Journal of Machine Learning Research 16 (2015) 1547-1572 Submitted 1/14; Revised 9/14; Published 8/15

Flexible High-Dimensional Classification Machines and Their Asymptotic Properties

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

Lingsong Zhang LINGSONG@PURDUE.EDU
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Editor: Massimiliano Pontil

Abstract

Classification is an important topic in statistics and machine learning with great potential in many real applications. In this paper, we investigate two popular large-margin classification methods, Support Vector Machine (SVM) and Distance Weighted Discrimination (DWD), under two contexts: the high-dimensional, low-sample size data and the imbalanced data. A unified family of classification machines, the FLexible Assortment MachinE (FLAME) is proposed, within which DWD and SVM are special cases. The FLAME family helps to identify the similarities and differences between SVM and DWD. It is well known that many classifiers overfit the data in the high-dimensional setting; and others are sensitive to the imbalanced data, that is, the class with a larger sample size overly influences the classifier and pushes the decision boundary towards the minority class. SVM is resistant to the imbalanced data issue, but it overfits high-dimensional data sets by showing the undesired data-piling phenomenon. The DWD method was proposed to improve SVM in the high-dimensional setting, but its decision boundary is sensitive to the imbalanced ratio of sample sizes. Our FLAME family helps to understand an intrinsic connection between SVM and DWD, and provides a trade-off between sensitivity to the imbalanced data and overfitting the high-dimensional data. Several asymptotic properties of the FLAME classifiers are studied. Simulations and real data applications are investigated to illustrate theoretical findings.

Keywords: classification, Fisher consistency, high-dimensional low-sample size asymptotics, imbalanced data, support vector machine

1. Introduction

Classification refers to predicting the class label, $y \in \mathcal{C}$, of a data object based on its covariates, $\boldsymbol{x} \in \mathcal{X}$. Here $\mathcal{C}$ is the space of class labels, and $\mathcal{X}$ is the space of the covariates. Usually we consider $\mathcal{X} \equiv \mathbb{R}^d$, where d is the number of variables or the dimension. See Duda et al. (2001) and Hastie et al. (2009) for a comprehensive introduction to many popular classification methods. When $\mathcal{C} = \{+1, -1\}$, this is an important class of classification

problems, called binary classification. The classification rule for a binary classifier usually has the form $\phi(\boldsymbol{x}) = \text{sign}\{f(\boldsymbol{x})\}$, where $f(\boldsymbol{x})$ is called the discriminant function. Linear classifiers are the most important and the most commonly used classifiers, as they are often easy to interpret in addition to reasonable classification performance. We focus on linear classifier in this article. In the above formula, linear classifiers correspond to $f(\boldsymbol{x}; \boldsymbol{\omega}, \beta) = \boldsymbol{x}^T\boldsymbol{\omega}+\beta$. The sample space is divided into halves by the *separating hyperplane*, also known as the *classification boundary*, defined by $\{\boldsymbol{x} : f(\boldsymbol{x}) \equiv \boldsymbol{x}^T\boldsymbol{\omega} + \beta = 0\}$. Note that the coefficient vector $\boldsymbol{\omega} \in \mathbb{R}^d$ defines the normal vector, and hence the orientation, of the classification boundary; and the intercept term $\beta \in \mathbb{R}$ defines the location of the classification boundary.

In this paper, two popular classification methods, Support Vector Machine (SVM; Cortes and Vapnik, 1995; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000) and Distance Weighted Discrimination (DWD; Marron et al., 2007) are investigated under two important contexts: the High-Dimensional, Low-Sample Size (HDLSS) data and the imbalanced data. Both methods are large-margin classifiers (Smola et al., 2000), that seek separating hyperplanes which maximize certain notions of *gap* (that is, distances) between the two classes. The investigation of the performance of SVM and DWD motivates the notion of a unified family of classifiers, the FLexible Assortment MachinE (FLAME), which connects the two classifiers, and helps to understand their connections and differences.

There is a large literature in statistics and machine learning on large-margin classifiers. For example, Wahba (1999) studied kernel SVM in Reproducing Kernel Hilbert Spaces. Lin (2004) introduced and proved Fisher consistency for SVM. Bartlett et al. (2006) quantified the excess risk of a loss function in a learning problem including the case of large-margin classification. On the methodology level, Shen et al. (2003) invented ψ-learning; Wu and Liu (2007) introduced robust SVM. Recently, Liu, Zhang, and Wu (2011) studies a unified class of classifiers which connected hard classification and soft classification (probability estimation).

It is worth mentioning that the FLAME family is not proposed as a better classification method to replace SVM or DWD. Instead, it is proposed as a unified machine, which is very helpful to investigate the trade-off between generalization errors and overfitting. A single parameter will be used to control the trade-off, of which DWD and SVM sit on the two ends.

1.1 Motivation: Pros and Cons of SVM and DWD

SVM is a very popular classifier in statistics and machine learning. It has been shown to have Fisher consistency, that is, when sample size goes to infinity, its decision rule converges to the Bayes rule (Lin, 2004). SVM has several nice properties: 1) Its dual formulation is relatively easy to implement (through Quadratic Programming). 2) SVM is robust to the model specification, which makes it very popular in various real applications. However, when being applied to HDLSS data, it has been observed that a large portion of the data (usually the support vectors, to be properly defined later) lie on two hyperplanes parallel to the SVM classification boundary. This is known as the *data-piling* phenomenon (Marron et al., 2007; Ahn and Marron, 2010). Data-piling of SVM indicates a type of overfitting. Other overfitting phenomenon of SVM under the HDLSS context include:

1. The angle between the SVM direction and the Bayes rule direction is usually large.

2. The variability of the sampling distribution of the SVM direction $\boldsymbol{\omega}$ is very large (Zhang and Lin, 2013). Moreover, because the separating hyperplane is decided only by the support vectors, the SVM direction tends to be unstable, in the sense that small turbulence or measurement error to the support vectors can lead to a big change of the estimated direction.
3. In some cases, the out-of-sample classification performance may not be optimal due to the suboptimal direction of the estimated SVM discrimination direction.

DWD is a recently developed classifier to improve SVM in the HDLSS setting. It uses a different notion of gap from SVM. While SVM is to maximize the smallest distance between classes, DWD is to maximize a special average distance (harmonic mean) between classes. It has been shown in many earlier simulations that DWD largely overcomes the overfitting (data-piling) issue and it usually gives a better discrimination direction.

On the other hand, the intercept term β of the DWD method is sensitive to the sample size ratio between the two classes, that is, to the imbalanced data (Qiao et al., 2010). Note that, even though a good discriminant direction $\boldsymbol{\omega}$ is more important in revealing the structure of the data, the classification/prediction performance heavily depends on the intercept β, more than on the direction $\boldsymbol{\omega}$. As shown in Qiao et al. (2010), usually the β term of the SVM classifier is not sensitive to the sample size ratio, while the β term of the DWD method will become too large (or too small) if the sample size of the positive class (or negative class) is very large.

In summary, both methods have pros and cons. SVM has a greater stochastic variability and usually overfits the data by showing data-piling phenomena, but is less sensitive to the imbalanced data issue. DWD usually overcomes the overfitting/data-piling issue, and has a smaller sampling variability, but is very sensitive to the imbalanced data. Driven by their similarity, we propose a unified class of classifiers, FLAME, in which the above two classifiers are special cases. FLAME provides a framework to study the connections and differences between SVM and DWD. Each FLAME classifier has a parameter θ which is used to control the performance balance between overfitting the HDLSS data and the sensitivity to the imbalanced data. It turns out that the DWD method is FLAME with $\theta = 0$; and that the SVM method corresponds to FLAME with $\theta = 1$. The optimal θ depends on the trade-off among several factors: stochastic variability, overfitting and resistance against the imbalanced data. In this paper, we also propose an approach to select θ, where the resulting FLAME have the potential to achieve a balanced performance between the SVM and DWD methods.

1.2 Outline

The rest of the paper is organized as follows. Section 2 provides toy examples and highlights the strengths and drawbacks of SVM and DWD on classifying the HDLSS and imbalanced data. We develop the FLAME method in Section 3, which is motivated by the investigation of the loss functions of SVM and DWD. Section 4 provides suggestions on choosing the parameters. Three types of asymptotic results for the FLAME classifier are studied in Section 5. Section 6 demonstrates its properties using simulation experiments. A real application study is conducted in Section 7. Some concluding remarks and discussions are

made in Section 8. Technical proofs of theorems and propositions are included in Online Appendix 1.

2. Comparison of SVM and DWD

In this section, we use several toy examples to illustrate the strengths and drawbacks of SVM and DWD under two contexts: HDLSS data and imbalanced data.

2.1 Overfitting HDLSS Data

We use several simulated examples to compare SVM and DWD. The results show that the stochastic variability of the SVM direction is usually larger than that of the DWD method, and SVM directions are deviated farther away from Bayes rule directions. In addition, the new proposed FLAME machine (see details in Section 3) is also included in the comparison, and it turns out that FLAME with a mediocre θ is between the above two methods.

Figure 1 shows the comparison results between SVM, DWD and FLAME (with tuning parameter $\theta = 1/2$). We simulate 10 samples with the same underlying distribution. Each simulated data set contains 12 variables and two classes, with 120 observations in each class. The two classes have mean difference on only the first three dimensions and the within-class covariances are diagonal, that is, the variables are independent. For each simulated data set, we plot the first three components of the resulting discriminant directions from SVM, DWD and FLAME (after normalizing the 3D vectors to have unit L_2 norms), as shown in Figure 1. It clearly shows that the DWD directions (the blue down-pointing triangles) are the closest ones to the *true* Bayes rule direction (the cyan diamond marker) among the three approaches. In addition, the DWD directions have the smallest variation (that is, more stable) over different samples. The SVM directions (the red up-pointing triangles) are farthest from the *true* Bayes rule direction and have a larger variation than the other two methods. To highlight the direction variabilities of the three methods, we introduce a novel measure for the variation (instability) of the discriminant directions: the trace of the sample covariance of the resulting direction vectors over the 10 replications, which we name as *dispersion*. The dispersion for the DWD method (0.0031) is much smaller than that of the SVM method (0.0453), as highlighted in the figure as well. The new FLAME classifiers usually have a performance between DWD and SVM. Figure 1 shows the results of a specific FLAME ($\theta = 0.5$, the magenta squares), which are better than SVM but worse than DWD.

