Statistica Sinica Preprint No: SS-2020-0253	
Title	Covariance-engaged Classification of Sets via Linear Programming
Manuscript ID	SS-2020-0253
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0253
Complete List of Authors	Zhao Ren, Sungkyu Jung and Xingye Qiao
Corresponding Author	Sungkyu Jung
E-mail	sungkyu@snu.ac.kr

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Covariance-engaged Classification of Sets via Linear Programming

Zhao Ren[1], Sungkyu Jung[2] and Xingye Qiao[3]

[1]*University of Pittsburgh,* [2]*Seoul National University,* [3]*Binghamton University*

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 uses 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. Its theoretical properties are investigated for both the independent case and various (short-range and long-range dependent) time series structures among the observations within each set. The convergence rates of the estimation errors and the risk of the CLIPS classifier are established to show that having multiple observations in a set leads to faster convergence rates than in the standard classification situation in which there is only one observation in the set. The applicable domains in which the CLIPS classifier outperforms its competitors are highlighted in a comprehensive simulation study. Finally, we illustrate the usefulness of the proposed methods in classifying real image data in histopathology.

Key words and phrases: Bayes risk, ℓ_1-minimization, Quadratic discriminant analysis, Set classification, Sparsity.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

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 training data to classify future observations. Some popular classification 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.

Advances in technology have enabled the production of large amounts of data in areas such as the 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). Specifically, Miedema et al. (2012) collected 348 independent cells, each containing observations of varying numbers (tens to hundreds) of nuclei. Here, each cell, rather than each nucleus, is labeled as either normal or cancerous. Each observation of nuclei contains 51 measurements of shape and texture features. A statistical task herein is to classify the whole set of observations from a single set (or all nuclei in a single cell) as normal or cancerous. This problem was referred to as *set classification* by Ning and Karypis (2009) and studied by Wang et al. (2012) and Jung and Qiao (2014). The problem appears in the image-based pathology literature (Samsudin and Bradley, 2010; Wang et al., 2010; Cheplygina et al., 2015; Shifat-E-Rabbi et al., 2020) and in face recognition, based on pictures obtained from multiple cameras, sometimes called image set classification (Arandjelovic and Cipolla, 2006; Wang et al., 2012). The approaches to set classification in the literature are combinations of feature engineering, off-the-shelf

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

classifiers (mostly the support vector machine), and consensus learning (either majority or weighted voting). To the best of the authors' knowledge, there is no theoretical justification for set classification. Set classification is not identical to multiple-instance learning (MIL) (Maron and Lozano-Pérez, 1998; Chen et al., 2006; Ali and Shah, 2010; Carbonneau et al., 2018), as shown by Kuncheva (2010). A key difference is that in set classification, a label is given to sets, whereas observations in a set have different labels in the MIL setting.

While conventional classification methods predict a class label for each observation, care is needed in generalizing the methods 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 between the covariances in order 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 as either $\mathcal{Y} = 1$ or $\mathcal{Y} = 2$. In particular, we propose set classifiers that extend a quadratic discriminant analysis to the set-classification setting, and that are designed to work well in the set classification of high-dimensional data with distributions similar to those in Fig. 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

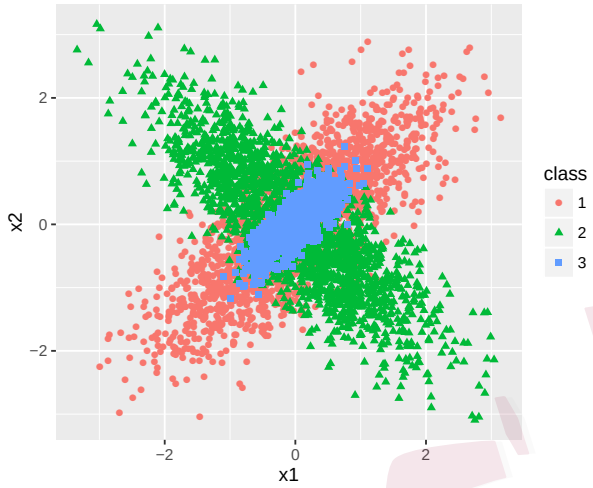

Figure 1: A two-dimensional toy example showing classes with no difference in the mean or marginal variance.

To provide a fundamental understanding of the set-classification problem, we establish a Bayesian optimal decision rule under normality and homogeneity (independent and identically distributed; i.i.d.) assumptions. This Bayes rule uses the covariance structure of the testing set of future observations. We show in Section 2 that it becomes much easier to accurately classify 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.

Based on the Bayesian optimal decision rule, we propose new methods of set classification in Section 3. 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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

parameters to be estimated ($\approx p^2$) is large, we assume that only a few parameters drive the difference between the two classes. With this sparsity assumption, we propose estimating the parameters in the classifier using linear programming, referring to the resulting classifiers as Covariance-engaged LInear Programming Set (CLIPS) classifiers. Specifically, the quadratic and linear parameters in the Bayes rule can be estimated efficiently under the sparse structure, owing to the extra observations in the training set resulting from having sets of observations. Our estimation approaches are closely related to and built upon the successful estimation strategies of Cai et al. (2011) and Cai and Liu (2011). To estimate the constant parameter, we perform a logistic regression with only one unknown, given the estimates of the quadratic and linear parameters. This allows us to implement the CLIPS classifier with high computation efficiency.

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

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; Jiang et al., 2018; Qin, 2018; Zou, 2019; Gaynanova and Wang, 2019; Cai and Zhang, 2019; Pan and Mai, 2020). As a byproduct of our theoretical study, we show that the new QDA method enjoys better theoretical properties than those

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

of some state-of-the-art sparse QDA methods, such as that of Fan et al. (2015).

The advantages of our set classifiers are demonstrated in comprehensive simulation studies. Moreover, in Section 5, we provide an application to histopathology where we classify sets of nucleus images as normal or cancerous tissue. The proofs of the main results and the technical lemmas can be found in the Supplementary Material, as well as a study on the case where the observations in a set demonstrate certain spatial and temporal dependent structures. There, we use various (both short- and long-range) dependent time series structures within each set by considering a very general vector linear process model.

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 vary between 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.

To formally introduce the set-classification problem and study its fundamental properties, we start with a setting in which 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 independently in the same way as $(\mathcal{X}, \mathcal{Y})$. To describe the generating process

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

of $(\mathcal{X}, \mathcal{Y})$, we assume that the random variables M and $\mathcal{Y}$ are independent, denote the marginal class probabilities by $\pi_1 = \text{pr}(\mathcal{Y} = 1)$ and $\pi_2 = \text{pr}(\mathcal{Y} = 2)$, and denote the marginal distribution of the set size M by p_M. In other words, the class membership $\mathcal{Y}$ cannot be predicted based only on the set size M. Conditioned on $M = m$ and $\mathcal{Y} = y$, the observations $X_1, X_2, \ldots, X_M$ in the set $\mathcal{X}$ are independent, and each is distributed as f_y.

