Support vector machine in high-dimension, low-sample-size settings

Yugo Nakayama

February 2020

Support vector machine in high-dimension, low-sample-size settings

Yugo Nakayama
(Doctoral Program in Mathematics)

Submitted to the Graduate School of
Pure and Applied Sciences
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in
Science
at the
University of Tsukuba

Contents

Preface

With the development of modern science, it has become possible to observe large-scale data. One of the features of such data is a high-dimension, low-sample-size. We call such data HDLSS. A divergence condition $d/n \to \infty$ is met for HDLSS data, where d is the data dimension and n is the sample size. HDLSS data is observed in many areas of modern science such as genetic microarrays, medical imaging, text recognition, finance, chemometrics, and so on. Researches on HDLSS data have been actively studied in various fields such as multivariate analysis and machine learning. Many methods of multivariate analysis rely on the large sample theory, so that we cannot apply some of them for high-dimensional data analysis. On the other hand, we can use machine learning methods for low-dimensional and high-dimensional data. However, their asymptotic properties seem not to have been sufficiently studied in the HDLSS context. In order to analyze HDLSS data, we need further analyses for multivariate analysis and machine learning.

Aoshima and Yata (2011) is one of the pioneer researches in high-dimensional data analysis, and they gave a broad perspective of high-dimensional statistical analysis such as a test of equality of two covariance matrices, classification and so on along with sample size determination to ensure prespecified accuracy for each inference. Regarding the classification problem, Aoshima and Yata (2014) gave the misclassification rate adjusted classifier for multiclass, high-dimensional data in which misclassification rates are no more than specified thresholds. Aoshima and Yata (2011, 2015b) gave geometric classifiers based on a geometric representation of HDLSS data. Ahn and Marron (2010) considered a classifier based on the maximal data piling direction. Aoshima and Yata (2019a) considered the distance-based classifier by using data transformation based on the eigenstructure. Noting that non-sparse situations often occur in high-dimensional settings, Aoshima and Yata (2019b) considered a family of quadratic classifiers and discussed asymptotic properties and optimality of the classifies under high-dimension, non-sparse settings.

In the field of machine learning, there are many studies about the classification in the context of supervised learning. For example, the support vector machine (SVM) has been an efficient tool for classification and pattern recognition in many areas. Hall et al. (2005) and Qiao and Zhang (2015) investigated the versatility of the linear SVM (LSVM) for high-dimensional data. Hall et al. (2005), Chan and Hall (2009) and Qiao and Zhang (2015) investigated asymptotic properties of the LSVM in the HDLSS context and showed a consistency property in the sense that the misclassification rates of the LSVM tend to zero as $d \to \infty$ under certain strict conditions in the HDLSS context. Chan and Hall (2009) gave scale-adjusted of the average distance, nearest neighbor and distance-based classifiers, including the LSVM. Huang (2017) investigated the SVM in the high-dimension, large-sample-size context as $d/n \to c > 0$. As long as we know, asymptotic properties of nonlinear SVMs seem not to have been sufficiently studied in the HDLSS context.

In this thesis, we consider tests of covariance matrix structures and asymptotic properties of the SVM in the HDLSS framework. This thesis consists of four chapters.

In Chapter 1, we consider a test of the sphericity for high-dimensional covariance matrices. This chapter is organized by the findings of Yata et al. (2018). We construct a test statistic by using the

extended cross-data-matrix (ECDM) methodology proposed by Yata and Aoshima (2013). We show that the ECDM test statistic is based on an unbiased estimator of a sphericity measure. In addition, the ECDM test statistic enjoys consistency properties and the asymptotic normality in high-dimensional settings. We propose a new test procedure based on the ECDM test statistic and evaluate its asymptotic size and power theoretically and numerically. We give a two-stage sampling scheme so that the test procedure can ensure a prespecified level both for the size and power. We apply the test procedure to detect divergently spiked noise in high-dimensional statistical analysis. We analyze gene expression data by the proposed test procedure.

