereafter, we assume event $\mathcal{E}_0$ holds.

We follow the notation introduced in the proof of Theorem 11 on the set of observations $(\mathcal{X}^\dagger, \mathcal{Y}^\dagger)$ and define $\tilde{Z} = \log(\hat{\pi}_1/\hat{\pi}_2)/M^\dagger + \bar{x}^T\tilde{\beta} + \bar{x}^T\tilde{\nabla}\bar{x}/2 + \mathrm{tr}(\tilde{\nabla}S)/2$, which is used to approximate $Z = \log(\pi_1/\pi_2)/M^\dagger + \bar{x}^T\beta + \bar{x}^T\nabla\bar{x}/2 + \mathrm{tr}(\nabla S)/2$, where $\bar{x}$ and S are the sample mean and covariance of the set $\mathcal{X}^\dagger$. Then we define the event $\mathcal{E}_z = \{|\tilde{Z} - Z| \leq C_z \Xi_N\}$. Lemma 18 applied to $(\mathcal{X}^\dagger, \mathcal{Y}^\dagger)$ and the second part of Condition 5 imply that on event $\mathcal{E}_0$ uniformly for all $k = 1, 2$ and $m \in [c_m m_0, C_m m_0]$, we have $\mathrm{pr}(\mathcal{E}_z | \mathcal{Y}^\dagger = k, M^\dagger = m) \geq 1 - C'_g p^{-L}$.

Without loss of generality, we focus on the case $k = 1$. Recall $\tilde{R}_{k,m}$ relies on the estimators $\tilde{\beta}_0, \tilde{\beta}, \tilde{\nabla}, \hat{\pi}_1$ and $\hat{\pi}_2$ and hence is random. On the event $\mathcal{E}_0$, we have that

$$\begin{aligned}
\tilde{R}_{1,m} &= \text{pr}\left(\tilde{Z} + \tilde{\beta}_0 \leq 0 | \mathcal{Y}^\dagger = 1, M^\dagger = m\right) \\
&= \text{pr}\left(Z + \beta_0 \leq Z - \tilde{Z} + \beta_0 - \tilde{\beta}_0 | \mathcal{Y}^\dagger = 1, M^\dagger = m\right) \\
&= \text{pr}\left(Z + \beta_0 \leq Z - \tilde{Z} + \beta_0 - \tilde{\beta}_0, \mathcal{E}_z | \mathcal{Y}^\dagger = 1, M^\dagger = m\right) + \text{pr}(\mathcal{E}_z^c | \mathcal{Y}^\dagger = 1, M^\dagger = m) \\
&\leq C_g' p^{-L} + \text{pr}\left(Z + \beta_0 \leq (C_z + 1)\, \Xi_N, \mathcal{E}_z | \mathcal{Y}^\dagger = 1, M^\dagger = m\right) \\
&\leq C_g' p^{-L} + \text{pr}\left(Z + \beta_0 \leq (C_z + 1)\, C_\Xi \xi_N | \mathcal{Y}^\dagger = 1, M^\dagger = m\right) \\
&= C_g' p^{-L} + F_{1,m}((C_z + 1)\, C_\Xi \xi_N), \qquad (43)
\end{aligned}$$

where the first inequality follows from the conditional probability $\text{pr}(\mathcal{E}_z | \mathcal{Y}^\dagger = k, M^\dagger = m) \geq 1 - C_g' p^{-L}$ and the definition of the event $\mathcal{E}_z$, the second inequality is due to the event $\mathcal{E}_0$, and the last equality follows from the definition of the cumulative distribution function $F_{1,m}(t)$.