2.1 Covariance-engaged Set Classifiers

Suppose 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 the testing set), and its true class label is $\mathcal{Y}^\dagger$. The Bayes optimal decision rule classifies the set $\mathcal{X}^\dagger = \{x_1, \ldots, x_m\}$ as Class 1 if the conditional class probability of Class 1 is greater than that of Class 2; that is, $\text{pr}(\mathcal{Y}^\dagger = 1 \mid M^\dagger = m,\ X_j^\dagger = x_j,\ j = 1, \ldots, m) > 1/2$. This is equivalent to $\pi_1 p_M(m) \prod_{j=1}^m f_1(x_j) > \pi_2 p_M(m) \prod_{j=1}^m f_2(x_j)$, owing to the 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 rule depends on the quantity

$$\begin{aligned} g(x_1, \ldots, x_m) &= \frac{1}{m} \log \left\{ \frac{\pi_1 p_M(m) \prod_{j=1}^m f_1(x_j)}{\pi_2 p_M(m) \prod_{j=1}^m f_2(x_j)} \right\} \\ &= \frac{1}{m} \log(\pi_1/\pi_2) - \frac{1}{2} \log(|\Sigma_1|/|\Sigma_2|) - \frac{1}{2} \mu_1^T \Sigma_1^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma_2^{-1} \mu_2 \\ &\quad + (\Sigma_1^{-1} \mu_1 - \Sigma_2^{-1} \mu_2)^T \bar{x} + \frac{1}{2} \bar{x}^T (\Sigma_2^{-1} - \Sigma_1^{-1}) \bar{x} + \frac{1}{2} \text{tr}\{(\Sigma_2^{-1} - \Sigma_1^{-1}) S\}. \quad (2.1) \end{aligned}$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

of observations m and the i.i.d. observations x_j, for $j = 1, \ldots, m$. The Bayes rule can be expressed as

$$\phi_B(\mathcal{X}^\dagger) = 2 - \mathbb{1}\{g(x_1, \ldots, x_m) > 0\}, \text{ where} \tag{2.2}$$
$$g(x_1, \ldots, x_m) = \frac{1}{m}\log(\pi_1/\pi_2) + \beta_0 + \beta^T\bar{x} + \bar{x}^T\nabla\bar{x}/2 + \mathrm{tr}(\nabla S)/2,$$

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 (2.2) uses the summary statistics m, $\bar{x}$, and S of $\mathcal{X}^\dagger$.

We refer to (2.2) and any estimated version of it as a covariance-engaged set classifier. In Section 3, several estimation approaches for β_0, β, and ∇ are proposed. In this section, we discuss a rationale for considering (2.2).

The covariance-engaged set classifier (2.2) resembles the conventional QDA classifier. As a natural alternative to (2.2), one may consider the sample mean $\bar{x}$ as a representative of the testing set, and apply the QDA to $\bar{x}$ directly to make a prediction. In other words, we 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

$$\phi_{B,\bar{x}}(\mathcal{X}^\dagger) = 2 - \mathbb{1}\{g_{\mathrm{QDA}}(\bar{x}) > 0\}, \text{ where} \tag{2.3}$$
$$g_{\mathrm{QDA}}(\bar{x}) = \frac{1}{m}\log(\pi_1/\pi_2) + \beta_0' + \beta^T\bar{x} + \bar{x}^T\nabla\bar{x}/2,$$

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 (2.2) and (2.3) is that the term $\mathrm{tr}(\nabla S)/2$ is absent from (2.3). Indeed, the advantage of (2.2) over (2.3) comes from the extra information in the sample covariance S of $\mathcal{X}^\dagger$. In the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

regular classification setting, (2.2) coincides with (2.3), because $\mathrm{tr}(\nabla S)/2$ 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:

$$\phi_{MV}(\mathcal{X}^\dagger) = 2 - \mathbb{1}\left\{\frac{1}{m}\sum_{j=1}^{m}\mathrm{sign}[g_{\mathrm{QDA}}(x_j)] > 0\right\}, \tag{2.4}$$

where $\mathrm{sign}(t) = 1, 0, -1$ for $t > 0,\ t = 0$, and $t < 0$ respectively, and $g_{\mathrm{QDA}}(x_j)$ is given in (2.3) with $\bar{x}$ replaced by x_j (and m by one). In contrast, because $g(\mathcal{X}^\dagger) = \frac{1}{m}\sum_{j=1}^{m} g_{\mathrm{QDA}}(x_j)$, our classifier (2.2) predicts the class label using a weighted vote of individual QDA decisions. In this sense, the majority voting scheme (2.4) can be viewed as a discretized version of (2.2). In Section 5, we demonstrate that our set classifier (2.2) performs significantly better than (2.4).

Remark 1. We have assumed that M and $\mathcal{Y}$ are independent in this 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 remains the same, except that these distributions replace two identical factors $p_M(m)$ in the first equality of (2.1). If $p_{M1}(m)$ and $p_{M2}(m)$ are significantly different, then the classification is easier, because one can make a decision based on the observed value of m. Here, we consider only the more difficult setting where $\mathcal{Y}$ and M are independent.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

2.2 Bayes Risk

In this section, we describe an advantage of having a set of observations for prediction, rather than o having a single observation. For this, we suppose for now that the parameters μ_k and Σ_k, for $k = 1, 2$, are known and make the following assumptions. Denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ as the greatest and smallest eigenvalues, respectively, 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 $C_e^{-1} \leqslant \lambda_{\min}(\Sigma_k) \leqslant \lambda_{\max}(\Sigma_k) \leqslant C_e$, 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 is considered.

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

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 (2.2), 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Theorem 1. *Suppose that Conditions 1–3 hold. If $D_p^2 m_0$ is sufficiently large, then $R_B \leqslant 4\exp\left(-c' m_0 D_p^2\right)$, 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 1 provides a Bayes risk bound $R_B \leqslant 4\exp\left(-c' D_p^2\right)$ for the theoretical QDA classifier in the regular classification setting. To guarantee a small Bayes risk for the 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.

We have discussed an alternative classifier using the sample mean $\bar{x}$ as a representative of the testing set, leading to $\phi_{B,\bar{x}}$ (2.3). The following proposition quantifies its risk, which has a slower rate than that of the Bayes classifier R_B.