Besides the advantage in terms of the stochastic variability and the deviation from the true direction, DWD outperforms SVM in terms of stability in the presence of small perturbations. In Figure 2, we use a two-dimensional example to illustrate this phenomenon. We simulate a perfectly separable 2-dimensional data set. The theoretical Bayes rule decision boundary is shown as the thick black line. The dashed red line and the dashed dotted blue line are the SVM and the DWD classification boundaries respectively before the perturbation. We then move one observation in the positive group slightly (from the solid triangle to the solid diamond as shown in the figure). This perturbation leads to a visible change of the SVM direction (shown as the dotted red line), but a smaller change for DWD (shown as the solid blue line). Note that all four hyperplanes are capable of classifying this training

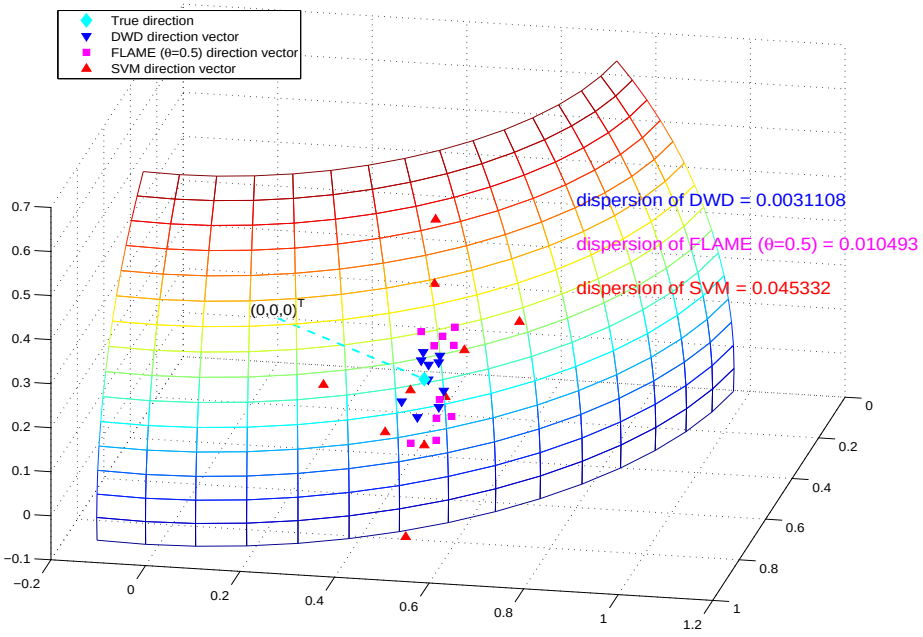

Figure 1: The true population mean difference direction vector (the cyan dashed line and diamond marker; equivalent to the Bayes rule direction), the DWD directions (blue down-pointing triangles), the FLAME directions with $\theta = 0.5$ (magenta squares), and the SVM directions (red up-pointing triangles) for 10 realizations of simulated data. Each direction vector has norm 1 and thus is depicted as a point on the 3D unit sphere. On average, all machines have their discriminant direction vectors scattering around the true direction. The DWD directions are the closest to the true direction and have the smallest variation. The SVM directions have the largest variation and are farthest from the true direction. The variation of the intermediate FLAME direction vectors is between the two machines above. The variation (dispersion) of a machine is also measured by the trace of the sample covariance calculated from the 10 resulting direction vectors for the 10 simulations.

data set perfectly. But it may not be true for an out-of-sample test set. This example shows the unstableness of SVM.

2.2 Sensitivity to Imbalanced Data

In the last subsection, we have shown that DWD outperforms SVM in estimating the discrimination direction, that is, DWD directions are closer to the Bayes rule discrimination directions and usually have a smaller variability. However, it was found that the location of DWD classification boundary, which is characterized by the intercept β, is sensitive to the sample size ratio between the two classes (Qiao et al., 2010).

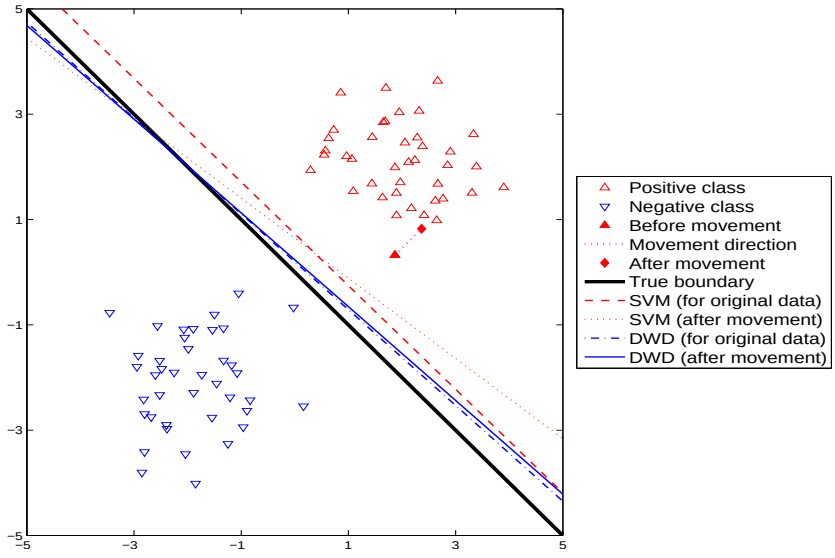

Figure 2: A 2D example shows that the unstable SVM boundary has changed due to a small turbulence of a support vector (the solid red triangle and diamond) while the DWD boundary remains almost still.

Usually, a good discriminant direction $\boldsymbol{\omega}$ helps to reveal the profiling difference between two classes of populations. But the classification/prediction performance heavily depends on the location coefficient β. We define the *imbalance factor* $m \geq 1$ as the sample size ratio between the majority class and the minority class. It turns out that β in the SVM classifier is not sensitive to m. However, the β term for the DWD method is very sensitive to m. We also notice that, as a consequence, the DWD separating hyperplane will be pushed toward the minority class, when the ratio m is close to infinity, that is, DWD classifiers intend to ignore the minority class. In this section, we use another toy example to better illustrate the impact of the imbalanced data on both the estimated β and the classification performance.

Figure 3 uses a one-dimensional example, so that estimating $\boldsymbol{\omega}$ is not needed. This also corresponds to a multivariate data set, where $\boldsymbol{\omega}$ is estimated correctly first, after which the data set is projected to $\boldsymbol{\omega}$ to form the one-dimensional data. In this plot, the x-coordinates of the red dots and the blue dots are the values of the data while the y-coordinates are random jitters for better visualization. The red and blue curves are the kernel density estimations for both classes. In the top subplot of Figure 3, where $m = 1$ (that is, the balanced data), both the DWD (blue lines) and SVM (red lines) boundaries are close to the Bayes rule boundary (black solid line), which sits at 0. In the bottom subplot, the sample size of the red class is tripled, which corresponds to $m = 3$. Note that the SVM boundary moves a little towards the minority (blue) class, but still fairly close to the true boundary. The DWD boundary, however, is pushed towards the minority. Although this does not impose immediate problems for the training data set, the DWD classifier will suffer from a great loss of classification performance when it is applied to an out-of-sample data set. It can be shown that when m goes to infinity, the DWD classification boundary will tends to

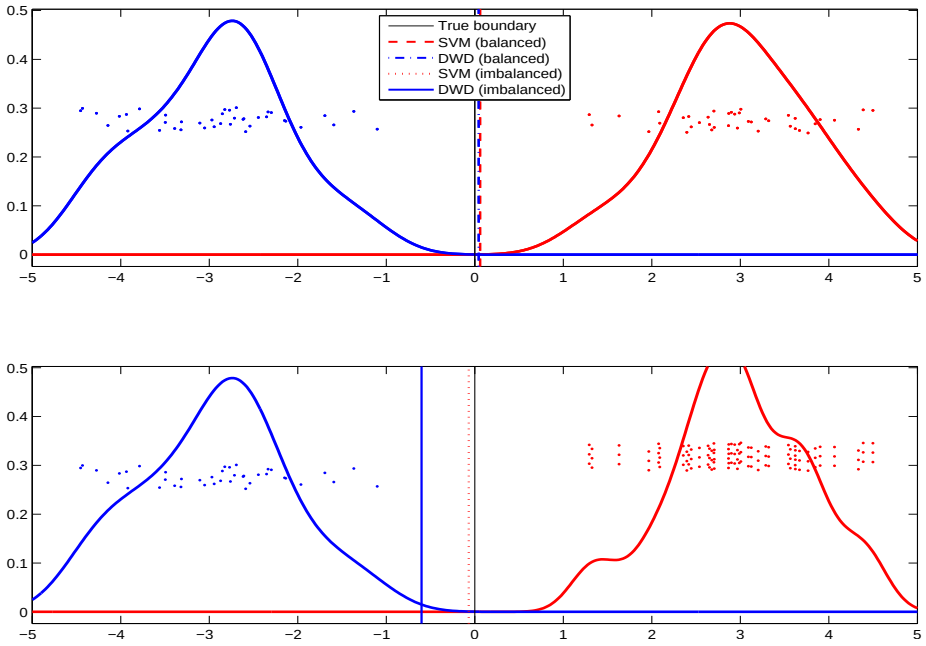

Figure 3: A 1D example shows that the DWD boundary is pushed towards the minority class (blue) when the majority class (red) has tripled its sample size.

negative infinity, which totally ignores the minority group (see our Theorem 4). However, SVM will not suffer from the imbalanced data issue. One reason is that SVM only needs a small fraction of data (called support vectors) to estimate both $\boldsymbol{\omega}$ and β, which mitigate the imbalanced data issue naturally.

Imbalanced data issues have been investigated in both statistics and machine learning. See an extensive survey in Chawla et al. (2004). Recently, Owen (2007) studied the asymptotic behavior of infinitely imbalanced binary logistic regression. In addition, Qiao and Liu (2009) and Qiao et al. (2010) proposed to use adaptive weighting approaches to overcome the imbalanced data issue.

In summary, the performance of DWD and SVM is different in the following ways: 1) The SVM direction usually has a larger variation and deviates farther from the Bayes rule direction than the DWD direction, which are indicators of overfitting HDLSS data. 2) The SVM intercept is not sensitive to the imbalanced data, but the DWD intercept is. These observations have motivated us to investigate their similarity and differences. In the next section, a new family of classifier will be proposed, which unifies the above two classifiers.

3. FLAME Family

In this section, we introduce FLAME, a family of classifiers, through a thorough investigation of the loss functions of SVM and DWD in Section 3.1. The formulation and implementation of the FLAME classifiers are given in Section 3.2.

3.1 SVM and DWD Loss Functions

The key factors that drive the very distinct performances of the SVM and the DWD methods are their associated loss functions (see Figure 4.)