In Chapter 2, we consider asymptotic properties of the hard-margin LSVM (hmLSVM) in HDLSS settings. This chapter is organized by the findings of Nakayama et al. (2017). We show that the LSVM holds a consistency property in which misclassification rates tend to zero as the dimension goes to infinity under certain severe conditions. We show that the LSVM is very biased in HDLSS settings and its performance is affected by the bias directly. In order to overcome such difficulties, we propose a bias-corrected LSVM (BC-LSVM). We show that the BC-LSVM gives preferable performances in HDLSS settings. We also discuss the LSVMs in multiclass HDLSS settings. Finally, we check the performance of the classifiers in real data analyses.

In Chapter 3, we investigate behaviors of the soft-margin LSVM (smLSVM) for the regularization parameter. This chapter is organized by the findings of Nakayama (2019). We show that the smLSVM cannot handle imbalanced classification and the smLSVM is very biased in HDLSS settings. In order to overcome such difficulties, we propose a robust LSVM (RSVM). We show that the RSVM gives preferable performances in HDLSS settings.

In Chapter 4, we study asymptotic properties of nonlinear SVMs in HDLSS settings. This chapter is organized by the findings of Nakayama et al. (2019). We propose a bias-corrected SVM (BC-SVM) which is robust against imbalanced data in a general framework. In particular, we investigate asymptotic properties of the BC-SVM having the Gaussian kernel and compare them with the ones having the linear kernel. We show that the performance of the BC-SVM is influenced by the scale parameter involved in the Gaussian kernel. We discuss a choice of the scale parameter yielding a high performance and examine the validity of the choice by numerical simulations and real data analyses.

Acknowledgements

I would like to express my deepest appreciation to my supervisor, Professor Makoto Aoshima. He always encouraged me and gave his enthusiastic guidance, helpful support and patience to me. This dissertation would not have been possible without his advice and stimulating comments. He also provided me the tremendous research environment. I could engage in my research project strenuously. I would also like to thank Associate Professor Kazuyoshi Yata for his variable comments. Finally, I am deeply grateful to my parents for their patience and supports.

Chapter 1

Tests of covariance structures in the HDLSS context

In this chapter, we consider tests of covariance structures in the HDLSS context. This chapter is organized by Yata et al. (2018).

We consider a test of the sphericity for high-dimensional covariance matrices. When $n > d$ and d is fixed, Nagao (1973) and others gave test statistics for the sphericity by using the large sample theory. Ledoit and Wolf (2002) investigated asymptotic properties of the test statistics when $d/n \to c > 0$. Since the conventional test statistics do not work for HDLSS data, Srivastava et al. (2011) gave a test statistic under $n,\ d \to \infty$ and $n/d \to 0$. However, the test statistic is heavily biased for high-dimensional data unless the population data follows Gaussian. On the other hand, Chen et al. (2010) gave a test statistic based on the U-statistic for high-dimensional data. In this chapter, we shall also pursue a non-parametric approach, however we produce a new test statistic by using the ECDM methodology. The ECDM method was developed by Yata and Aoshima (2013) and was motivated by the cross-data-matrix (CDM) method due to Yata and Aoshima (2012b). One of the advantages of the ECDM method is that one can produce an unbiased estimator having a small variance at a low computational cost even for ultra high-dimensional data. In addition, the ECDM method possesses a high versatility in high-dimensional data analysis.

In Section 1.2, we introduce the ECDM methodology and produce a test statistic for the sphericity by using it. We show that the ECDM test statistic is based on an unbiased estimator of a sphericity measure. In addition, the ECDM test statistic enjoys consistency properties and the asymptotic normality in high-dimensional settings. We propose a new test procedure based on the ECDM test statistic and evaluate its asymptotic size and power theoretically.

In Section 1.3, we give a two-stage sampling scheme so that the test procedure can ensure a prespecified level both for the size and power.

In Section 1.4, we apply the test procedure to detect divergently spiked noise in high-dimensional statistical analysis.