Proposition 1. *Suppose that Conditions 1–3 hold. Denote the risk of classifier $\phi_{B,\bar{x}}$ in (2.3) as $R_{\bar{x}}$. Assume $\|\nabla\|_F^2 + m_0\|\delta\|^2$ is sufficiently large. Then, $R_{\bar{x}} \leqslant 4\exp\left(-c'(\|\nabla\|_F^2 + m_0\|\delta\|^2)\right)$, for some small constant $c' > 0$ depending on C_e, c_m, and C_π only. In addition, the rate on the exponent cannot be improved in general, that is, $R_{\bar{x}} \geqslant \exp\left(-c''(\|\nabla\|_F^2 + m_0\|\delta\|^2)\right)$, for some small constant $c'' > 0$.*

Remark 2. Compared with the result in Theorem 1, the above proposition implies that the classifier $\phi_{B,\bar{x}}$ 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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

natural $\bar{x}$-based classification rule relies only on the first moment of the data set $\mathcal{X}^\dagger$, while the sufficient statistics, the first two moments, are used fully by the covariance-engaged classifier in (2.2).

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 $p \leqslant c_0 m_0 N$, 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 estimated consistently using 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, owing to the homogeneity assumption. Note that 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 into (2.2). It is natural to use the maximum likelihood estimators, given n_k,

$$\hat{\mu}_k = \sum_{(i,j):\mathcal{Y}_i=k} X_{ij}/n_k \text{ and } \hat{\Sigma}_k = \sum_{(i,j):\mathcal{Y}_i=k} \{(X_{ij} - \hat{\mu}_k)(X_{ij} - \hat{\mu}_k)^T\}/n_k. \tag{3.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

For the classification of $\mathcal{X}^\dagger = \{X_1^\dagger, \ldots, X_{M^\dagger}^\dagger\}$, with $M^\dagger = m$ and $X_i^\dagger = x_i$, the set classifier (2.2) is estimated as

$$\hat{\phi}(\mathcal{X}^\dagger) = 2 - \mathbb{1}\left\{\frac{1}{m}\log(\hat{\pi}_1/\hat{\pi}_2) + \hat{\beta}_0 + \hat{\beta}^T\bar{x} + \bar{x}^T\hat{\nabla}\bar{x}/2 + \mathrm{tr}(\hat{\nabla}S)/2 > 0\right\}, \tag{3.2}$$

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 (3.2), we have assumed $p < n_k$, so that $\hat{\Sigma}_k$ is invertible.

The generalization error of the set classifier (3.2) 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 2 reveals a theoretical property of $\hat{R}$ in a moderate-dimensional setting that allows p, N, and m_0 to grow jointly. This includes the traditional setting in which p is fixed.

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

In Theorem 2, 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 the QDA, where $M_i = M^\dagger \equiv 1$ and $m_0 = 1$, was obtained in Li and Shao (2015) under a stronger assumption $p^2/(ND_p^2) \to 0$.

For high-dimensional data where $p = p(N) \gg Nm_0$, and hence $p > n_k$ with probability one for $k = 1, 2$, by Condition 2, it is problematic to plug in the estimators (3.1) because $\hat{\Sigma}_k$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

is rank deficient with probability one. A simple remedy is to use a diagonalized or enriched version of $\hat{\Sigma}_k$, defined by $\hat{\Sigma}_{k(d)} = \text{diag}\{(\hat{\sigma}_{k,ii})_{i=1,\ldots,p}\}$ 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 the best of our knowledge, no theoretical guarantee has been obtained without some structural assumptions.

3.2 A Direct Approach using 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. The naive plug-in method proposed in Section 3.1 ignores this 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 (2.2) implies that the dimension of the problem can be significantly reduced, which makes consistency possible, even in a high-dimensional setting.

We propose directly estimating the quadratic term ∇, the linear term β, and the constant β_0 coefficients, taking advantage of the assumed sparsity. Because the estimates are calculated efficiently using linear programming, the resulting classifiers are called 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 techniques developed in the literature on 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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

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 using a two-step thresholded estimator.

As the first step, we adopt the CLIME estimator (Cai et al., 2011) to obtain the initial estimators $\tilde{\Omega}_1$ and $\tilde{\Omega}_2$ of the precision matrices Σ_1^{-1} and Σ_2^{-1}, respectively. 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, respectively, of a $p \times p$ matrix B. The CLIME estimators are defined as

$$\tilde{\Omega}_k = \underset{\Omega \in \mathbb{R}^{p\times p}}{\operatorname{argmin}} \|\Omega\|_1 \text{ subject to } \|\hat{\Sigma}_k \Omega - I\|_\infty < \lambda_{1,N},\ k = 1, 2, \tag{3.3}$$

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

$$\tilde{\nabla}_{ij} = \min\{\breve{\nabla}_{ij}, \breve{\nabla}_{ji}\}, \breve{\nabla}_{ij} = (\tilde{\Omega}_{2,ij} - \tilde{\Omega}_{1,ij}) \mathbb{1}\Big\{\Big|\tilde{\Omega}_{2,ij} - \tilde{\Omega}_{1,ij}\Big| > \lambda'_{1,N}\Big\}, \tag{3.4}$$

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

Although this thresholded CLIME difference estimator is obtained by first individually estimating Σ_k^{-1}, note that the estimation accuracy depends only on the sparsity of their difference ∇, rather than on the sparsity of either Σ_1^{-1} or Σ_2^{-1}, under a relatively mild sparsity condition in terms of their matrix ℓ_1 norms. We show in Theorem 3 in Section 4 that if the true precision matrix difference ∇ is negligible, $\tilde{\nabla} = 0$ with high probability. When $\tilde{\nabla} = 0$, our method described in (3.8) becomes a linear classifier adaptively. The computation of $\tilde{\nabla}$ (3.4) is fast, because the first step (CLIME) can be recast as a linear program, and the second step is a simple thresholding procedure.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Remark 3. 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 minimizing the vector ℓ_1 norm of ∇. Specifically, they proposed $\tilde{\nabla}^{ZCL} \in \operatorname{argmin}_B \|B\|_1$, subject to $\|\hat{\Sigma}_1 B\hat{\Sigma}_2 - \hat{\Sigma}_1 + \hat{\Sigma}_2\|_\infty \leqslant \lambda''_{1,n}$, where $\lambda''_{1,n}$ is some thresholding level. This method, however, is computationally expensive (because it has $O(p^2)$ number of linear constraints when cast to linear programming) and can only handle a relatively small size of p. Cai and Zhang (2019) further considered a symmetric version of the above direct estimation, and solved it using a primal-dual interior point method. See also Jiang et al. (2018). We chose to use (3.4), mainly because of the fast computation.

Next we estimate the linear coefficient vector $\beta = \beta_1 - \beta_2$, where $\beta_k = \Sigma_k^{-1}\mu_k$, for $k = 1, 2$. In the literature on 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 are placed on both β_1 and β_2 (see, e.g., Cai and Liu, 2011; Fan et al., 2015). In the latter case, β is also sparse because it is the difference between two sparse vectors. For the estimation of β, we propose a new method that 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 4 for details.)