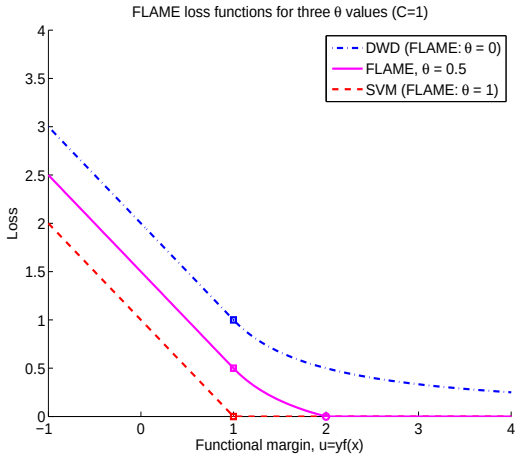

Figure 4: FLAME loss functions for three θ values: $\theta = 0$ (equivalent to SVM/Hinge loss), $\theta = 0.5$, $\theta = 1$ (equivalent to DWD). The parameter C is set to be 1.

Figure 4 displays the loss functions of SVM, DWD and FLAME with some specific tuning parameters. SVM uses the Hinge loss function, $H(u) = (1-u)_+$ (the red dashed curve in Figure 4), where u corresponds to the functional margin $u \equiv yf(\boldsymbol{x})$. Note that the functional margin u can be viewed as the distance of vector $\boldsymbol{x}$ from the separating hyperplane (defined by $\{\boldsymbol{x} : f(\boldsymbol{x}) = 0\}$). When $u > 0$ and is large, the data vector is correctly classified and is far from the separating hyperplane; when $u < 0$, the data vector is wrongly classified. Note that when $u > 1$, the corresponding Hinge loss equals zero. Thus, only those observations with $u \leq 1$ contribute to the estimation of $\boldsymbol{\omega}$ and β. These observations are called *support vectors*. Hence, SVM is insensitive to the observations that are far away from the decision boundary, which is the reason that it is less sensitive to the imbalanced data issue. However, the influence by only the support vectors makes the SVM solution subject to overfitting (data-piling). This can be explained by the following: in the optimization process of SVM, the functional margins for the vectors are pushed towards a region with small loss, that is, functional margins u are encouraged to be large. But once a vector is pushed to the point where $u = 1$, the optimization mechanism lacks further incentive to continue pushing it towards a larger function margin as the Hinge loss cannot be further reduced for this vector. Therefore many data vectors are piling along the hyperplane corresponding to $u = 1$. Data-piling is bad for generalization because a small turbulence to the support vectors could lead to a big difference of the estimated discriminant direction vector (recall the examples in Section 2.1).

The DWD method corresponds to a different DWD loss function,

$$V(u) = \begin{cases} 2\sqrt{C} - Cu & \text{if } u \le \frac{1}{\sqrt{C}}, \\ 1/u & \text{otherwise.} \end{cases} \tag{1}$$

Here C is a pre-defined constant. Figure 4 shows the DWD loss function with $C = 1$. It is clear that the DWD loss function is very similar to the SVM loss function when u is small (both are linearly decreasing with respect to u). The major difference is that the DWD loss is always positive. This property will make the DWD method behave in a very different way than SVM. As there is always an incentive to make the function margin to be larger (and the loss to be smaller), the DWD loss function kills data-piling, and mitigates the overfitting issue for HDLSS data.

On the other hand, the DWD loss function makes the DWD method very sensitive to the imbalanced data issue. This is because now that each observation will have some influence, the larger class will have a larger influence. The decision boundary of the DWD method tends to ignore the smaller class, because sacrificing the smaller class (boundary being closer to the smaller class and farther from the larger class) can lead to a dramatic reduction of the loss from the larger class, which ultimately lead to a minimized overall loss.

3.2 FLAME

We propose to borrow strengths from both methods to simultaneously deal with both the imbalanced data and the overfitting (data-piling) issues. We first highlight the connections between the DWD loss and an modified version of the Hinge loss (of SVM). Then we modify the DWD loss so that samples far from the classification boundary will have zero loss.

Let $f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\omega} + \beta$. The formulation of SVM can be rewritten (see details in Appendix A) in the form of $\underset{\boldsymbol{\omega},\beta}{\operatorname{argmin}} \sum_i H^*(y_i f(\boldsymbol{x}_i))$, s.t. $\|\boldsymbol{\omega}\|^2 \le 1$ where the modified Hinge loss function H^* is defined as

$$H^*(u) = \begin{cases} \sqrt{C} - Cu & \text{if } u \le \frac{1}{\sqrt{C}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Comparing the DWD loss (1) and this modified Hinge loss (2), one can easily see their connections: for $u \le \frac{1}{\sqrt{C}}$, the DWD loss is greater than the Hinge loss of SVM by an exact constant $\sqrt{C}$, and for $u > \frac{1}{\sqrt{C}}$, the DWD loss is $1/u$ while the SVM Hinge loss equals 0. Clearly the modified Hinge loss (2) is the result of soft-thresholding the DWD loss at $\sqrt{C}$. In other words, SVM can be seen as a special case of DWD where the losses of those vectors with $u = y_i f(\boldsymbol{x}_i) > 1/\sqrt{C}$ are shrunken to zero. To allow different levels of soft-thresholding, we propose to use a new loss function which (soft-)thresholds the DWD loss function by constant $\theta\sqrt{C}$ where $0 \le \theta \le 1$, that is, a fraction of $\sqrt{C}$. The new loss function is

$$L(u) = \left[V(u) - \theta\sqrt{C}\right]_+ = \begin{cases} (2-\theta)\sqrt{C} - Cu & \text{if } u \le \frac{1}{\sqrt{C}}, \\ 1/u - \theta\sqrt{C} & \text{if } \frac{1}{\sqrt{C}} \le u < \frac{1}{\theta\sqrt{C}}, \\ 0 & \text{if } u \ge \frac{1}{\theta\sqrt{C}}, \end{cases} \tag{3}$$

that is, to reduce the DWD loss by a constant, and truncate it at 0. The magenta solid curve in Figure 4 is the FLAME loss when $C = 1$ and $\theta = 0.5$. This simple but useful modification unifies the DWD and SVM methods. When $\theta = 1$, the new loss function (when $C = 1$) reduces to the SVM Hinge loss function; while when $\theta = 0$, it remains as the DWD loss.

Note that $L(u) = 0$ for $u > 1/(\theta\sqrt{C})$. Thus, those data vectors with very large functional margins will still have zero loss. For DWD loss (corresponding to $\theta = 0$), note that $1/(\theta\sqrt{C}) = \infty$. Thus no data vector can have zero loss. For SVM loss, all the data vector with $u > 1/(\theta\sqrt{C}) = 1/\sqrt{C}$ will have zero loss. Training a FLAME classifier with $0 < \theta < 1$ can be interpreted as sampling a portion of data which are farther from the boundary than $1/\theta\sqrt{C}$ and assign zero loss to them. Alternatively, it can be viewed as sampling data that are closer to the boundary than $1/\theta\sqrt{C}$ and assign positive loss to them. Note that the larger θ is, the fewer data are sampled to have positive loss. As one can flexibly choose θ, the new classification method with this new loss function is called the FLexible Assortment MachinE (FLAME).

FLAME can be implemented by a Second-Order Cone Programming algorithm (Toh et al., 1999; Tütüncü et al., 2003). Let $\theta \in [0, 1]$ be the FLAME parameter. The proposed method minimizes $\min_{\boldsymbol{\omega},b,\boldsymbol{\xi}} \sum_{i=1}^{n} \left(\frac{1}{r_i} + C\xi_i - \theta\sqrt{C} \right)_+$. A slack variable $\varphi_i \geq 0$ can be introduced to absorb the $(\cdot)_+$ function. The optimization of the FLAME can be written as

$$\begin{aligned} &\min_{\boldsymbol{\omega},b,\boldsymbol{\xi}} \sum_i \varphi_i, \\ \text{s.t.} \quad &\left(\frac{1}{r_i} + C\xi_i - \theta\sqrt{C} \right) - \varphi_i \leq 0, \ \varphi_i \geq 0, \\ &r_i = y_i(\boldsymbol{x}_i^T \boldsymbol{\omega} + \beta) + \xi_i, \ r_i \geq 0 \text{ and } \xi_i \geq 0, \\ &\|\boldsymbol{\omega}\|^2 \leq 1. \end{aligned}$$

A MATLAB routine has been implemented and is available at the authors' personal websites. See Online Appendix 1 for more details on the implementation.

4. Choice of Parameters

There are two tuning parameters in the FLAME model: one is the C, inherited from the DWD loss, which controls the amount of allowance for misclassification; the other is the FLAME parameter θ, which controls the level of soft-thresholding. Similar to the discussion in DWD (Marron et al., 2007), the classification performance of FLAME is insensitive to different values of C. In addition, it can be shown for any C, FLAME is Fisher consistent, by applying the general results in Lin (2004). Thus, the default value for C as proposed in Marron et al. (2007) will be used in FLAME. As the property and the performance of FLAME depends on the choice of the θ parameter, it is important to select the right amount of thresholding. In this section, we introduce a way of choosing the second parameter θ, which is motivated by a theoretical consideration and is heuristically meaningful as well.

Having observed that the DWD discrimination direction is usually closer to the Bayes rule direction, but its location term β is sensitive to the imbalanced data issue, we propose

the following data-driven approach to select an appropriate θ. Without loss of generality, we assume that the negative class is the majority class with sample size n_- and the positive class is the minority class with sample size n_+. We point out that the main reason that DWD is sensitive to the imbalanced data issue is that it uses all vectors in the *majority* class to build up a classifier. A heuristic strategy to correct this would be to force the optimization to use the same number of vectors from both classes (so as to mimic a balanced data set) to build up a classifier: we first apply DWD to the data set, and calculate the distances of all data in the majority (negative) class to the current DWD classification boundary; we then train FLAME with a carefully chosen parameter θ which assigns positive loss to the closet $n_+(< n_-)$ data vectors in the majority (negative) class to the classification boundary. As a consequence, each class will have exactly n_+ vectors which have positive loss. In other words, while keeping the least imbalance (because we have the same numbers of vectors from both classes that have influence over the optimization), we obtain a model with the least possible overfitting (because $2n_+$ vectors have influence, instead of only the limited support vectors as in SVM.)

In practice, since the new FLAME classification boundary using the θ value chosen above may be different from the initial DWD classification boundary, the n_+ closest points to the FLAME classification boundary may not be the same n_+ closest points to the DWD boundary. This means that it is not guaranteed that exactly n_+ points from the majority class will have positive loss. However, one can expect that reasonable approximation can be achieved. Moreover, an iterative scheme for finding θ is introduced as follows in order to minimize such discrepancy.

For simplicity, we let $(\boldsymbol{x}_i, y_i)$ with index i be an observation from the positive/minority class and $(\boldsymbol{x}_j, y_j)$ with index j be an observation from the negative/majority class.