In Section 1.5, we give simulation studies to investigate the performance of the proposed test procedure. Finally, we analyze gene expression data by the proposed test procedure.

1.1 Introduction

Suppose we take samples, $\boldsymbol{x}_j,\ j=1,\ldots,n$, of size n (≥ 4) from a population, which are independent and identically distributed (i.i.d.) as a d-variate distribution. We assume that $\boldsymbol{x}_j$ has an unknown mean vector $\boldsymbol{\mu}$ and unknown covariance matrix $\boldsymbol{\Sigma}$. We denote the eigenvalue decomposition of $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma}=\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues, $\lambda_1\geq\cdots\geq\lambda_d\geq 0$, and $\boldsymbol{H}$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{x}_j=\boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{z}_j+\boldsymbol{\mu}$, where $\boldsymbol{z}_j=(z_{1j},\ldots,z_{dj})^T$ is considered as a sphered data vector having the zero mean vector and identity covariance matrix. Let $\sigma=\mathrm{tr}(\boldsymbol{\Sigma})/d$. We assume that $\sigma\in(0,\infty)$ as $d\to\infty$. For a function, $f(\cdot)$, "$f(d)\in(0,\infty)$ as $d\to\infty$" implies that $\liminf_{d\to\infty}f(d)>0$ and $\limsup_{d\to\infty}f(d)<\infty$. We consider the following model:

$$\boldsymbol{x}_j=\boldsymbol{\Gamma}\boldsymbol{w}_j+\boldsymbol{\mu}, \tag{1.1}$$

where $\boldsymbol{\Gamma}=(\boldsymbol{\Gamma}_1,\ldots,\boldsymbol{\Gamma}_q)$ is a $d\times q$ matrix for some $q>0$ such that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T=\boldsymbol{\Sigma}$, and $\boldsymbol{w}_j=(w_{1j},\ldots,w_{qj})^T$, $j=1,\ldots,n$, are i.i.d. random vectors having $E(\boldsymbol{w}_j)=\boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{w}_j)=\boldsymbol{I}_q$. Here, $\boldsymbol{I}_q$ denotes the identity matrix of dimension q. Let $\mathrm{Var}(w_{rj}^2)=M_r$, $r=1,\ldots,q$. We assume that $M_r\in(0,\infty)$ as $d\to\infty$ for all r. Similar to Bai and Saranadasa (1996) and Aoshima and Yata (2019a), we assume that

(A-i) $E(w_{rj}^2w_{sj}^2)=E(w_{rj}^2)E(w_{sj}^2)=1$ and $E(w_{rj}w_{sj}w_{tj}w_{uj})=0$ for all $r\neq s,t,u$.

We assume the following assumption instead of (A-i) as necessary:

(A-ii) $E(w_{r_1j}^{\alpha_1}w_{r_2j}^{\alpha_2}\cdots w_{r_vj}^{\alpha_v})=E(w_{r_1j}^{\alpha_1})E(w_{r_2j}^{\alpha_2})\cdots E(w_{r_vj}^{\alpha_v})$ for all $r_1\neq r_2\neq\cdots\neq r_v\in[1,q]$ and $\alpha_i\in[1,4]$, $i=1,\ldots,v$, where $v\leq 8$ and $\sum_{i=1}^v\alpha_i\leq 8$.

See Chen and Qin (2010) about (A-ii). Note that (A-ii) implies (A-i). When $\boldsymbol{x}_j$ is Gaussian, it holds that $\boldsymbol{\Gamma}=\boldsymbol{H}\boldsymbol{\Lambda}^{1/2}$ and $\boldsymbol{w}_j=\boldsymbol{z}_j$ in (1.1). Note that (A-ii) is naturally satisfied when $\boldsymbol{x}_j$ is Gaussian because the elements of $\boldsymbol{z}_j$ are independent and $M_r=2$ for all r. We assume the following HDLSS divergence condition:

(A-iii) $d,n\to\infty$ and $n/d\to 0$.