The true parameter β_k satisfies $\Sigma_k\beta_k - \mu_k = 0$. However, owing to the rank-deficiency of $\hat{\Sigma}_k$, there are either none or infinitely many θ_k 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 (3.1). We relax this constraint and seek a possibly nonsparse pair (θ_1, θ_2) with the smallest ℓ_1 norm difference. We estimate the coefficients β

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

by $\tilde{\beta} = \tilde{\beta}_1 - \tilde{\beta}_2$, where

$$(\tilde{\beta}_1, \tilde{\beta}_2) = \underset{(\theta_1,\theta_2):\|\theta_k\|_1 \leqslant L_1}{\operatorname{argmin}} \|\theta_1 - \theta_2\|_1 \text{ subject to } \|\hat{\Sigma}_k\theta_k - \hat{\mu}_k\|_\infty < \lambda_{2,N},\ k = 1, 2, \tag{3.5}$$

where L_1 is some sufficiently large constant, introduced only to ease the theoretical evaluations. In practice, the constraint $\|\theta_k\|_1 \leqslant L_1$ can be removed without affecting the solution. Our procedure (3.5) can be recast as a linear programming problem (see, e.g., Candes and Tao, 2007; Cai and Liu, 2011) and is computationally efficient.

The direct estimation approach for $\beta = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2$ above is a natural extension of Cai and Liu (2011), in which a direct estimation of $\Sigma^{-1}(\mu_1 - \mu_2)$ for the LDA ($\Sigma = \Sigma_1 = \Sigma_2$) was considered. Note that by centering the quadratic Bayes discriminant function $g_{\text{QDA}}(\cdot)$, alternative sparse linear coefficient vectors have been considered in the literature on QDA. For example, Jiang et al. (2018) proposed estimating $(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2)$, while Li and Shao (2015), Fan et al. (2015), and Cai and Zhang (2019) proposed estimating $\Sigma_2^{-1}(\mu_1 - \mu_2)$, both of which are location-invariant. Although β considered in our approach is not location-invariant, we emphasize that the sparsity conditions for the three different linear coefficient vectors are not comparable, because their interpretations differ. Other direct estimation approaches of the linear coefficient vector have also been considered in related discriminant analyses, see, for example, Clemmensen et al. (2011), Witten and Tibshirani (2011) and ,Mai et al. (2012, 2019).

Finally, we consider the estimation of the constant coefficient β_0. The conditional class probability $\eta(x_1, \ldots, x_m) = \text{pr}(\mathcal{Y} = 1 \mid M = m,\ X_i = x_i,\ i = 1, \ldots, m)$ that a set belongs

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

to Class 1 given $\mathcal{X} = \{x_1, \ldots, x_m\}$ can be evaluated by the following logit function:

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

where $\bar{x}$ and S are the sample mean and thecovariance of the set $\{x_1, \ldots, x_m\}$, respectively. Having obtained our estimators $\tilde{\nabla}$ and $\tilde{\beta}$ from (3.4) and (3.5), respectively, and estimated $\hat{\pi}_1$ and $\hat{\pi}_2$ by N_1/N and N_2/N, respectively, from the training data, only the scalar β_0 is undecided. We may estimate $\tilde{\beta}_0$ by conducting a simple logistic regression with a dummy independent variable M_i, and offset $\log(\hat{\pi}_1/\hat{\pi}_2) + M_i\left(\bar{X}_i^T\tilde{\beta} + \bar{X}_i^T\tilde{\nabla}\bar{X}_i/2 + \mathrm{tr}(\tilde{\nabla}S_i)/2\right)$ for the ith set of observations in the training data, where M_i, $\bar{X}_i$, and S_i are the sample size, sample mean, and sample covariance, respectively, of the ith set. In particular, we solve

$$\tilde{\beta}_0 = \underset{\theta_0\in\mathbb{R}}{\mathrm{argmin}}\ \ell(\theta_0 \mid \{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N, \tilde{\beta}, \tilde{\nabla}), \text{where the negative log-likelihood is} \tag{3.6}$$

$$\ell(\theta_0 \mid \{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N, \tilde{\beta}, \tilde{\nabla}) \tag{3.7}$$
$$= \frac{1}{N}\sum_{i=1}^N\left((\mathcal{Y}_i - 2)M_i\left(\theta_0 + \frac{\log(\hat{\pi}_1/\hat{\pi}_2)}{M_i} + \bar{X}_i^T\tilde{\beta} + \bar{X}_i^T\tilde{\nabla}\bar{X}_i/2 + \mathrm{tr}(\tilde{\nabla}S_i)/2\right)\right.$$
$$\left. + \log\left[1 + \exp\left\{M_i\left(\theta_0 + \frac{\log(\hat{\pi}_1/\hat{\pi}_2)}{M_i} + \bar{X}_i^T\tilde{\beta} + \bar{X}_i^T\tilde{\nabla}\bar{X}_i/2 + \mathrm{tr}(\tilde{\nabla}S_i)/2\right)\right\}\right]\right)$$

Because there is only one independent variable in the logistic regression above, the optimization can be easily and efficiently solved. Alternative ways of estimating the constant coefficient in the literature on QDA include a simple plug-in estimator (Cai and Zhang, 2019) and using the idea of cross-validation (Jiang et al., 2018).

For the purpose of evaluating theoretical properties, we apply the sample splitting technique (Wasserman and Roeder, 2009; Meinshausen and Bühlmann, 2010). Specifically, we

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

randomly choose the first batch of $N_1/2$ and $N_2/2$ sets from two classes in the training data to obtain the estimators $\tilde{\nabla}$ and $\tilde{\beta}$ using (3.4) and (3.5), respectively. Then, $\tilde{\beta}_0$ is estimated based on the second batch, along with $\tilde{\nabla}$ and $\tilde{\beta}$, using (3.6). We plug all estimators in (3.4), (3.5), and (3.6) into the Bayes decision rule (2.2) and obtain the CLIPS classifier,

$$\tilde{\phi}(\mathcal{X}^\dagger) = 2 - \mathbb{1}\left\{\frac{\log(\hat{\pi}_1/\hat{\pi}_2)}{m} + \tilde{\beta}_0 + \tilde{\beta}^T\bar{x} + \bar{x}^T\tilde{\nabla}\bar{x}/2 + \mathrm{tr}(\tilde{\nabla}S)/2 > 0\right\}, \tag{3.8}$$

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

4. Theoretical Properties of the CLIPS classifier

In this section, we derive the theoretical properties of the estimators from (3.4)–(3.6), as well as generalization errors for the CLIPS classifier (3.8). In particular, we demonstrate the advantages of having sets of independent observations, in contrast to the classical QDA setting with individual observations under the homogeneity assumption of Section 2. Parallel results under various time series structures can be found in the Supplementary Material.