In addition, by the definition of the deterministic value $R_{Bk,m}$, we have

$$R_{B1,m} = \text{pr}\left(Z + \beta_0 \leq 0 | \mathcal{Y}^\dagger = 1, M^\dagger = m\right) = F_{1,m}(0). \qquad (44)$$

By our assumption, the quantity $(C_z + 1)\, C_\Xi \xi_N$ is sufficiently small and hence less than δ_0. It follows from (43)-(44) and definition of d_N that on the event $\mathcal{E}_0$, we have that

$$\begin{aligned}
\tilde{R}_{1,m} - R_{B1,m} &\leq C_g' p^{-L} + \sup_{t \in [-\delta_0, \delta_0]} \left| F_{1,m}'(t) \right| (C_z + 1)\, C_\Xi \xi_N \\
&\leq C_g' p^{-L} + (C_z + 1)\, C_\Xi \xi_N d_N.
\end{aligned}$$

Similarly we can show that same upper bound applies to $\tilde{R}_{2,m} - R_{B2,m}$ uniformly for all $m \in [c_m m_0, C_m m_0]$. Therefore on the event $\mathcal{E}_0$, we obtain that $\tilde{R} \leq R_B + C_g' p^{-L} + (C_z + 1)\, C_\Xi \xi_N d_N$, which completes our proof. ■

Appendix 2: Proofs of Supporting Lemmas

Proof of Lemma 14

Proof Recall $n_1 = \sum_{i=1}^N M_i \mathbb{1}\{\mathcal{Y}_i = 1\}$ with $\mathbb{1}\{\mathcal{Y}_i = 1\}$ i.i.d. Bernoulli with probability $\pi_1 \in [C_\pi, 1 - C_\pi]$ and $M_i \in [c_m m_0, C_m m_0]$ with probability 1. Hoeffding's inequality (e.g. Vershynin, 2012, Proposition 5.10) implies that there exists some constant C' depending on C_π and L only such that (iv) holds, i.e. $|\pi_1 - \hat{\pi}_1| \leq C' \sqrt{\frac{\log p}{N}}$ with probability at least $1 - p^{-L}$. Consequently, $n_1 \geq cNm_0$ for some constant c depending on c_m, C_π and L with probability at least $1 - p^{-L}$ given $\log p \leq c_0 N$ and Condition 3. Similar results apply to $\hat{\pi}_2$ and n_2. From now on, we condition on the above event and only need to show (i)-(iii) hold with probability at least $1 - p^{-L}$.

Since $\Sigma_k^{-1/2}(\hat{\mu}_k - \mu_k) \sim N(0, \frac{1}{n_k} I_p)$, the tail probability of Chi-squared distribution (Laurent and Massart, 2000, e.g.) implies that for any $0 < t < 1$, $\mathrm{pr}(|\|\sqrt{n_k}\Sigma_k^{-1/2}(\hat{\mu}_k - \mu_k)\|^2/p - 1| \geq t) \leq 2\exp(pt^2/8)$. Hence, by picking a small t (e.g. $t = 0.1$) as well as Condition 1 and $n_k > cNm_0$, we obtain the result (ii) holds with probability at least $1 - O(p^{-L})$.

In addition, it follows from the Davidson-Szarek bound (e.g. Davidson and Szarek, 2001, Theorem II.7) that for each k, there exists some constant $C > 0$ depending on C_e, L such that $\|\Sigma_k - \hat{\Sigma}_k\|_{\ell_2} < C\sqrt{p/(Nm_0)}$ with probability at least $1 - 2p^{-L}$, given Condition 1 and the fact $p < c_0 Nm_0$ with a sufficiently small c_0. Here $\|\cdot\|_{\ell_2}$ denotes the matrix spectral norm. Consequently, the assumption $p < c_0 Nm_0$ and Condition 1, together with a union bound argument, implies the result (i). Result (iv) also follows, noting that $\|\cdot\|_F \leq \sqrt{p}\|\cdot\|_{\ell_2}$. ■

Proof of Lemma 15

Proof We follow the same argument on bounding n_1 and n_2 as that at the beginning of the proof of Lemma 14. In particular, given $\log p \leq c_0 N$, we have $\mathrm{pr}(n_k \geq cNm_0) = 1 - p^{-L}$ for $k = 1, 2$ and some constant $c > 0$.