Algorithm 1 ***(Adaptive parameter)***

1. *Initiate* $\theta_0 = 0$.
2. *For* $k = 0, 1, \cdots,$
   (a) *Solve FLAME solutions* $\boldsymbol{\omega}(\theta_k)$ *and* $\beta(\theta_k)$ *given parameter* θ_k.
   (b) *Let* $\theta_{k+1} = \max\left(\theta_k,\ \left\{g_{(n_+)}(\theta_k)\sqrt{C}\right\}^{-1}\right)$, *where* $g_j(\theta_k)$ *is the functional margin* $u_j \equiv y_j(\boldsymbol{x}_j^T\boldsymbol{\omega}(\theta_k) + \beta(\theta_k))$ *of the* j*th vector in the negative/majority class and* $g_{(l)}(\theta_k)$ *is the* l*th order statistic of these functional margins.*
3. *When* $\theta_k = \theta_{k-1}$, *the iteration stops.*

The goal of this algorithm is to make $g_{(n_+)}(\theta_k)$ to be the greatest functional margin among all the data vectors that have positive loss in the negative/majority class. To achieve this, we calibrate θ by aligning $g_{(n_+)}(\theta_k)$ to the turning point $u = 1/(\theta\sqrt{C})$ in the definition of the FLAME loss (3), that is $g_{(n_+)}(\theta_k) = 1/(\theta\sqrt{C}) \Rightarrow \theta = \left(g_{(n_+)}(\theta_k)\sqrt{C}\right)^{-1}$.

We define the equivalent sample objective function of FLAME for the iterative algorithm above, $s(\boldsymbol{\omega}, \beta, \theta) = \dfrac{1}{n_+ + n_-}\left[\sum_{i=1}^{n_+} L((\boldsymbol{x}_i^T\boldsymbol{\omega} + \beta), \theta) + \sum_{j=1}^{n_-} L(-(\boldsymbol{x}_j'\boldsymbol{\omega} + \beta), \theta)\right] + \dfrac{\lambda}{2}\|\boldsymbol{\omega}\|^2$. Then the convergence of this algorithm is shown in Theorem 1. The proofs of all the theorems and propositions in this article are included in Online Appendix 1.

Theorem 1 *In Algorithm 1, $s(\boldsymbol{\omega}_k, \beta_k, \theta_k)$ is non-increasing in k. As a consequence, Algorithm 1 converges to a stationary point $s(\boldsymbol{\omega}_\infty, \beta_\infty, \theta_\infty)$ where $s(\boldsymbol{\omega}_k, \beta_k, \theta_k) \geq s(\boldsymbol{\omega}_\infty, \beta_\infty, \theta_\infty)$. Moreover, Algorithm 1 terminates finitely.*

Ideally, one would hope to get an optimal parameter θ^* which satisfies $\theta^* = \left(g_{(n_+)}(\theta^*)\sqrt{C}\right)^{-1}$. In practice, θ_∞ will approximate θ^* very well. In addition, we notice that one-step iteration usually gives decent results for simulation examples and some real examples.

5. Theoretical Properties

In this section, several important theoretical properties of the FLAME classifier are investigated. We first prove the Fisher consistency (Lin, 2004) of the FLAME in Section 5.1. As one focus of this paper is imbalanced data classification, the asymptotic properties for FLAME under extremely imbalanced data setting is studied in Section 5.2. Lastly, a novel HDLSS asymptotics where n is fixed and $d \to \infty$, the other focus of this article, is studied in Section 5.3.

5.1 Fisher Consistency and Large Sample Asymptotics

Fisher consistency is a very basic property for a classifier. A classifier is Fisher consistent implies that the minimizer of the conditional risk of the classifier given observation $\boldsymbol{x}$ has the same sign as the Bayes rule, $\underset{k\in\{+1,-1\}}{\operatorname{argmax}} \mathrm{P}(Y = k|\boldsymbol{X} = \boldsymbol{x})$. It has been shown that both SVM and DWD are Fisher consistent (Lin, 2004; Qiao et al., 2010). The following proposition states that the FLAME classifier is Fisher consistent too.

Proposition 2 *Let f^* be the global minimizer of the expected loss $\mathbb{E}[L(Yf(\boldsymbol{X}), \theta)]$, where $L(\cdot)$ is the loss function for the FLAME classifier, given parameters C and θ. Then* $\operatorname{sign}(f^*(\boldsymbol{x})) = \operatorname{sign}(\mathrm{P}(Y = +1|\boldsymbol{X} = \boldsymbol{x}) - 1/2)$.

Fisher consistency is also known as classification-calibrated, notably by Bartlett et al. (2006). With this weakest possible condition on the loss function, they extended the results of Zhang (2004) and showed that there was a nontrivial upper bound on the excess risk. Moreover, they were able to derive faster rates of convergence in some low noise settings. In particular, for a classification-calibrated loss function $L(\cdot)$, there exists a function $\psi : [-1, 1] \mapsto [0, \infty)$ so that $\psi(R_{0-1}(f) - R^*_{0-1}) \leq R_L(f) - R^*_L$ or $c(R_{0-1}(f) - R^*_{0-1})^\alpha \psi\left(\frac{(R_{0-1}(f)-R^*_{0-1})^\alpha}{2c}\right) \leq R_L(f) - R^*_L$ for some constant $c > 0$ with certain low noise parameter α, where $R_{0-1}(f)$ and $R_L(f)$ are the risk of the prediction function f with respect to the 0-1 loss and the loss function L respectively, and R^*_{0-1} and R^*_L are the corresponding Bayes risk and "optimal L-risk" respectively. The techniques in Zhang (2004) and Bartlett et al. (2006) can be directly applied to the FLAME classifier. The form of the ψ transform above which establishes the relations between the two excess risks, being applied to the current article, is given by Proposition 3.

Proposition 3 *The ψ-transform of the FLAME loss function with parameters C and θ is*

$$\psi(\gamma) = (2 - \theta)\sqrt{C} - H((1 + \gamma)/2),$$

where

$$H(\eta) = \begin{cases} \sqrt{C}\min(\eta, 1-\eta)(2+\frac{1}{\theta}-\theta), \ \textit{if } \eta < \frac{\theta^2}{1+\theta^2} \ \textit{ or } \eta > \frac{1}{1+\theta^2}, \\ \sqrt{C}[2\min(\eta, 1-\eta) - \theta + 2\sqrt{\eta(1-\eta)}], \ \textit{otherwise.} \end{cases} \tag{4}$$

These results provide bounds for the excess risk $R_{0-1}(f) - R_{0-1}^*$ in terms of the excess L-risk $R_L(f) - R_L^*$. Combined with a bound on the excess L-risk, they can give us a bound on the excess risk. Recent works for SVM have focused on fast rates of convergence. Vito et al. (2005) studied classification problems as inverse problems; Steinwart and Scovel (2007) studied the convergence properties of the standard SVM with Gaussian kernels; Blanchard et al. (2008) used a method called "localization". See also Chen et al. (2004) for another relevant work for the q-soft margin SVM.

5.2 Asymptotics under Imbalanced Setting

In this subsection, we investigate the asymptotic performance of SVM, DWD and FLAME. The asymptotic setting we focus on is when the minority sample size n_+ is fixed and the majority sample size $n_- \to \infty$, which is similar to the setting in Owen (2007). We will show that DWD is sensitive to the imbalanced data, while FLAME with proper choices of parameter θ and SVM are not.

Let $\overline{\boldsymbol{x}}_+$ be the sample mean of the positive/minority class. Theorem 4 shows that in the imbalanced data setting, when the size of the negative/majority class grows while that of the positive/minority class is fixed, the intercept term for DWD tends to negative infinity, in the order of $\sqrt{m}$. Therefore, DWD will classify all the observations to the negative/majority class, that is, the minority class will be 100% misclassified.

Theorem 4 *Let n_+ be fixed. Assume that the conditional distribution of the negative majority class $F_-(\boldsymbol{x})$ surrounds $\overline{\boldsymbol{x}}_+$ by the definition given in Owen (2007), and that γ is a constant satisfying $\inf_{\|\boldsymbol{\omega}\|=1} \int_{(\boldsymbol{x}-\overline{\boldsymbol{x}}_+)'\boldsymbol{\omega}>0} dF_-(\boldsymbol{x}) > \gamma \geq 0$, then the DWD intercept $\widehat{\boldsymbol{\beta}}$ satisfies*

$$\widehat{\boldsymbol{\beta}} < -\sqrt{\frac{\gamma}{C}m} - \overline{\boldsymbol{x}}_+^T\boldsymbol{\omega} = -\sqrt{\frac{n_-\gamma}{n_+C}} - \overline{\boldsymbol{x}}_+^T\boldsymbol{\omega}.$$

In Section 4, we have introduced an iterative approach to select the parameter θ. Theorem 5 shows that with the optimal parameter θ^* found by Algorithm 1, the discriminant direction of FLAME is in the same direction of the vector that joins the sample mean of the positive class and the *tilted* population mean of the negative class. Moreover, in contrast to DWD, the intercept term of FLAME in this case is finite.

Theorem 5 *Suppose that $n_- \gg n_+$ and $\boldsymbol{\omega}^*$ and β^* are the FLAME solutions trained with the parameter θ^* that satisfies $\theta^* = \left(g_{(n_+)}(\theta^*)\sqrt{C}\right)^{-1}$. Then $\boldsymbol{\omega}^*$ and β^* satisfy that*

$$\boldsymbol{\omega}^* = \frac{C}{(1+m)\lambda}\left[\overline{\boldsymbol{x}}_+ - \frac{\int(\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2}\boldsymbol{x}dF_-(\boldsymbol{x} \mid E)}{\int(\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2}dF_-(\boldsymbol{x} \mid E)}\right], \tag{5}$$

where E is the event that $[Y(\boldsymbol{X}^T\boldsymbol{\omega}^* + \beta^*)]^{-1} \geq \theta^*\sqrt{C}$ where $(\boldsymbol{X}, Y)$ is a random sample from the negative/majority class, and that

$$\int (\boldsymbol{x}^T\boldsymbol{\omega}^* + \beta^*)^{-2} dF_-(\boldsymbol{x} \mid E) = \frac{n_+}{n^o} C, \text{where } 0 < n^o \leq n_+.$$

Note that event E is $[Y(\boldsymbol{X}^T\boldsymbol{\omega}^* + \beta^*)]^{-1} \geq \theta^*\sqrt{C}$, which implies that the second term in (5) focuses on data vectors in the negative class with positive loss since their functional margins are less than $1/(\theta\sqrt{C})$. Recall that this is precisely the interpretation of FLAME (see Section 3.2), namely, to sample a subset of the majority class to have positive loss, so as to make the problem less imbalanced.

Remark: As a consequence of Theorem 5, when $m = n_-/n_+ \to \infty$, we have $\|\boldsymbol{\omega}^*\| \to 0$. Since the right-hand-side of the last equation above is positive and finite, β^* does not diverge. In addition, since $\mathrm{P}(\overline{E}) \to 1$ with probability converging to 1, $\beta^* < -1/(\theta\sqrt{C})$.