To establish the statistical properties of the thresholded CLIME difference estimator $\tilde{\nabla}$ defined in (3.4), we assume that the true quadratic parameter $\nabla = \Sigma_2^{-1} - \Sigma_1^{-1}$ has no more than s_q nonzero entries,

$$\nabla \in \mathcal{FM}_0(s_q) = \{A = (a_{ij}) \in \mathbb{R}^{p\times p}, \text{symmetric} : \sum_{i,j=1}^{p} \mathbb{1}\{a_{ij} \neq 0\} \leqslant s_q\}. \tag{4.1}$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Theorem 3. *Suppose Conditions 1–3 hold. Moreover, assume* $\nabla \in \mathcal{FM}_0(s_q)$, *and* $\|\Sigma_k^{-1}\|_{\ell_1} \leqslant C_{\ell 1}$, *with some constant* $C_{\ell 1} > 0$, *for* $k = 1, 2$, *and* $\log p \leqslant c_0 N$, *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*

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

as long as $\lambda_{1,N} \geqslant CC_{\ell 1}\sqrt{\frac{\log p}{Nm_0}}$ *and* $\lambda'_{1,N} \geqslant 8C_{\ell 1}\lambda_{1,N}$ *in (3.4), 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 4. The parameter space $\mathcal{FM}_0(s_q)$ can be extended easily to 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 3 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 consider only ℓ_0 balls in this work.

Remark 5. Theorem 3 implies that the error bounds of estimating ∇ under the vector ℓ_1 norm and the Frobenius norm both 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 3, provided in the Supplementary Material, partially follows from Cai et al. (2011).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

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

$$\beta \in \mathcal{F}_0(s_l) = \{\alpha = (a_j) \in \mathbb{R}^p : \sum_{j=1}^{p} \mathbb{1}\{\alpha_j \neq 0\} \leqslant s_l\}. \tag{4.2}$$

Theorem 4 gives the rates of convergence of the linear coefficient estimator $\tilde{\beta}$ in (3.5) 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 4. *Suppose Conditions 1–3 hold. Moreover, assume that* $\beta \in \mathcal{F}_0(s_l)$, $\log p \leqslant c_0 N$, $\|\beta_k\|_1 \leqslant C_\beta$, *and* $\|\mu_k\| \leqslant C_\mu$, *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*

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

as long as $\lambda_{2,N} \geqslant C' \sqrt{\frac{\log p}{N m_0}}$ *in (3.5), where* $\max\{\|\Sigma_1^{-1}\|_{\ell_1}, \|\Sigma_2^{-1}\|_{\ell_1}\} \leqslant C_{\ell 1}$ *and* C'', C' *depend on* $L, C_e, c_m, C_\pi, C_\beta$, *and* C_μ *only.*

Remark 6. The parameter space $\mathcal{F}_0(s)$ can be extended easily into an ℓ_q ball or weak ℓ_q ball with $0 < q < 1$ as well, and the results in Theorem 4 remain valid with appropriate sparsity parameters. We focus on $\mathcal{F}_0(s)$ to ease the analysis.

Lastly, we derive the rate of convergence for estimating the constant coefficient β_0. Because $\tilde{\beta}_0$ is obtained by maximizing the log-likelihood function after plugging $\tilde{\beta}$ and $\tilde{\nabla}$ into (3.6), the behavior of our estimator $\tilde{\beta}_0$ critically depends on the accuracy of estimating β and ∇. Theorem 5 provides the result for $\tilde{\beta}_0$ based on certain general initial estimators $\tilde{\beta}$ and $\tilde{\nabla}$, with the following mild condition.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Condition 4. The expectation of the conditional variance of the 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 5. *Suppose Conditions 1–4 hold,* $\log p \leqslant c_0 N$ *with some sufficiently small constant* $c_0 > 0$*, and* $\|\mu_k\| \leqslant C_\mu$ *with some constant* $C_\mu > 0$*, for* $k = 1, 2$*. In addition, we have some initial estimators* $\tilde{\beta}$*,* $\tilde{\nabla}$*,* $\hat{\pi}_1$*, and* $\hat{\pi}_2$ *such that* $m_0(1 + \sqrt{(\log p)/m_0})\|\tilde{\beta} - \beta\| + m_0(1 + (\log p)/m_0)\|\tilde{\nabla} - \nabla\|_1 + \max_{k=1,2} |\pi_k - \hat{\pi}_k| \leqslant C_p$ *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*

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

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

Remark 7. Condition 4 is determined by our data-generating process stated in Section 2.1. It is satisfied when the classification problem is nontrivial. 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 one for $\mathcal{X}$, which is often seen in the literature on logistic regression. See, for example, Fan and Lv (2013) and Fan et al. (2015).

Theorems 3, 4, and 5 demonstrate the estimation accuracy for the quadratic, linear, and constant coefficients, respectively, in our CLIPS classifier (3.8). We conclude this section by establishing an oracle inequality for its generalization error by providing a rate of convergence of the excess risk. To this end, we define the generalization error of the 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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

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 first introduce some notation related to the Bayes decision rule in (2.2). Recall that given $M^\dagger = m$, the Bayes decision rule $\phi_B(\mathcal{X}^\dagger)$ depends solely 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 zero 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 3, 4, and 5, with probability at least $1 - O(p^{-L})$, our estimators satisfy that

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

where $\kappa_N := (1 + (\log p)/m_0)s_q\lambda'_{1,N} + (1 + \sqrt{(\log p)/m_0})C_{\ell 1}\sqrt{s_l}\lambda_{2,N} + \sqrt{(\log p)/(Nm_0^2)}$. The quantity $\kappa_N d_N$ is the key to obtaining the oracle inequality. Condition 5 guarantees that the assumptions of Theorem 5 are satisfied with high probability in our settings.

Condition 5. Suppose $\kappa_N m_0 \leqslant c_0$ and $\kappa_N d_N \leqslant c_0$, with some sufficiently small constant $c_0 > 0$.

Theorem 6 reveals the oracle property of the CLIPS classifier, and provides a rate of convergence of the excess risk, that is, the generalization error of the CLIPS classifier less the Bayes risk R_B defined in Section 2.2.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Theorem 6. *Suppose that the assumptions of Theorems 3 and 4 hold, and that Conditions 4–5 also hold. Then, with probability at least* $1 - O(p^{-L})$*, we have the oracle inequality*

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

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

Theorem 6 implies that, with high probability, the generalization error of the CLIPS classifier is close to the Bayes risk with a rate of convergence no slower than $\kappa_N d_N$. In particular, whenever 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 $\kappa_N d_N$ in the order of

$$(1 + \sqrt{(\log p)/m_0})\sqrt{\log p/(m_0 N)}\sqrt{s_l} + (1 + (\log p)/m_0)\sqrt{\log p/(m_0 N)}s_q. \tag{4.3}$$