Note the distribution of each X_{ij} is independent of n_k. From now on, we condition on n_1 and n_2. Write $X_{ij} = \mathbb{E}X_{ij} + U_{ij}$, where $U_{ij} \sim N(0, \Sigma_{\mathcal{Y}_i})$. Then we have $\hat{\Sigma}_k = (\frac{1}{n_k}\sum_{(i,j):\mathcal{Y}_i=k} U_{ij}U_{ij}^T) - (\mu_k - \hat{\mu}_k)(\mu_k - \hat{\mu}_k)^T$. Since $\hat{\mu}_k - \mu_k \sim N(0, \frac{1}{n_k}\Sigma_k)$, tail probability of normal distribution with union bound implies that for any $L > 0$, there exists some constant $C_1 > 0$ depending on L only such that for $k = 1, 2$,

$$\mathrm{pr}(\|\hat{\mu}_k - \mu_k\|_\infty \geq C_1\sqrt{\frac{(\max_j \sigma_{k,jj})\log p}{n_k}}) \leq p^{-L}. \tag{45}$$

Moreover, since $\mathbb{E}\frac{1}{n_k}\sum_{(i,j):\mathcal{Y}_i=k} U_{ij}U_{ij}^T = \Sigma_k$ and each entry of $U_{ij}U_{ij}^T$ is sub-exponentially distributed, Bernstein's inequality (e.g. Vershynin, 2012, Proposition 5.16) with union bound implies that there exists some constant $C_2 > 0$ depending on L such that

$$\mathrm{pr}(\|\frac{1}{n_k}\sum_{(i,j):\mathcal{Y}_i=k} U_{ij}U_{ij}^T - \Sigma_k\|_\infty \geq C_2\sqrt{\frac{(\max_j \sigma_{k,jj})^2\log p}{n_k}}) \leq p^{-L}. \tag{46}$$

Combining (45) and (46) and probabilities of $n_k \geq cNm_0$, we have obtained both results (i) and (ii) with probability at least $1 - 4p^{-L}$, where the constant $C' > 0$ depends on c_m, C_π, C_e and L only. ■

Proof of Lemma 16

Proof We follow the same argument on bounding n_1 and n_2 as that at the beginning of the proof of Lemma 14. In particular, given $\log p \leq c_0 N$, we have $\mathrm{pr}(n_k \geq cNm_0) = 1 - p^{-L}$ for $k = 1, 2$ and some constant $c > 0$.

Write $X_{ij} = \mathbb{E}X_{ij} + U_{ij}$, where $U_{ij} \sim N(0, \Sigma_{\mathcal{Y}_i})$. We have $\hat{\Sigma}_k = (\frac{1}{n_k}\sum_{(i,j):\mathcal{Y}_i=k} U_{ij}U_{ij}^T) - (\mu_k - \hat{\mu}_k)(\mu_k - \hat{\mu}_k)^T$. Result (i) of Lemma 15 implies that there exists some constant $C_1 > 0$ such that

$$\mathrm{pr}(\|\hat{\mu}_k - \mu_k\|_\infty \geq C_1\sqrt{\frac{\log p}{Nm_0}}) = O(p^{-L}). \tag{47}$$

According to our assumptions, we have $\|\Sigma_k^{-1}\mu_k\| \leq \lambda_{\min}^{-1}(\Sigma_k)\|\mu_k\| \leq C_e C_\mu$. We condition on n_1 and n_2. Then the normality of $\hat{\mu}_k - \mu_k \sim N(0, \Sigma_k/n_k)$ yields that for $k = 1, 2$ and some constant C'' depending on L only, we have $\left|(\mu_k - \hat{\mu}_k)^T \Sigma_k^{-1}\mu_k\right| \geq C''\lambda_{\max}(\Sigma_k)C_e C_\mu\sqrt{\frac{\log p}{n_k}}$ with probability at least $1 - p^{-L}$. Taking union bound with the event $n_k \geq cNm_0$, we obtain that there exists some constant $C_2 > 0$ such that

$$\mathrm{pr}(\left|(\mu_k - \hat{\mu}_k)^T \Sigma_k^{-1}\mu_k\right| \geq C_2\sqrt{\frac{\log p}{Nm_0}}) \leq 2p^{-L}. \tag{48}$$