The following theorem shows the performance of SVM under the imbalanced data context, which completes our comparisons between SVM, DWD and FLAME.

Theorem 6 *Suppose that $n_- \gg n_+$. The solutions $\widehat{\boldsymbol{\omega}}$ and $\widehat{\beta}$ to SVM satisfy that*

$$\widehat{\boldsymbol{\omega}} = \frac{1}{(1+m)\lambda} \left\{ \overline{\boldsymbol{x}}_+ - \int \boldsymbol{x} dF_-(\boldsymbol{x} \mid G) \right\},$$

where G is the event that $1 - Y(\boldsymbol{X}^T\widehat{\boldsymbol{\omega}} + \widehat{\beta}) > 0$ where $(\boldsymbol{X}, Y)$ is a random sample from the negative/majority class, and that

$$\mathrm{P}(\overline{G}) = \mathrm{P}(1 + \boldsymbol{X}^T\widehat{\boldsymbol{\omega}} + \widehat{\beta} \leq 0) = 1 - 1/m.$$

Remark: The last statement in Theorem 6 means that with probability converging to 1, $\widehat{\beta} \leq -1$. However, note this is the only restriction that SVM solution has for the intercept term (recall that the counterpart in DWD is $\widehat{\boldsymbol{\beta}} < -\sqrt{\frac{\gamma}{C}m} - \overline{\boldsymbol{x}}_+^T\boldsymbol{\omega}$).

5.3 High-Dimensional, Low-Sample Size Asymptotics

HDLSS data are emerging in many areas of scientific research. The HDLSS asymptotics is a recently developed theoretical framework. Hall et al. (2005) gave a geometric representation for the HDLSS data, which can be used to study these new "n fixed, $d \to \infty$" asymptotic properties of binary classifiers such as SVM and DWD. Ahn et al. (2007) weakened the conditions under which the representation holds. Qiao et al. (2010) improved the conditions and applied this representation to investigate the performance of the weighted DWD classifier. Bolivar-Cime and Marron (2013) compared several binary classification methods in the HDLSS setting under the same theoretical framework. The same geometric representation can be used to analyze FLAME. See summary of some previous HDLSS results in Online Appendix 1. We develop the HDLSS asymptotic properties of the FLAME family by providing conditions in Theorem 7 under which the FLAME classifiers always correctly classify HDLSS data.

We first introduce the notations and give some regularity assumptions, then state the main theorem. Let $k \in \{+1, -1\}$ be the class index. For the kth class and given a fixed n_k,

consider a sequence of random data matrices $\boldsymbol{X}_1^k, \boldsymbol{X}_2^k, \cdots \boldsymbol{X}_d^k, \cdots$, indexed by the number of rows d, where each column of $\boldsymbol{X}_d^k$ is a random observation vector from $\mathbb{R}^d$ and each row represents a variable. Assume that each column of $\boldsymbol{X}_d^k$ comes from a multivariate distribution with dimension d and with covariance matrix $\boldsymbol{\Sigma}_d^k$ independently. Let $\lambda_{1,d}^k \geq \cdots \geq \lambda_{d,d}^k$ be the eigenvalues of the covariance, and $\left(\sigma_d^k\right)^2 = d^{-1}\sum_{i=1}^d \lambda_{i,d}^k$ the average eigenvalue. The eigenvalue decomposition of $\boldsymbol{\Sigma}_d^k$ is $\boldsymbol{\Sigma}_d^k = \boldsymbol{V}_d^k \boldsymbol{\Lambda}_d^k \left(\boldsymbol{V}_d^k\right)^T$. We may define the square root of $\boldsymbol{\Sigma}_d^k$ as $\left(\boldsymbol{\Sigma}_d^k\right)^{1/2} = \boldsymbol{V}_d^k \left(\boldsymbol{\Lambda}_d^k\right)^{1/2}$, and the inverse square root $\left(\boldsymbol{\Sigma}_d^k\right)^{-1/2} = \left(\boldsymbol{\Lambda}_d^k\right)^{-1/2}\left(\boldsymbol{V}_d^k\right)^T$. With minimal abuse of notation, let $\mathbb{E}(\boldsymbol{X}_d^k)$ denote the expectation of columns of $\boldsymbol{X}_d^k$. Lastly, the $n^k \times n^k$ *dual* sample covariance matrix is denoted by $\boldsymbol{S}_{D,d}^k = d^{-1}\{\boldsymbol{X}_d^k - \mathbb{E}(\boldsymbol{X}_d^k)\}^T\{\boldsymbol{X}_d^k - \mathbb{E}(\boldsymbol{X}_d^k)\}$.

Assumption 1 *There are five components:*

(i) Each column of $\boldsymbol{X}_d^k$ has mean $\mathbb{E}(\boldsymbol{X}_d^k)$ and the covariance matrix $\boldsymbol{\Sigma}_d^k$ of its distribution is positive definite.

(ii) The entries of $\boldsymbol{Z}_d^k \equiv \left(\boldsymbol{\Sigma}_d^k\right)^{-\frac{1}{2}}\{\boldsymbol{X}_d^k - \mathbb{E}(\boldsymbol{X}_d^k)\} = \left(\boldsymbol{\Lambda}_d^k\right)^{-\frac{1}{2}}\left(\boldsymbol{V}_d^k\right)^T\{\boldsymbol{X}_d^k - \mathbb{E}(\boldsymbol{X}_d^k)\}$ are independent.

(iii) The fourth moment of each entry of each column is uniformly bounded by $M > 0$ and the Wishart representation holds for each dual sample covariance matrix $\boldsymbol{S}_{D,d}^k$ associated with $\boldsymbol{X}_d^k$, that is,

$$d\boldsymbol{S}_{D,d}^k = \left\{\left(\boldsymbol{Z}_d^k\right)^T\left(\boldsymbol{\Lambda}_d^k\right)^{1/2}\left(\boldsymbol{V}_d^k\right)^T\right\}\left\{\boldsymbol{V}_d^k\left(\boldsymbol{\Lambda}_d^k\right)^{1/2}\boldsymbol{Z}_d^k\right\} = \sum_{i=1}^d \lambda_{i,d}^k \boldsymbol{W}_{i,d}^k,$$

where $\boldsymbol{W}_{i,d}^k \equiv \left(Z_{i,d}^k\right)^T Z_{i,d}^k$ and $Z_{i,d}$ is the ith row of $\boldsymbol{Z}_d^k$ defined above. It is called Wishart representation because if $\boldsymbol{X}_d^k$ is Gaussian, then each $\boldsymbol{W}_{i,d}^k$ follows the Wishart distribution $\mathcal{W}_{n^k}(1, \mathbf{I}_{n^k})$ independently.

(iv) The eigenvalues of $\boldsymbol{\Sigma}_d^k$ are sufficiently diffused, in the sense that

$$\epsilon_d^k = \frac{\sum_{i=1}^d (\lambda_{i,d}^k)^2}{(\sum_{i=1}^d \lambda_{i,d}^k)^2} \to 0 \quad as \quad d \to \infty. \tag{6}$$

(v) The sum of the eigenvalues of $\boldsymbol{\Sigma}_d^k$ is the same order as d, in the sense that $\left(\sigma_d^k\right)^2 = O(1)$ and $1/\left(\sigma_d^k\right)^2 = O(1)$.

Assumption 2 *The distance between the two population expectations satisfies,*

$$d^{-1}\left\|\mathbb{E}(\boldsymbol{X}_d^{(+1)}) - \mathbb{E}(\boldsymbol{X}_d^{(-1)})\right\|^2 \to \mu^2, \text{ as } d \to \infty.$$

Moreover, there exist constants σ^2 and τ^2, such that

$$\left(\sigma_d^{(+1)}\right)^2 \to \sigma^2, \text{ and } \left(\sigma_d^{(-1)}\right)^2 \to \tau^2.$$

Let $\nu^2 \equiv \mu^2 + \sigma^2/n_+ + \tau^2/n_-$. The following theorem gives the sure classification condition for FLAME, which includes SVM and DWD as special cases.

Theorem 7 *Without loss of generality, assume that $n_+ \le n_-$. The situation of $n+ > n_-$ is similar and omitted.*

- *If either one of the following three conditions is satisfied,*
  1. *for $\theta \in \left[0, (1+\sqrt{m^{-1}})/(\nu\sqrt{dC})\right)$, $\mu^2 > (n_-/n_+)^{\frac{1}{2}}\sigma^2/n_+ - \tau^2/n_- > 0$;*
  2. *for $\theta \in \left[(1+\sqrt{m^{-1}})/(\nu\sqrt{dC}), 2/(\nu\sqrt{dC})\right)$, $\mu^2 > T - \tau^2/n_- > 0$ where $T := \left(1/(2\theta\sqrt{dC}) + \sqrt{1/(4\theta^2 dC) + \sigma^2/n_+}\right)^2 - \sigma^2/n_+$;*
  3. *for $\theta \in \left[2/(\nu\sqrt{dC}), 1\right]$, $\mu^2 > \sigma^2/n_+ - \tau^2/n_- > 0$,*

  then for a new data point $\boldsymbol{x}_0^+$ from the positive class ($+1$),
  $\mathrm{P}(\boldsymbol{x}_0^+$ is correctly classified by FLAME$) \to 1$, as $d \to \infty$.
  Otherwise, the probability above $\to 0$.
- *If either one of the following three conditions is satisfied,*
  1. *for $\theta \in \left[0, (1+\sqrt{m^{-1}})/(\nu\sqrt{dC})\right)$, $(n_-/n_+)^{\frac{1}{2}}\sigma^2/n_+ - \tau^2/n_- > 0$;*
  2. *for $\theta \in \left[(1+\sqrt{m^{-1}})/(\nu\sqrt{dC}), 2/(\nu\sqrt{dC})\right)$, $T - \tau^2/n_- > 0$;*
  3. *for $\theta \in \left[2/(\nu\sqrt{dC}), 1\right]$, $\sigma^2/n_+ - \tau^2/n_- > 0$,*

  then for any $\mu > 0$, for a new data point $\boldsymbol{x}_0^-$ from the negative class (-1),
  $\mathrm{P}(\boldsymbol{x}_0^-$ is correctly classified by FLAME$) \to 1$, as $d \to \infty$.

Remark: Theorem 7 has two parts. The first part gives the conditions under which FLAME correctly classifies a new data point from the positive class, and the second part is for the negative class. Each part lists three conditions based on three disjoint intervals of parameter θ. Note the first and third intervals of each part generalize results which were shown to hold only for DWD and SVM before (*c.f.* Theorem 1 and Theorem 2 in Hall et al., 2005). In particular, it shows that all the FLAME classifiers with θ falling into the first interval behave like DWD asymptotically. Similarly, all the FLAME classifiers with θ falling into the third interval behave like SVM asymptotically. This partially explains the shape of the within-group error curve that we will show in Figure 6 (see also Figures A.2 and A.3 in Online Appendix 1), which we will discuss in the next section.