In this paper, we are interested in testing the sphericity of $\boldsymbol{\Sigma}$:

$$H_0:\boldsymbol{\Sigma}=\sigma\boldsymbol{I}_d\quad\text{vs.}\quad H_1:\boldsymbol{\Sigma}\neq\sigma\boldsymbol{I}_d. \tag{1.2}$$

We give a two-stage test procedure which can ensure a prespecified level both for the size and power. Most interestingly, we apply the test procedure to detect divergently spiked noise in high-dimensional statistical analysis.

Let $\boldsymbol{\Sigma}_*=\boldsymbol{\Sigma}-\sigma\boldsymbol{I}_d$ and $\Delta=||\boldsymbol{\Sigma}_*||_F^2=\mathrm{tr}(\boldsymbol{\Sigma}^2)-\sigma^2$, where $||\cdot||_F$ is the Frobenius norm. Note that $\Delta=0$ under H_0 in (1.2) and Δ is regarded as a sphericity measure. See Ahn et al. (2007) for the sphericity measure.

We produce a test statistic by using the extended cross-data-matrix (ECDM) methodology. We show that the ECDM test statistic is based on an unbiased estimator of a sphericity measure. In addition, the ECDM test statistic enjoys consistency properties and the asymptotic normality in high-dimensional settings. We propose a new test procedure based on the ECDM test statistic and evaluate its asymptotic size and power theoretically and numerically. We give a two-stage sampling scheme so that the test procedure can ensure a prespecified level both for the size and power. We apply the test procedure to detect divergently spiked noise in high-dimensional statistical analysis. We analyze gene expression data by the proposed test procedure. Finally, in Section 1.5, we give simulation studies and real data analysis to check performances of our test procedures.

1.2 Test for sphericity

We give an unbiased estimator of Δ by using the extended cross-data-matrix (ECDM) methodology.

1.2.1 Extended cross-data-matrix methodology

In this section, we introduce the ECDM methodology. The ECDM methodology was developed by Yata and Aoshima (2013) as an extension of the CDM method due to Yata and Aoshima (2012b). One of the advantages of the ECDM method is that one can produce an unbiased estimator having a small variance at a low computational cost even for ultra high-dimensional data. See Section 2.5 of Yata and Aoshima (2013) for the details. Let $n_{(1)} = \lceil n/2 \rceil$ and $n_{(2)} = n - n_{(1)}$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. Let

$$\boldsymbol{V}_{n(1)(k)} = \begin{cases} \{\lfloor k/2 \rfloor - n_{(1)} + 1, \ldots, \lfloor k/2 \rfloor\} & \text{if } \lfloor k/2 \rfloor \geq n_{(1)}, \\ \{1, \ldots, \lfloor k/2 \rfloor\} \cup \{\lfloor k/2 \rfloor + n_{(2)} + 1, \ldots, n\} & \text{otherwise;} \end{cases}$$

$$\boldsymbol{V}_{n(2)(k)} = \begin{cases} \{\lfloor k/2 \rfloor + 1, \ldots, \lfloor k/2 \rfloor + n_{(2)}\} & \text{if } \lfloor k/2 \rfloor \leq n_{(1)}, \\ \{1, \ldots, \lfloor k/2 \rfloor - n_{(1)}\} \cup \{\lfloor k/2 \rfloor + 1, \ldots, n\} & \text{otherwise} \end{cases}$$

for $k = 3, \ldots, 2n-1$, where $\lfloor x \rfloor$ denotes the largest integer $\leq x$. Let $\#\boldsymbol{S}$ denote the number of elements in a set $\boldsymbol{S}$. Note that $\#\boldsymbol{V}_{n(l)(k)} = n_{(l)}$, $l = 1, 2$, $\boldsymbol{V}_{n(1)(k)} \cap \boldsymbol{V}_{n(2)(k)} = \emptyset$ and $\boldsymbol{V}_{n(1)(k)} \cup \boldsymbol{V}_{n(2)(k)} = \{1, \ldots, n\}$ for $k = 3, \ldots, 2n-1$. Also, note that

$$i \in \boldsymbol{V}_{n(1)(i+j)} \quad \text{and} \quad j \in \boldsymbol{V}_{n(2)(i+j)} \quad \text{for} \quad i < j \ (\leq n). \tag{1.3}$$

See Figure 1.1.