By our choice of $\lambda_{2,N}$, we have that $\lambda_{2,N}/2 > (C_1 + C_2)\sqrt{(\log p)/(Nm_0)}$. Consequently, given equations (47)-(48), decomposition of Σ_k and $\log p = o(N)$, to conclude (β_1, β_2) is feasible, i.e. $\left\|\hat{\Sigma}_k\beta_k - \hat{\mu}_k\right\| < \lambda_{2,N}$, $k = 1, 2$, we only need to show with probability $1 - O(p^{-L})$ that

$$\|(\frac{1}{n_k}\sum_{(i,j):\mathcal{Y}_i=k} U_{ij}U_{ij}^T)\Sigma_k^{-1}\mu_k - \mu_k\|_\infty < \frac{1}{2}\lambda_{2,N}. \tag{49}$$

Note that the rth coordinate is $\frac{1}{n_k}\sum_{(i,j):\mathcal{Y}_i=k}\left(U_{ij,r}U_{ij}^T\Sigma_k^{-1}\mu_k - \mu_{k,r}\right)$, the sum of i.i.d. centered sub-exponential variable since each summand is the product of two normal variables $U_{i,j}$ and $U_i^T\Sigma_k^{-1}\mu_k$. Moreover, the sub-exponential variable has constant parameter since $U_{ij}^T\Sigma_k^{-1}\mu_k$ and $U_{ij,r}$ have bounded variance. Thus Bernstein's inequality (e.g. Vershynin, 2012, Proposition 5.16) with union bound over all coordinates and the event $n_k \geq cNm_0$ implies that there exists some constant $C_3 > 0$ such that

$$\mathrm{pr}(\|(\frac{1}{n_k}\sum_{(i,j):\mathcal{Y}_i=k} U_{ij}U_{ij}^T)\Sigma_k^{-1}\mu_k - \mu_k\|_\infty > C_3\sqrt{\frac{\log p}{Nm_0}}) \leq 2p^{-L}. \tag{50}$$

By picking a large constant C' in our choice of $\lambda_{2,N}$, we obtain $\lambda_{2,N}/2 > C_3\sqrt{(\log p)/(Nm_0)}$, which completes the proof of (49). ■

Proof of Lemma 18

Proof It is sufficient to show that for any realization of $\{\mathcal{Y}_i\}_{i=1}^N$ and $\{M_i\}_{i=1}^N$, equation (39) is valid for each i with probability at least $1 - O(p^{-L-1})$. Indeed, this fact, together with the union bound argument and $p \geq N$ implies the desired result. The first inequality of (39) follows from the definitions of $\tilde{Z}_i$ and Z_i directly. We show the second inequality holds in the remaining of proof with probability at least $1 - O(p^{-L-1})$ for the fixed i. Without loss of generality, we assume $\mathcal{Y}_i = 1$ and $M_i = m_0 c_m$.

Recall that the initial estimators satisfy $\max_{k=1,2}|\pi_k - \hat{\pi}_k| \leq C_p$ with a sufficiently small constant C_p. Consequently, we have that $\hat{\pi}_1, \hat{\pi}_2 \in [C_\pi/2, 1 - C_\pi/2]$ by Condition 3, which

further yields $\frac{1}{m_0 c_m}\left|\log\left(\frac{\hat\pi_1\pi_2}{\hat\pi_2\pi_1}\right)\right| \le C_{z1}\max_{k=1,2}|\pi_k-\hat\pi_k|/m_0$ with some universal constant C_{z1} depending on c_m and C_π only by the boundedness of $\hat\pi_1/\hat\pi_2$.

To deal with the term $|\bar X_i^T(\tilde\beta-\beta)|$, we note that $\bar X_i \sim N(\mu_1, \Sigma_1/(m_0c_m))$, which implies that $|\bar X_i^T(\tilde\beta-\beta)| \le \|\tilde\beta-\beta\|\cdot\|\mu_1\| + \|\tilde\beta-\beta\|\left(C_e/(m_0c_m)\right)^{1/2}|D|$, where $D\sim N(0,1)$. According to the tail probability of standard normal distribution, we obtain that with probability at least $1-O(p^{-L-1})$, that $|D|\le C_z'\sqrt{\log p}$ where C_z' only depends on L. This fact, together with the assumption $\|\mu_1\|\le C_\mu$ further implies that $|\bar X_i^T(\tilde\beta-\beta)| \le C_{z2}\|\tilde\beta-\beta\|(1+\sqrt{(\log p)/m_0})$ with probability $1-O(p^{-L-1})$, where $C_{z2} = ((C_e/c_m)^{1/2}C_z' + C_\mu)$.