In the first part, the condition for other FLAMEs (with θ in the second interval) is weaker than the DWD-like FLAMEs (in the first interval), but stronger than the SVM-like FLAMEs (in the third interval). This means that it is easier to classify a new data point from the positive/minority class by SVM, than by an intermediate FLAME, which is easier than by DWD. Note that when $n_+ \le n_-$, the hyperplane for FLAME is in general closer to the positive class.

In terms of classifying data points from the negative class, the order of the difficulties among DWD, FLAME and SVM reverses.

6. Simulations

FLAME is not only a unified representation of DWD and SVM, but also introduces a new family of classifiers which has the potential of avoiding the overfitting HDLSS data issue and the sensitivity to imbalanced data issue. In this section, we use simulations to show the performance of FLAME at various parameter levels.

6.1 Measures of Performance

Before we introduce our simulation examples, we first introduce the performance measures in this paper. Note that the Bayes rule classifier can be viewed as the "gold standard" classifier. In our simulation settings, we assume that data are generated from two multivariate normal distributions with different mean vectors $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ and same covariance matrices $\boldsymbol{\Sigma}$. This setting leads to the following Bayes rule,

$$\text{sign}(\boldsymbol{x}^T\boldsymbol{\omega}_B + \beta_B) \text{ where } \boldsymbol{\omega}_B = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-) \text{ and } \beta_B = -\frac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)'\boldsymbol{\omega}_B. \quad (7)$$

Five performance measures are evaluated in this paper:

1. The *mean within-class error* (MWE) for an out-of-sample test set, which is defined as

$$MWE = \frac{1}{2n_+}\sum_{i=1}^{n_+}\mathbb{1}(\widehat{Y}_i^+ \neq Y_i^+) + \frac{1}{2n_-}\sum_{j=1}^{n_-}\mathbb{1}(\widehat{Y}_j^- \neq Y_j^-)$$

2. The *deviation* of the estimated intercept β from the Bayes rule intercept β_B: $|\beta - \beta_B|$.
3. *Dispersion*: a measure of the stochastic variability of the estimated discrimination direction vector $\boldsymbol{\omega}$. The dispersion measure was introduced in Section 1, as the trace of the sample covariance of the resulting discriminant direction vectors: disperson = $\text{Var}([\boldsymbol{\omega}_r]_{r=1:R})$ where R is the number of repeated runs.
4. *Angle* between the estimated discrimination direction $\boldsymbol{\omega}$ and the Bayes rule direction $\boldsymbol{\omega}_B$: $\angle(\boldsymbol{\omega}, \boldsymbol{\omega}_B)$.
5. *RankComp*$(\boldsymbol{\omega}, \boldsymbol{\omega}_B)$: In general, for two direction vectors $\boldsymbol{\omega}$ and $\boldsymbol{\omega}^*$, RankComp is defined as the proportion of the pairs of variables, among all $d(d-1)/2$ pairs, whose relative importances (in terms of their absolute values) given by the two directions are different, that is,

$$\text{RankComp}(\boldsymbol{\omega}, \boldsymbol{\omega}^*) \equiv \frac{1}{d(d-1)/2}\sum_{1\le i<j\le d}\mathbb{1}\left\{(|\omega_i| - |\omega_j|) \times (|\omega_i^*| - |\omega_j^*|) < 0\right\},$$

   where ω_i and ω_i^* are the ith components of the vectors $\boldsymbol{\omega}$ and $\boldsymbol{\omega}^*$ respectively. The RankComp measure can be viewed as a discretized analog to the angle between two vectors, and it provides more insights in the ranking of variables that a direction vector may suggest. We report the RankComp between the estimated direction $\boldsymbol{\omega}$ and the Bayes rule direction $\boldsymbol{\omega}_B$ to measure their closeness.

We will investigate these measures based on different dimensions d and different imbalance factors m.

6.2 Effects of Dimensions and Imbalanced Data

In Section 1, a specific FLAME ($\theta = 0.5$) has been compared with SVM ($\theta = 1$) and DWD ($\theta = 0$) in Figure 1, and on average, its discriminant directions are closer to the Bayes rule direction $\boldsymbol{\omega}_B$ compared to the SVM directions, but are less close than the DWD directions. In this subsection, we will further investigate the performance of FLAME with several

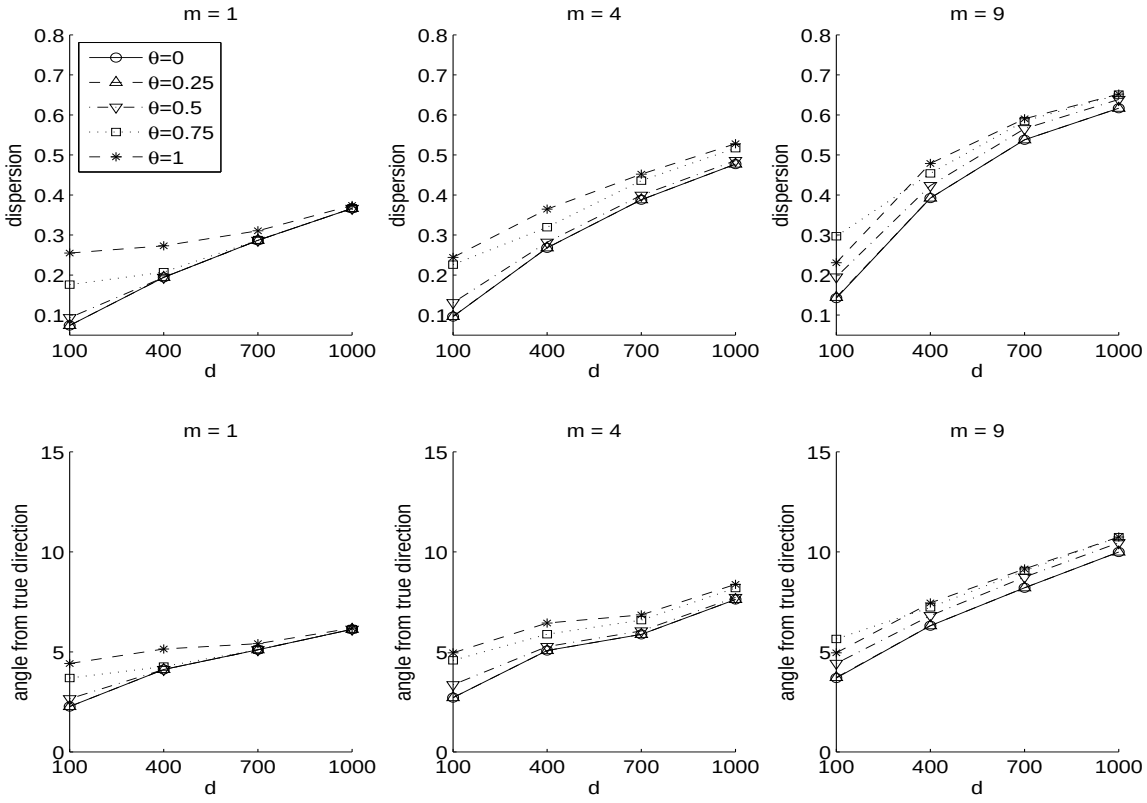

Figure 5: The dispersions (top row) and the angles between the FLAME direction and the Bayes direction (bottom row) for 50 runs of simulations, where the imbalance factors m are 1, 4 and 9 (the left, center and right panels), in the increasing dimension setting ($d = 100, 400, 700, 1000$; shown on the x-axes). The FLAME machines have $\theta = 0, 0.25, 0.5, 0.75, 1$ which are depicted using different curve styles (the first and the last cases correspond to DWD and SVM, respectively.) Note that with θ and the dimension d increase, both the dispersion and the deviation from the Bayes direction increase. The emergence of the imbalanced data (the increase of m) does not much deteriorate the FLAME directions except for large d.

different values of θ, and compare them with DWD and SVM under various simulation settings.

Figure 5 shows the comparison results under the same simulation setting with various combinations of (d, m)'s. In this simulation setting, data are from multivariate normal distributions with identity covariance matrix $MVN_d(\boldsymbol{\mu}_\pm, \mathbf{I}_d)$, where $d = 100, 400, 700$ and 1000. We let $\boldsymbol{\mu}_0 = c(d, d-1, d-2, \cdots, 1)^T$ where $c > 0$ is a constant which scales $\boldsymbol{\mu}_0$ to have norm 2.7. Then we let $\boldsymbol{\mu}_+ = \boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_- = -\boldsymbol{\mu}_0$. The imbalance factor varies among 1, 4 and 9 while the total sample size is 240. For each experiment, we repeat the simulation 50

times, and plot the average performance measure in Figure 5. The Bayes rule is calculated according to (7). It is obvious that when the dimension increases, both the dispersion and the angle increase. They are indicators of overfitting HDLSS data. When the imbalance factor m increases, the two measures increase as well, although not as much as when the dimension increases. More importantly, it shows that when θ decreases (from 1 to 0, or equivalently FLAME changes from SVM to DWD), the dispersion and the angle both decrease, which is promising because it shows that FLAME improves SVM on the overfitting issue.

6.3 Effects of Tuning Parameters with Covariance

We also investigate the effect of different covariance structures, since independence structure among variables as in the last subsection is less common in real applications. We investigate three covariance structures: independent, interchangeable and block-interchangeable covariance. Data are generated from two multivariate normal distributions $MVN_{300}(\boldsymbol{\mu}_{\pm}, \boldsymbol{\Sigma})$ with $d = 300$. We fist let $\boldsymbol{\mu}_1 = (75, 74, 73, \cdots, 1, 0, 0, \cdots, 0)'$, then scale it by multiply a constant c such that the Mahalanobis distance between $\boldsymbol{\mu}_+ = c\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_- = -c\boldsymbol{\mu}_1$ equals 5.4, that is, $(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-) = 5.4$. Note that this represents a reasonable signal-to-noise ratio.

We consider the FLAME machines with different parameter θ from a grid of 11 values $(0, 0.1, 0.2, \cdots, 1)$, and apply them to nine simulated examples (three different imbalance factors $(m = 2, 3, 4) \times$ three covariance structures). For the independent structure example, $\boldsymbol{\Sigma} = \mathbf{I}_{300}$; For the interchangeable structure example, $\boldsymbol{\Sigma}_{ii} = 1$ and $\boldsymbol{\Sigma}_{ij} = 0.8$ for $i \neq j$; For the block-interchangeable structure example, we let $\boldsymbol{\Sigma}$ be a block diagonal matrix with five diagonal blocks, the sizes of which are 150, 100, 25, 15, 10, and each block is an interchangeable covariance matrix with diagonal entries 1 and off-diagonal entries 0.8.