The advantage of having large m_0 can be understood by investigating (4.3) as a function of m_0. Indeed, the leading term of (4.3) is

$$\begin{aligned}
&\frac{\log p}{m_0^{3/2}}\sqrt{\frac{\log p}{N}}s_q, \ \text{ if } m_0 \leqslant \log p \cdot \min\{1, \frac{s_q^2}{s_l}\};\\
&\frac{\sqrt{\log p}}{m_0}\sqrt{\frac{\log p}{N}}\sqrt{s_l}, \ \text{ if } \log p \cdot \frac{s_q^2}{s_l} \leqslant m_0 \leqslant \log p;\\
&\sqrt{\frac{1}{m_0}}\sqrt{\frac{\log p}{N}}(\sqrt{s_l} + s_q), \ \text{ if } \log p \leqslant m_0.
\end{aligned}$$

To illustrate the decay rate, we assume $s_l \geqslant s_q^2$. Then, as m_0 increases, the error decreases at the order of $m_0^{3/2}$ up to a certain point $\log p \cdot \frac{s_q^2}{s_l}$, and then decreases at the order of m_0 up to another point $\log p$. When m_0 is large enough that $m_0 \geqslant \log p$, the error decreases at the order of $\sqrt{m_0}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

To further emphasize the advantage of having sets of observations, we compare a general case $m_0 = m^*$, where $\log p \leqslant m^*$, with the special case that $m_0 = 1$, that is, the regular QDA situation. Then, the quantity κ_N with m^* has a faster decay rate, with a factor of order between $\sqrt{m^* \log p}$ and $\sqrt{m^*} \log p$ (depending on the relationship between s_l and s_q), compared to the $m_0 = 1$ case, owing to the extra observations within each set.

The above discussion reveals that in a high-dimensional setting, the benefit of the set classification cannot be simply explained by having $N^* = Nm_0$ independent observations instead of having only N individual observations, as in the classical QDA setting. Indeed, if we have N^* individual observations in the classical QDA setting, then the implied rate of convergence would be either $\log p \sqrt{\frac{\log p}{Nm_0}} s_q$ (if $\log p \cdot s_q^2 \geqslant s_l$) or $\sqrt{\log p} \sqrt{\frac{\log p}{Nm_0}} \sqrt{s_l}$ (otherwise), which is slower than that provided in equation (4.3).

Remark 8. Note that even in the special QDA situation where $m_0 = 1$, owing to the sharper analysis, our result is still new, and the established rate of convergence $(\log p)/N^{1/2}\sqrt{s_l} + (\log p)^{3/2}/N^{1/2} s_q$ in Theorem 6 is at least as good as the $(\log p)^{3/2}/N^{1/2}(s_q + s_l)$ derived in the oracle inequality of Fan et al. (2015) under similar assumptions. Whenever $s_l > s_q$, our rate is even faster, with a factor of order $\sqrt{s_l \log p}$, than that in Fan et al. (2015).

Remark 9. The results in this section, including Theorem 6, demonstrate the advantages of the set-classification setting in contrast to the classical QDA setting. When multiple observations within each set have short-range dependence, the rates of convergence for estimating the key parameters and the oracle inequality resemble the results under the independent assumption. However, the results change significantly when there is a long-range dependence structure among multiple observations.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Remark 10. Cai and Zhang (2019) considered a sparse QDA using a constrained convex optimization approach, establishing a minimax rate of convergence $(s_l + s_q)(\log p \cdot \log^2 N)/N$ on the excess risk up to a logarithmic factor under similar sparsity assumptions. In contrast, our result in the special QDA situation has the rate of convergence discussed in Remark 8, which is slower for most scenarios under different assumptions. It would be interesting to investigate the optimal convergence rates for set classification under both short-range (including i.i.d.) and long-range dependence structures in future studies.

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 $\hat{\Sigma}_k$ (labeled as Plugin(d) and Plugin(e), respectively) introduced at the end of Section 3.1, and plug them into the Bayes rule (2.2), as done in (3.2). For comparison, 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 in Miedema et al. (2012). Then a support vector machine (SVM; Cortes and Vapnik, 1995) and a distance-weighted discrimination (DWD; Marron et al., 2007; Wang and Zou, 2018) are applied to the features to make predictions (labeled SVM and DWD, respectively). We use the R library `clime` to calculate the CLIME estimates, the R library `e1071` to calculate the SVM classifier, and the R library `sdwd` (Wang and Zou, 2016) to calculate the DWD classifier.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

5.1 Simulations

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

Scenario 1 We set the precision matrix for Class 1 to $\Sigma_1^{-1} = (1 + \sqrt{p})I_p$. 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 with values equal to ζ, and all 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 nonzero entries, the magnitudes of which are 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 five-by-five submatrix of Σ_1 has its off-diagonal elements set to ρ. The rest of the setting is the same as that in Scenario 1. In this scenario, both the difference in the covariance and the difference in the precision matrix are confined in the leading five-by-five submatrix, so that the majority of the 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Scenario 3 We set the precision matrix Σ_1 for Class 1 to be a Toeplitz matrix with the first row $(1-\rho^2)^{-1}(\rho^0, \rho^1, \rho^2, \ldots, \rho^{p-1})$. 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 one, that is, a matrix with nonzero entries that are confined to the main diagonal and one more diagonal on both sides. Because 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 that 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 the mean vectors (u), and 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 because the true value of β is a zero vector. This 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,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

and 3, respectively.

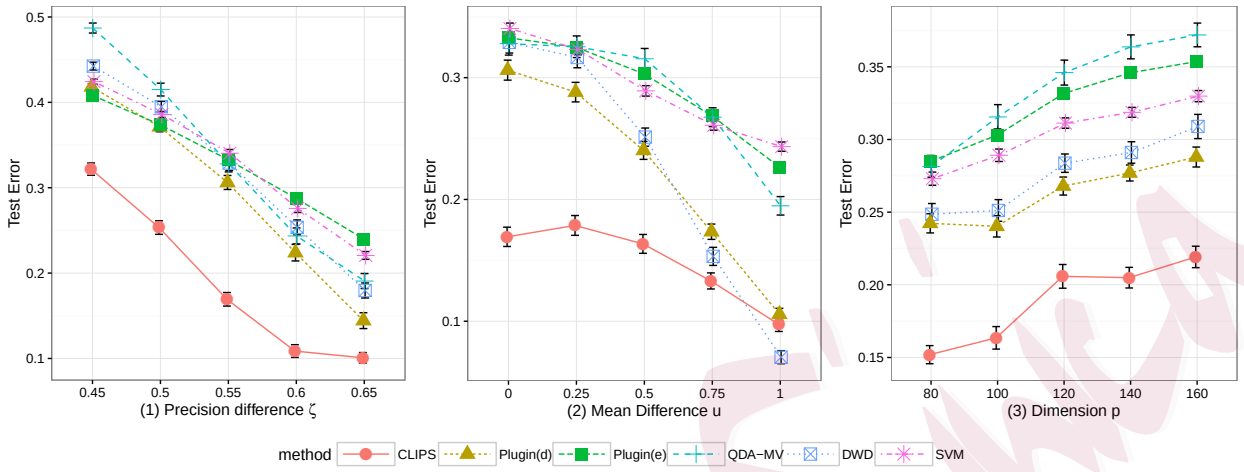

Figure 2: Set classification for Scenario 1. The three panels correspond to varying ζ, varying u, and varying p, respectively. The CLIPS classifier performs very well when the effect of the covariance dominates that of the mean difference.