Finally, we provide an upper bound for $\frac{1}{M_i}\left|\sum_{j=1}^{M_i} X_{ij}^T(\tilde\nabla-\nabla)X_{ij}/2\right|$. Since $X_{i1},\ldots,X_{iM_i}$ are i.i.d. copies of $N(\mu_1,\Sigma_1)$, we naturally decompose it into three terms as follows with $U_{ij} := X_{ij}-\mu_1 \sim N(0,\Sigma_1)$

$$\begin{aligned}
&\frac{1}{M_i}\left|\sum_{j=1}^{M_i} X_{ij}^T\left(\tilde\nabla-\nabla\right)X_{ij}/2\right| \\
\le\ & \frac{1}{M_i}\left|\sum_{j=1}^{M_i} U_{ij}^T\left(\tilde\nabla-\nabla\right)U_{ij}/2\right| + \left|\mu_1^T\left(\tilde\nabla-\nabla\right)\mu_1/2\right| + \frac{1}{M_i}\left|\sum_{j=1}^{M_i}\mu_1^T\left(\tilde\nabla-\nabla\right)U_{ij}\right| \qquad (51)
\end{aligned}$$

We deal with these three terms individually. First of all, $|\mu_1^T(\tilde\nabla-\nabla)\mu_1/2| \le C_\mu^2\|\tilde\nabla-\nabla\|_1/2$ by the assumption $\|\mu_1\|\le C_\mu$. Second, the term $(\sum_{j=1}^{M_i}\mu_1^T(\tilde\nabla-\nabla)U_{ij})/M_i$ follows a distribution of $N(0,\mu_1^T(\tilde\nabla-\nabla)\Sigma_1(\tilde\nabla-\nabla)\mu_1/(m_0c_m))$, which yields that with probability at least $1-O(p^{-L-1})$ that

$$\begin{aligned}
\frac{1}{M_i}\left|\sum_{j=1}^{M_i}\mu_1^T\left(\tilde\nabla-\nabla\right)U_{ij}\right| &\le \left(\mu_1^T\left(\tilde\nabla-\nabla\right)\Sigma_1\left(\tilde\nabla-\nabla\right)\mu_1/(m_0c_m)\right)^{1/2} C_z''\sqrt{\log p} \\
&\le C_\mu C_z'\left(C_e/c_m\right)^{1/2}\|\tilde\nabla-\nabla\|_1\sqrt{\frac{\log p}{m_0}},
\end{aligned}$$

where we have used tail probability of standard normal distribution and the last inequality follows from Condition 1. Third, by Hölder's inequality, we have

$$\begin{aligned}
\frac{1}{M_i}\left|\sum_{j=1}^{M_i} U_{ij}^T\left(\tilde\nabla-\nabla\right)U_{ij}/2\right| &= \left|\mathrm{tr}(\left(\tilde\nabla-\nabla\right)\sum_{j=1}^{M_i}U_{ij}^TU_{ij}/M_i)/2\right| \\
&\le \frac{1}{2}\left\|\tilde\nabla-\nabla\right\|_1\left\|\sum_{j=1}^{M_i}U_{ij}^TU_{ij}/M_i\right\|_\infty.
\end{aligned}$$

Since each entry of $\sum_{j=1}^{M_i}U_{ij}^TU_{ij}/M_i-\Sigma_1$ is the sum of centered sub-exponential variable with bounded parameter. The Bernstein's inequality (e.g. Vershynin, 2012, Proposition 5.16) with union bound over all p^2 entries implies that there exists some constant $C_z''>0$ depending on L and C_e only such that $\left\|\sum_{j=1}^{M_i}U_{ij}^TU_{ij}/M_i-\Sigma_1\right\|_\infty \le C_z''\sqrt{\frac{\log p}{c_m m_0}}$ with

probability at least $1-O(p^{-L-1})$. Therefore, we obtain that with probability $1-O(p^{-L-1})$,