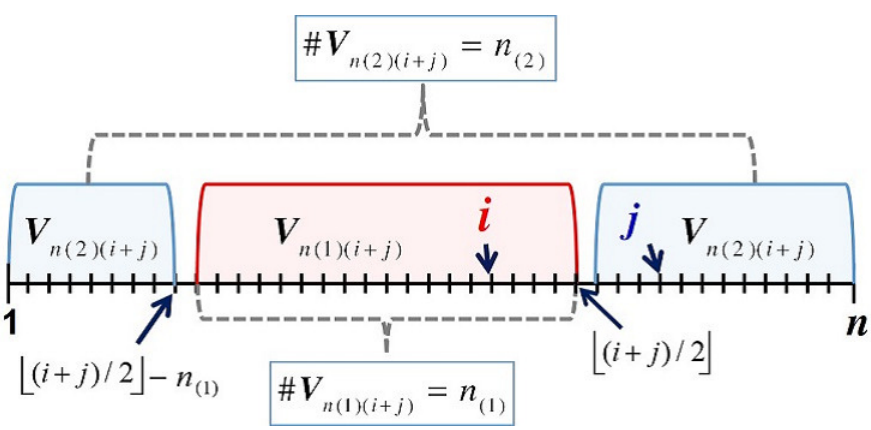


Figure 1.1: Illustration of (1.3) when $\lfloor (i+j)/2 \rfloor > n_{(1)}$.

Let

$$\overline{\boldsymbol{x}}_{(1)(k)} = n_{(1)}^{-1} \sum_{j \in \boldsymbol{V}_{n(1)(k)}} \boldsymbol{x}_j \quad \text{and} \quad \overline{\boldsymbol{x}}_{(2)(k)} = n_{(2)}^{-1} \sum_{j \in \boldsymbol{V}_{n(2)(k)}} \boldsymbol{x}_j$$

for $k = 3, \ldots, 2n-1$. From (1.3), we note that $(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{(1)(i+j)})$ and $(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{(2)(i+j)})$ are independent for all $i < j$. Then, Yata and Aoshima (2013) gave an estimator of $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ by the ECDM method as

$$W_n = \frac{2u_n}{n(n-1)} \sum_{i<j}^{n} \left\{ (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{(1)(i+j)})^T (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{(2)(i+j)}) \right\}^2, \tag{1.4}$$

where $u_n = n_{(1)}n_{(2)}/\{(n_{(1)}-1)(n_{(2)}-1)\}$. Note that $E(W_n) = \mathrm{tr}(\boldsymbol{\Sigma}^2)$. Aoshima and Yata (2015a) and Yata and Aoshima (2016) gave the following result.

Lemma 1.1 (Aoshima and Yata (2015a); Yata and Aoshima (2016))**.** *Assume (A-i). Then, it holds that as $d, n \to \infty$*

$$\mathrm{Var}\Big(\frac{W_n}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\Big) = \frac{4}{n^2}\{1+o(1)\} + O\Big(\frac{\mathrm{tr}(\boldsymbol{\Sigma}^4)}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2 n}\Big) \to 0.$$

1.2.2 Unbiased estimator of sphericity measure

We can give an estimator of $\sigma^2 d$ by the ECDM method as

$$U_n = \frac{2u_n}{dn(n-1)} \sum_{i<j}^{n} ||\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{n(1)(i+j)}||^2 \, ||\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{n(2)(i+j)}||^2, \tag{1.5}$$

where $||\cdot||$ denotes the Euclidean norm. Note that $E(U_n) = \mathrm{tr}(\boldsymbol{\Sigma})^2/d = \sigma^2 d$. We have the following result.