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, works reasonably well in this setting in which the major mode of variation is in the off-diagonal of the precision matrices. However, because 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 improves, the DWD method gains the most (it is even the best performing classifier when the mean difference u is as large as one). This may be because the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

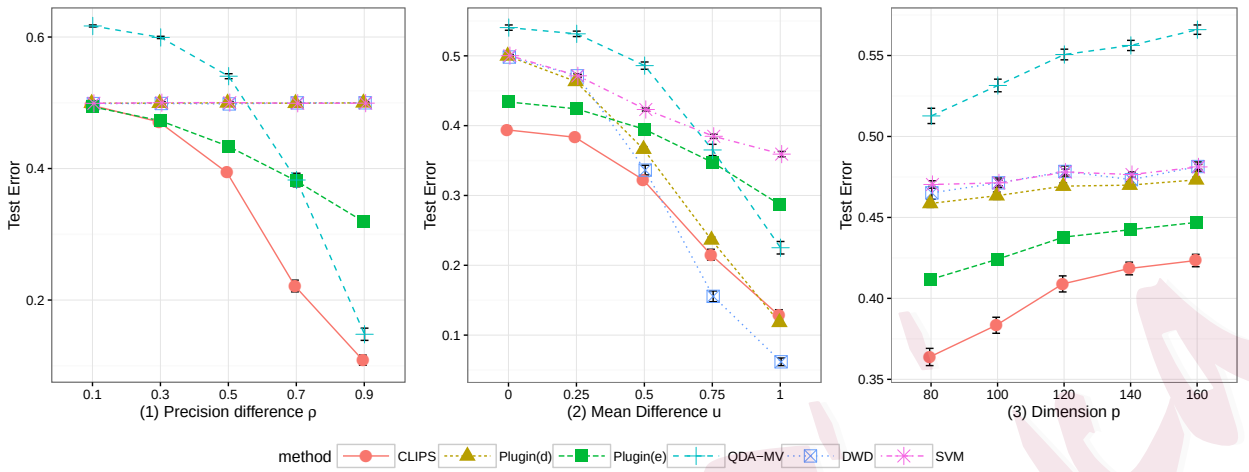

Figure 3: Set classification for Scenario 2. The three panels correspond to varying ρ, varying u, and varying p, respectively. The classifiers that do not engage the covariance perform poorly when there is no mean difference signal.

mean difference on which DWD relies, instead of the difference in the precision matrix, is sufficiently large to secure good performance in separating sets between two classes.

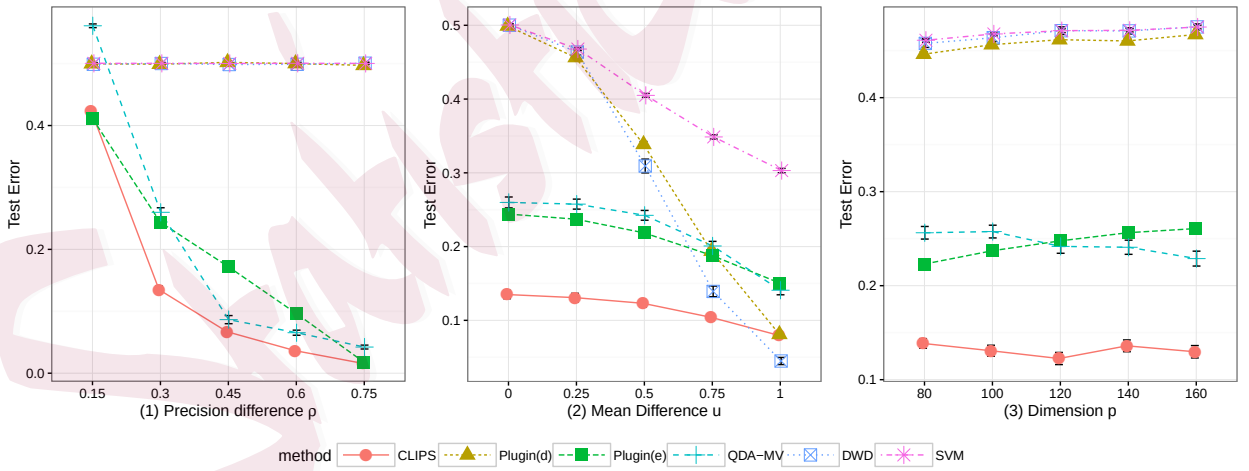

Figure 4: Set classification for Scenario 3. The three panels correspond to varying ρ, varying u, and varying p, respectively. As in Scenario 2, the classifiers that do not engage the covariance perform poorly when there is no mean difference signal.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

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

The results for Scenario 3 (Figure 4) are similar to those of Scenario 2, except that, this time, the advantage of the two covariance-engaged set classification methods, CLIPS and Plugin(e), seems to be more obvious when the mean difference is zero (see left panel). Moreover, the QDA-MV method 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 also depends on p (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 a biopsy, in which the 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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

to obtain segmented images of nuclei. The data we analyzed contain five sets of nuclei from normal liver tissues and five sets of nuclei from cancerous tissues. Each set contains 50 images. The data set is publicly available (https://faculty.virginia.edu/rohde/segmented-nuclei.zip) 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 a principal component analysis. Then, among the 50 images of each set, 16 images are retained as a training set, 16 are a tuning set, and another 16 are the 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)) and the QDA-MV method perform better than those methods that do not take the covariance matrices

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

into account, such as the DWD and SVM (note that they do consider the diagonal of the covariance matrix.)