Figure 6 shows the results of the interchangeable structure example. Since the results under different covariance structures are similar, those for the other two covariance structures are reported in Online Appendix 1 to save space (Figure A.2 for the independent structure, and Figure A.3 for the block-interchangeable covariance).

In each plot, we include the within-group error (top-left), the absolute value of the difference between the estimated intercept and the Bayes intercept $|\beta - \beta_B|$ (top-middle), the angle between the estimated direction and the Bayes direction $\angle(\boldsymbol{\omega}, \boldsymbol{\omega}_B)$ (bottom-left), the RankComp between the estimated direction and the Bayes direction (bottom-middle) and the dispersion of the estimated directions (bottom-right).

We can see that in Figure 6 (and Figures A.2 and A.3 in Online Appendix 1), when we increase θ from 0 to 1, that is, when the FLAME moves from the DWD end to the SVM end, the within-group error decreases. This is mostly due to the fact that the intercept term β comes closer to the Bayes rule intercept β_B. On the other hand, the estimated direction is deviating from the true direction (larger angle), is giving the wrong rank of the variables (larger RankComp), and is more unstable (larger dispersion). Similar phenomena hold for the other two covariance structures, with one exception in the block interchangeable setting (Figure A.3 in Online Appendix 1) where the RankComp first decreases then increases.

In the entire FLAME family, DWD represents one extreme which provides better estimation of the direction, is closer to the Bayes direction, provides the right order for all

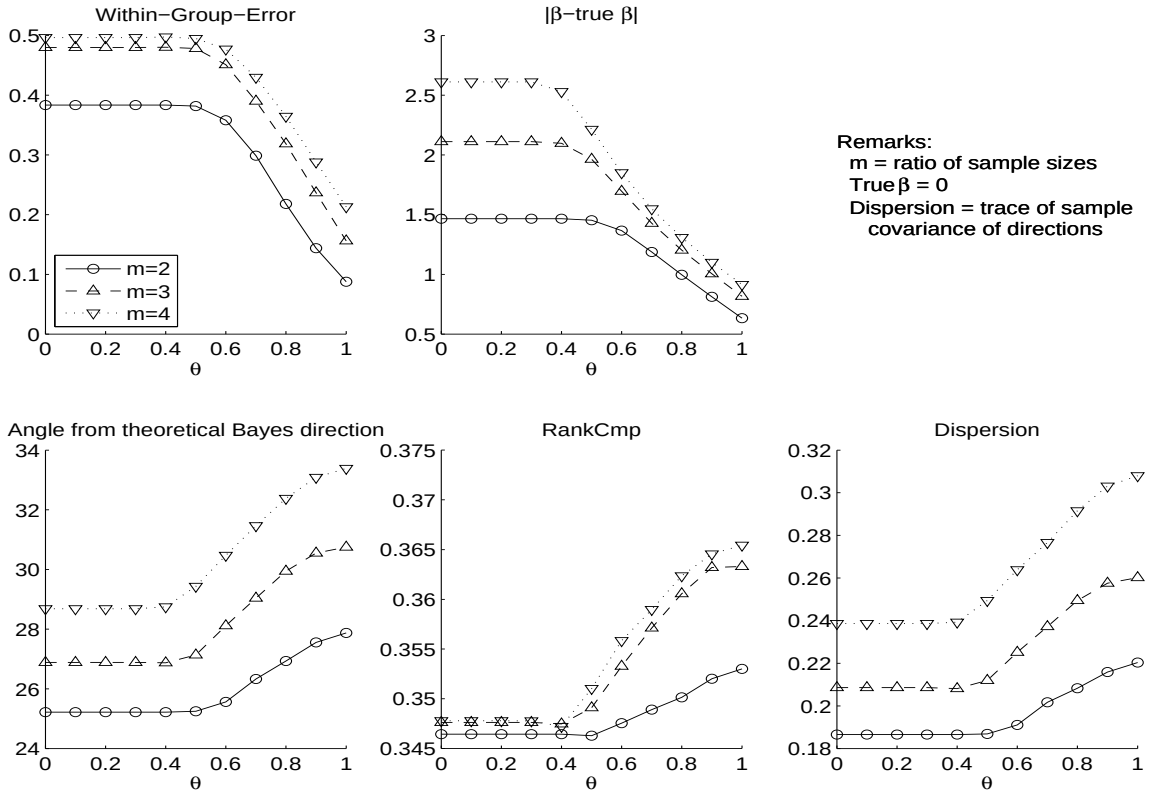

Figure 6: Interchangeable example. It can be seen that with FLAME turns from DWD to SVM (θ from 0 to 1), the within-class error decreases (top-left), thanks to the more accurate estimate of the intercept term (top-middle). On the other hand, this comes at the cost of larger deviation from the Bayes direction (bottom-left), incorrect rank of the importance of the variables (bottom-middle) and larger stochastic variability of the estimation directions (bottom-right).

variables, and is more stable. But it suffers from the inaccurate estimation of β in the presence of imbalanced data; SVM represents the other extreme, which is not sensible to imbalanced data and usually provides a good estimation of β, but is in general outperformed by DWD in terms of closeness to the Bayes optimal direction. In most situations, within the FLAME family, there is no single machine that is better than the both ends from the two aspects at the same time.

7. Real Data Application

In this section we demonstrate the performance of FLAME on a real example: the Human Lung Carcinomas Microarray Data set, which has been analyzed earlier in Bhattacharjee et al. (2001).

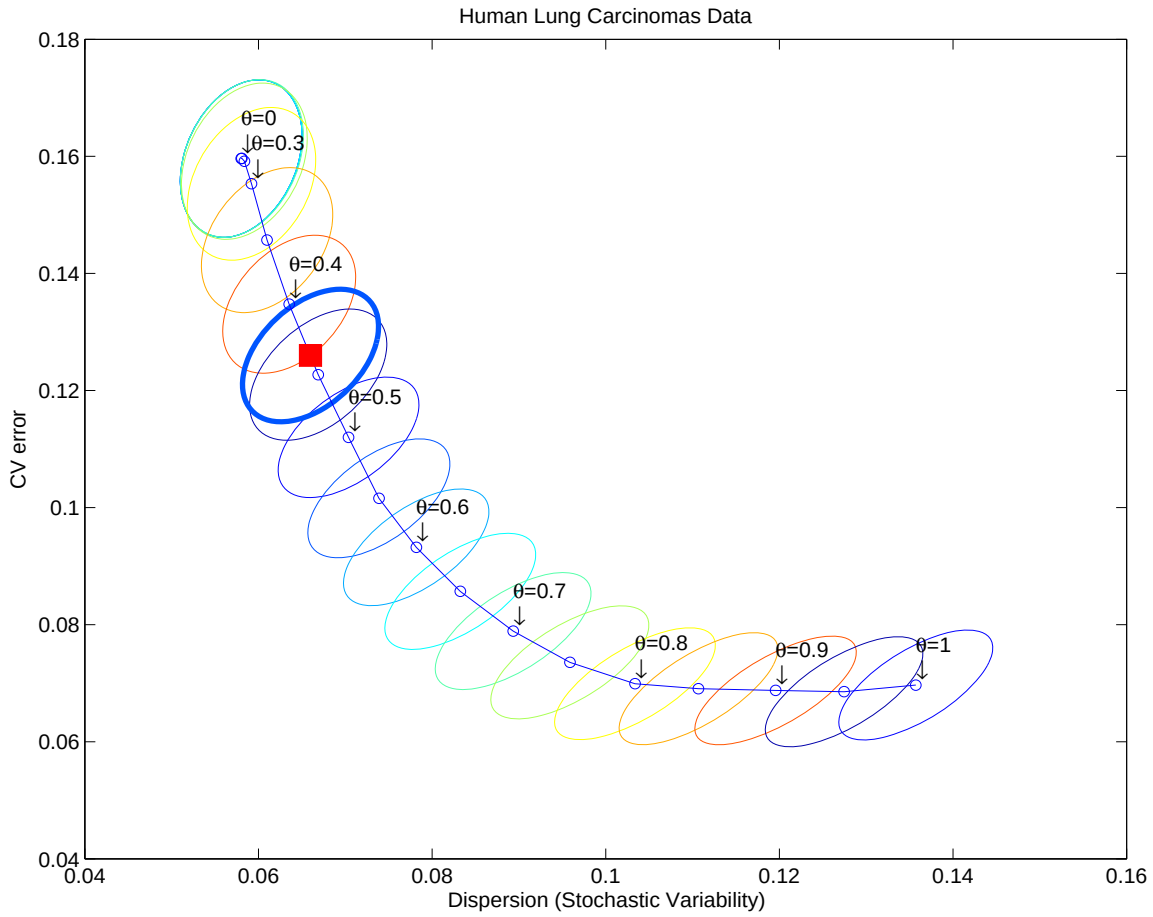

Figure 7: The dispersion and cross-validation error for the Human Lung Carcinomas Data set over 100 random splittings for different choices of θ values. The mean and standard error of the two measurements are depicted by ellipses as detour plots. The red square shows the performance for the adaptive parameter recommendation after one step. This plot shows clear trade-off between generalization error and stochastic variability.

The Human Lung Carcinomas Data set contains six classes: adenocarcinoma, squamous, pulmonary carcinoid, colon, normal and small cell carcinoma, with sample sizes of 128, 21, 20, 13, 17 and 6 respectively. Liu et al. (2008) used this data as a test set to demonstrate their proposed significance analysis of clustering approach. We combine the first two subclasses and the last four subclasses to form the positive and negative classes respectively. The sample sizes are 149 and 56 with imbalance factor $m = 2.66$. The original data contain 12,625 genes. We first filter genes using the ratio of the sample standard deviation and sample mean of each gene and keep 2,530 of them with large ratios (Dudoit et al., 2002; Liu et al., 2008).

We conduct five-fold cross-validations (CV) to evaluate the within-group error for the two classes over 100 random splits. In each split, we apply FLAME with 21 different θ values, ranging from $0, 0.05, 0.1, \ldots$ to 1. Because the true Bayes rule is unknown, we cannot evaluate the RankComp measure or the angle measure. Instead, we calculate the dispersion of the resulting direction vectors when conducting five-fold cross-validation. The adaptive value for θ (after one step) is calculated based on the DWD direction using all the samples in the data set, and the performance of the resulting FLAME is evaluated as well. We report the cross-validation error and the dispersion in a scatter plot.

Figure 7 shows the dispersion and cross-validation error. The mean and standard error of both measurements are depicted by ellipses as detour plots. This plot clears illustrates the existence of the trade-off between generalization error and stochastic variability.