$$\frac{1}{M_i}\left|\sum_{j=1}^{M_i} U_{ij}^T\left(\tilde{\nabla}-\nabla\right)U_{ij}/2\right| \leq (C_z''\sqrt{\frac{\log p}{c_m m_0}}+C_e)\|\tilde{\nabla}-\nabla\|_1/2,$$

where we have used $\|\Sigma_1\|_\infty \leq C_e$ by Condition 1. Combining the upper bounds of three terms above, we finally obtain that with probability $1-O(p^{-L-1})$,

$$\frac{1}{M_i}\left|\sum_{j=1}^{M_i} X_{ij}^T\left(\tilde{\nabla}-\nabla\right)X_{ij}/2\right| \leq C_{z3}(\sqrt{\frac{\log p}{m_0}}+1)\|\tilde{\nabla}-\nabla\|_1,$$

where constant $C_{z3} = C_\mu^2/2 + C_\mu C_z' \left(C_e/c_m\right)^{1/2} + (C_e + C_z''/\sqrt{c_m})/2$.

To complete our proof, we combine all bounds for $\frac{1}{m_0 c_m}\left|\log\left(\frac{\hat{\pi}_1\pi_2}{\hat{\pi}_2\pi_1}\right)\right|$, $|\bar{X}_i^T(\tilde{\beta}-\beta)|$ and $\frac{1}{M_i}|\sum_{j=1}^{M_i} X_{ij}^T(\tilde{\nabla}-\nabla)X_{ij}/2|$ with $C_z = C_{z1}+C_{z2}+C_{z3}$. ■

References

Felix Abramovich, Yoav Benjamini, David L Donoho, and Iain M Johnstone. Special invited lecture: adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

Saad Ali and Mubarak Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):288–303, 2010.

Francis Bach et al. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577, 2011.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Yixin Chen, Jinbo Bi, and James Ze Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

Veronika Cheplygina, David MJ Tax, and Marco Loog. On classification with bags, groups and sets. *Pattern Recognition Letters*, 59:11–17, 2015.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces*, 1:317–366, 2001.

Yingying Fan and Jinchi Lv. Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association*, 108(503):1044–1061, 2013.

Yingying Fan, Jiashun Jin, and Zhigang Yao. Optimal classification in sparse Gaussian graphic model. *The Annals of Statistics*, 41(5):2537–2571, 2013.

Yingying Fan, Yinfei Kong, Daoji Li, Zemin Zheng, et al. Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics*, 43(3):1243–1272, 2015.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Sungkyu Jung and Xingye Qiao. A statistical approach to set classification by feature selection with applications to classification of histopathology images. *Biometrics*, 70: 536–545, 2014.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

Quefeng Li and Jun Shao. Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, 25:457–473, 2015.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, pages 570–576, 1998.

JS Marron, Michael J Todd, and Jeongyoun Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Jayson Miedema, James Stephen Marron, Marc Niethammer, David Borland, John Woosley, Jason Coposky, Susan Wei, Howard Reisner, and Nancy E Thomas. Image and statistical analysis of melanocytic histology. *Histopathology*, 61(3):436–444, 2012.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, Harrison H Zhou, et al. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

Noor A. Samsudin and Andrew P. Bradley. Nearest neighbour group-based classification. *Pattern Recognition*, 43(10):3458–3467, October 2010. ISSN 0031-3203. doi: 10.1016/j.patcog.2010.05.010. URL http://dx.doi.org/10.1016/j.patcog.2010.05.010.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *In Compressed Sensing, Theory and Applications*, pages 210–268, 2012.

Boxiang Wang and Hui Zou. Sparse distance weighted discrimination. *Journal of Computational and Graphical Statistics*, 25(3):826–838, 2016.

Wei Wang, John A Ozolek, and Gustavo K Rohde. Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images. *Cytometry Part A*, 77(5):485–494, 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20099247.

Wei Wang, John A. Ozolek, Dejan Slepčev, Ann B. Lee, Cheng Chen, and Gustavo K. Rohde. An optimal transportation approach for nuclear structure-based pathology. *IEEE Transactions on Medical Imaging*, 30(3):621–631, March 2011.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286, 2010.

Sihai Dave Zhao, T Tony Cai, and Hongzhe Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.