To gain some insight into why the covariance-engaged set classifiers work and traditional

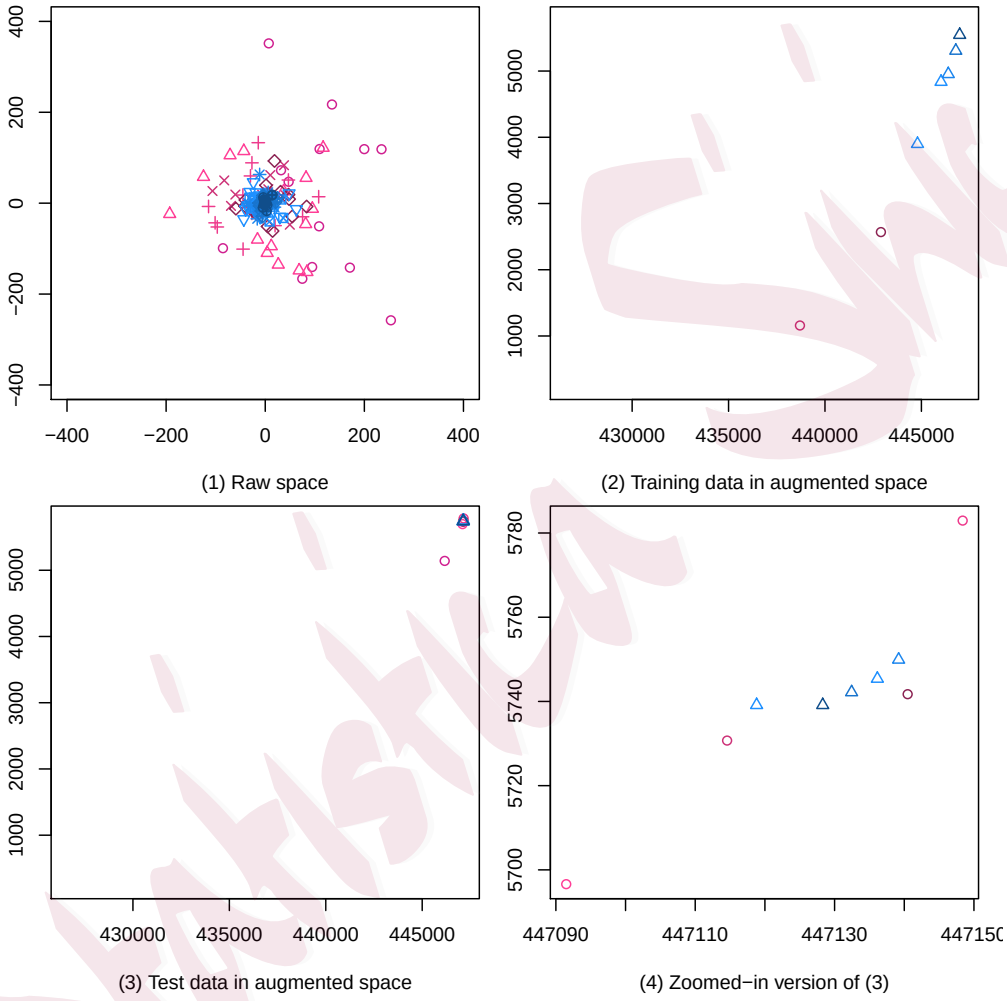

Figure 5: PCA scatter plots for the liver cell nucleus image data. Both classes are shown in different colors (blue and purple, or lighter and darker gray). (1): the elementary observations in the raw space; different sets are shown in different symbols. (2) and (3): the augmented space seen by the DWD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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

methods fail, we visualize the data set in Figure 5. Subfigure (1) shows a scatter plot of the first two principal components of all the elementary observations (ignoring the set memberships) in the data sets, in which blue (light gray) and violet (dark gray) depict the two different classes. Observations in the same set are shown using 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 the second moment, such as the variance, distinguishes the two classes.

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

Supplementary Material

The online Supplementary Material contains additional theoretical arguments, proofs of all results, and an additional data analysis.

Acknowledgments

This work was supported in part by the *National Science Foundation* (DMS-1812030),

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

National Research Foundation of Korea (No. 2019R1A2C2002256), and a collaboration grant from *Simons Foundation* (246649).

References

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

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

Arandjelovic, O. and Cipolla, R. (2006). Face set classification using maximally probable mutual modes. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 511–514. IEEE.

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

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

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

Cai, T. T. and Zhang, L. (2019). A convex optimization approach to high-dimensional sparse quadratic discriminant analysis. *arXiv preprint arXiv:1912.02872*.

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

Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

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

Cheplygina, V., Tax, D. M., and Loog, M. (2015). On classification with bags, groups and sets. *Pattern Recognition Letters*, 59:11–17.

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.

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

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

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

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

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

Gaynanova, I. and Wang, T. (2019). Sparse quadratic classification rules via linear dimension reduction. *Journal of multivariate analysis*, 169:278–299.

Jiang, B., Wang, X., and Leng, C. (2018). A direct approach for sparse quadratic discriminant analysis. *The Journal of Machine Learning Research*, 19(1):1098–1134.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Kuncheva, L. I. (2010). Full-class set classification using the hungarian algorithm. *International Journal of Machine Learning and Cybernetics*, 1(1-4):53–61.

Li, Q. and Shao, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, 25:457–473.

Mai, Q., Yang, Y., and Zou, H. (2019). Multiclass sparse discriminant analysis. *Statistica Sinica*, 29:97–111.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42.

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

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

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

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

Miedema, J., Marron, J. S., Niethammer, M., Borland, D., Woosley, J., Coposky, J., Wei, S., Reisner, H., and Thomas, N. E. (2012). Image and statistical analysis of melanocytic histology. *Histopathology*, 61(3):436–444.

Ning, X. and Karypis, G. (2009). The set classification problem and solution methods. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 847–858. SIAM.

Pan, Y. and Mai, Q. (2020). Efficient computation for differential network analysis with applications to quadratic discriminant analysis. *Computational Statistics & Data Analysis*, 144:106884.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Qin, Y. (2018). A review of quadratic discriminant analysis for high-dimensional data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(4):e1434.

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

Samsudin, N. A. and Bradley, A. P. (2010). Nearest neighbour group-based classification. *Pattern Recognition*, 43(10):3458–3467.

Shifat-E-Rabbi, M., Yin, X., Fitzgerald, C. E., and Rohde, G. K. (2020). Cell image classification: a comparative overview. *Cytometry Part A*, 97(4):347–362.

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

Wang, B. and Zou, H. (2018). Another look at distance-weighted discrimination. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):177–198.

Wang, R., Guo, H., Davis, L. S., and Dai, Q. (2012). Covariance discriminative learning: A natural and efficient approach to image set classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2496–2503. IEEE.

Wang, W., Ozolek, J. A., and Rohde, G. K. (2010). Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images. *Cytometry Part A*, 77(5):485–494.

Wang, W., Ozolek, J. A., Slepčev, D., Lee, A. B., Chen, C., and Rohde, G. K. (2011). An optimal transportation approach for nuclear structure-based pathology. *IEEE Transactions on Medical Imaging*, 30(3):621–631.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0253

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

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

Zhao, S. D., Cai, T. T., and Li, H. (2014). Direct estimation of differential networks. *Biometrika*, 101(2):253–268.

Zou, H. (2019). Classification with high dimensional features. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(1):e1453.

Zhao Ren

Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA

E-mail:zren@pitt.edu

Sungkyu Jung

Department of Statistics, Seoul National University, Gwanak-gu, Seoul 08826, Korea

E-mail: sungkyu@snu.ac.kr

Xingye Qiao

Department of Mathematical Sciences, Binghamton University, State University of New York, Binghamton, NY, 13902 USA

E-mail: qiao@math.binghamton.edu