This experiment shows that FLAME opens a new dimension to improve both the classification performance and the interpretative ability of the classifier. In particular, compared to SVM (FLAME with $\theta = 1$), we can probably choose $\theta = 0.7$ or 0.75 so that the stability of the classifier can be much improved at a very small cost of the generalization error. Compared to DWD (FLAME with $\theta = 0$), any increase in θ after 0.3 can lead to dramatic improvement of the cross-validation error, again with very minimal compromise of the stability. The optimal choice of θ seem to be ad-hoc, and depends on the preference of the user. Our adaptive parameter recommendation gives θ at around 0.44.

8. Conclusion and Discussion

In this paper, we thoroughly investigate SVM and DWD on their performance when applied to the HDLSS and imbalanced data. A novel family of binary classifiers called FLAME is proposed, where SVM and DWD are the two ends of the spectrum. On the DWD end, the estimation of the intercept term is deteriorated while it provides better estimation of the direction vector, and thus better handles the HDLSS data. On the hand, SVM is good at estimating the intercept term but not the direction and is subject to overfitting, and thus is more suitable for imbalanced data but not HDLSS data.

We conduct extensive study of the asymptotic properties of the FLAME family in three different flavors, the "d fixed, $n \to \infty$" asymptotics (Fisher consistency), the "d and n_+ fixed, $n_- \to \infty$" asymptotics (extremely imbalanced data), and the "n fixed, $d \to \infty$" asymptotics (the HDLSS asymptotics). These results explain the performance we have seen in the simulations and suggest that with a smart choice of θ, FLAME can properly handle both the HDLSS data and the imbalanced data, by improving the estimations of the direction and the intercept term.

The FLAME family can be immediately extended to multi-class classification, as was done for SVM and DWD such as in Weston and Watkins (1999); Crammer and Singer (2002); Lee et al. (2004) or Huang et al. (2013). Another natural extension is variable selection for FLAME.

The FLAME machines generalize the concepts of support vectors. In SVM, support vectors are referred to vectors that sit on or fall into the two hyperplanes corresponding to $u \leq 1$ (or $u \leq 1/\sqrt{C}$ for the modified version of Hinge loss (2)). In SVM, only support vectors have impacts on the final solution. DWD is the other extreme case where all the data vectors have some impacts. In the presence of imbalanced sample size, the fact that all the

data vectors influence the solution cause the optimization to ignore the minority class. The FLAME with $0 < \theta < 1$ is somewhere in the middle. For FLAME, part of the data vectors, more than the support vectors, but fewer than all the vectors, have impacts. Smart choice of θ means that one needs to include as many vectors, and as balanced influential samples, as possible. More vectors usually lead to mitigated overfitting, and balanced sample size of the influential vectors from two classes means that the sensitivity issue of the intercept term can be alleviated.

The authors are aware that it is possible to implement a two-step procedure to conduct binary linear classification. In the first step, a good direction is found, probably in the fashion of DWD; in the second step, a fine intercept is chosen by borrowing idea of SVM. This idea is elaborated in Qiao and Zhang (2015).

The choice of θ usually depends on the nature of the data and the scientific context. If the users prefer better classification performance over reasonable discrimination direction for interpretation of the data, θ may be chosen to be closer to 1. If the right direction is the first priority, then θ should be chosen to be closer to 0. Note that, under some circumstances, the primary goal is to obtain a direction vector which can provide a score $\boldsymbol{x}^T\boldsymbol{\omega}$ for each observation for further use, and the intercept parameter β is of no use at all. For example, some users may use a receiver operating characteristic (ROC) curve as a graphical tool to evaluate classification performance over different β value instead of using a single β value given by the classifier. In this case, a FLAME machine close to the DWD method may be ideal.

Qiao et al. (2010) considered the sample weighted versions of DWD and SVM. One could in theory extend the FLAME directly to the so-called weighted FLAME family. Such extension is quite straightforward. It is easy to see that all the classifiers in such a family are Fisher consistent with respect to the weighted 0-1 loss function. The intercept term from weighted FLAME does not diverge. Similar HDLSS asymptotic results to what are presented in the current article can be expected as well.

Acknowledgments

The first author's work was partially supported by Binghamton University Harpur College Dean's New Faculty Start-up Funds and Dean's Research Semester Awards for Junior Faculty, and a collaboration grant from the Simons Foundation (#246649). Both authors thank the Statistical and Applied Mathematical Sciences Institute (SAMSI) for the support and the hospitality during the 2012-13 Program on Statistical and Computational Methodology for Massive Data sets and the 2013-14 Program on Low-dimensional Structure in High-dimensional Systems.

Appendix A. Derivation of Modified Hinge Loss

Note that the original SVM formulation is $\underset{\tilde{\boldsymbol{\omega}},\tilde{\beta}}{\operatorname{argmin}} \sum \left(1 - y_i \tilde{f}(\boldsymbol{x}_i)\right)_+$, s.t. $\|\tilde{\boldsymbol{\omega}}\|^2 \leq C$, where $\tilde{f}(\boldsymbol{x}) = \boldsymbol{x}^T\tilde{\boldsymbol{\omega}} + \tilde{\beta}$. Here the coefficient vector $\tilde{\boldsymbol{\omega}}$ does not have unit norm. We let $\boldsymbol{\omega} = \tilde{\boldsymbol{\omega}}/\sqrt{C}$,

$\beta = \tilde{\beta}/\sqrt{C}$ and $f = \tilde{f}/\sqrt{C}$. Thus SVM solution is given by $\underset{\boldsymbol{\omega},\beta}{\operatorname{argmin}} \sum \left(1 - \sqrt{C} y_i f(\boldsymbol{x}_i)\right)_+$, s.t. $\|\boldsymbol{\omega}\|^2 \leq 1$, or equivalently, $\underset{\boldsymbol{\omega},\beta}{\operatorname{argmin}} \sum \left(\sqrt{C} - C y_i f(\boldsymbol{x}_i)\right)_+$, s.t. $\|\boldsymbol{\omega}\|^2 \leq 1$.

References

J. Ahn and J. S. Marron. The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259, 2010.

J. Ahn, J. S. Marron, K.M. Muller, and Y.Y. Chi. The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):760–766, 2007. doi: 10.1093/biomet/asm050. URL `http://biomet.oxfordjournals.org/content/94/3/760.short`.

P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. ISSN 0162-1459. doi: 10.1198/016214505000000907. URL `http://pubs.amstat.org/doi/abs/10.1198/016214505000000907`.

Arindam Bhattacharjee, William G Richards, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd, Javad Beheshti, Raphael Bueno, Michael Gillette, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24): 13790–13795, 2001. URL `http://www.pnas.org/content/98/24/13790.short`.

Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *The Annals of Statistics*, pages 489–531, 2008.

A Bolivar-Cime and J. S. Marron. Comparison of binary discrimination methods for high dimension low sample size data. *Journal of Multivariate Analysis*, 115:108–121, 2013.

N.V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004. doi: 10.1145/1007730.1007733.

Di-Rong Chen, Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Support vector machine soft margin classifiers: Error analysis. *Journal of Machine Learning Research*, 5:1143–1175, 2004.

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

K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. *Machine Learning*, 47(2):201–233, 2002.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, 2000.

R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, 2001.

S. Dudoit, J. Fridlyand, and T.P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002. doi: 10.1198/016214502753479248. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214502753479248.

Peter Hall, J. S. Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00510.x. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00510.x.

T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

Hanwen Huang, Yufeng Liu, Ying Du, Charles M Perou, D Neil Hayes, Michael J Todd, and J. S. Marron. Multiclass distance weighted discrimination. *Journal of Computational and Graphical Statistics*, 22(just-accepted):953–969, 2013.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. *Journal of the American Statistical Association*, 99(465):67–81, 2004. ISSN 0162-1459. doi: 10.1198/016214504000000098. URL http://pubs.amstat.org/doi/abs/10.1198/016214504000000098.

Y. Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68(1):73–82, 2004. ISSN 0167-7152. doi: 10.1016/j.spl.2004.03.002. URL http://www.sciencedirect.com/science/article/pii/S0167715204000707.

Y. Liu, D.N. Hayes, A. Nobel, and J. S. Marron. Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293, 2008. doi: 10.1198/016214508000000454. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214508000000454.

Y. Liu, H.H. Zhang, and Y. Wu. Hard or soft classification? large-margin unified machines. *Journal of the American Statistical Association*, 106(493):166–177, 2011. doi: 10.1198/jasa.2011.tm10319. URL http://pubs.amstat.org/doi/abs/10.1198/jasa.2011.tm10319.

J. S. Marron, M.J. Todd, and J. Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007. doi: 10.1198/016214507000001120. URL http://pubs.amstat.org/doi/abs/10.1198/016214507000001120.

A.B. Owen. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8:761–773, 2007. URL http://jmlr.csail.mit.edu/papers/v8/owen07a.html.

X. Qiao and Y. Liu. Adaptive weighted learning for unbalanced multicategory classification. *Biometrics*, 65(1):159–168, 2009. doi: 10.1111/j.1541-0420.2008.01017.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2008.01017.x/abstract.

X. Qiao, H.H. Zhang, Y. Liu, M.J. Todd, and J. S. Marron. Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statistical Association*, 105(489):401–414, 2010. doi: 10.1198/jasa.2010.tm08487. URL `http://pubs.amstat.org/doi/abs/10.1198/jasa.2010.tm08487`.

Xingye Qiao and Lingsong Zhang. Distance-weighted support vector machine. *Statistics and Its Interface*, 8:3, 2015.

Xiaotong Shen, George C Tseng, Xuegong Zhang, and Wing Hung Wong. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

Alexander J Smola, Peter L Bartlett, Bernhard Schölkopf, and Dale Schuurmans. *Advances in Large Margin Classifiers*, volume 1. MIT Press Cambridge, MA, 2000.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, pages 575–607, 2007.

K.C. Toh, M.J. Todd, and R.H. Tütüncü. Sdpt3–a matlab software package for semidefinite programming, version 1.3. *Optimization Methods and Software*, 11(1):545–581, 1999. URL `http://www.math.nus.edu.sg/~mattohkc/papers/guide.ps.Z`.

R.H. Tütüncü, K.C. Toh, and M.J. Todd. Solving semidefinite-quadratic-linear programs using sdpt3. *Mathematical Programming*, 95(2):189–217, 2003. URL `http://dx.doi.org/10.1007/s10107-002-0347-5`.

V.N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Ernesto D Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto D Giovannini, and Francesca Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, 2005.

Grace Wahba. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. *Advances in Kernel Methods-Support Vector Learning*, 6:69–87, 1999.

J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *European Symposium on Artificial Neural Networks*, pages 219–224, 1999.

Yichao Wu and Yufeng Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007. URL `http://amstat.tandfonline.com/doi/full/10.1198/016214507000000617`.

Lingsong Zhang and Xihong Lin. Some considerations of classification for high dimension low-sample size data. *Statistical Methods in Medical Research*, 22:537–550, 2013.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.