Lemma 1.2. *Assume (A-i). Then, it holds that as $d, n \to \infty$*

$$\mathrm{Var}\Big(\frac{U_n}{\sigma^2 d}\Big) = O\Big(\frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2 n}\Big) \to 0.$$

Finally, we construct an estimator of Δ by the ECDM method as

$$T_n = W_n - U_n. \tag{1.6}$$

We note that $E(T_n) = \Delta$ without any assumptions. We have the following result.

Lemma 1.3. *Assume (A-i). Then, it holds that as $d, n \to \infty$*

$$\mathrm{Var}(T_n) = 4\frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2}{n^2}\{1+o(1)\} + O\Big(\frac{\mathrm{tr}(\boldsymbol{\Sigma}^4)}{n^2} + \frac{\mathrm{tr}\{(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_*)^2\}}{n}\Big).$$

1.2.3 New test procedure for sphericity

For T_n given by (1.6), we have the following results.

Lemma 1.4. *Assume (A-i) and*

(A-iv) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{n\Delta} \to 0$ *under (A-iii).*

Then, it holds that under (A-iii)

$$\frac{T_n}{\Delta} = 1 + o_P(1).$$

Lemma 1.5. *Assume (A-ii) and*

(A-v) $\limsup \Big\{\dfrac{n\Delta}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\Big\} < \infty$ *under (A-iii).*

Then, it holds that under (A-iii)

$$\frac{T_n - \Delta}{2\mathrm{tr}(\boldsymbol{\Sigma}^2)/n} \Rightarrow N(0,1),$$

where "$\Rightarrow$" denotes the convergence in distribution and $N(0,1)$ denotes a random variable distributed as the standard normal distribution.

Note that $\mathrm{tr}(\boldsymbol{\Sigma}^2) = \sigma^2 d$ under H_0 in (1.2). From Lemma 1.5 we propose a test procedure for (1.2) by

$$\text{rejecting } H_0 \Longleftrightarrow \frac{nT_n}{2U_n} > z_\alpha, \tag{1.7}$$

where z_α is a constant such that $P\{N(0,1) > z_\alpha\} = \alpha$ with $\alpha \in (0, 1/2)$. Then, we have the following result.

Theorem 1.1. *Assume (A-ii) and (A-v). For the test by (1.7), we have that under (A-iii)*

$$\textit{Size} = \alpha + o(1) \quad \textit{and} \quad \textit{Power} = \Phi\Big(\frac{n\Delta}{2\mathrm{tr}(\boldsymbol{\Sigma}^2)} - z_\alpha\Big) + o(1), \tag{1.8}$$

where $\Phi(\cdot)$ denotes the c.d.f. of $N(0,1)$.

When (A-iv) is met, we have the following result.

Corollary 1.1. *Assume (A-i). Assume (A-iv) under H_1. For the test by (1.7), we have that under (A-iii)*

$$\textit{Power} = 1 + o(1).$$

Remark 1. *Chen et al. (2010) gave a test procedure for (1.2) based on the following statistic:*

$$T_{CZZ} = A_n - \mathrm{tr}(\boldsymbol{S}_n)^2/d,$$

where $\boldsymbol{S}_n$ is the sample covariance matrix having $E(\boldsymbol{S}_n) = \boldsymbol{\Sigma}$, and

$$\begin{aligned} A_n =& \frac{1}{n(n-1)} \sum_{j \neq j'}^{n} (\boldsymbol{x}_j^T \boldsymbol{x}_{j'})^2 - \frac{2}{n(n-1)(n-2)} \sum_{j \neq j' \neq j''}^{n} \boldsymbol{x}_{j'}^T \boldsymbol{x}_j \boldsymbol{x}_j^T \boldsymbol{x}_{j''} \\ &+ \frac{1}{n(n-1)(n-2)(n-3)} \sum_{j \neq j' \neq l \neq l'}^{n_i} \boldsymbol{x}_j^T \boldsymbol{x}_{j'} \boldsymbol{x}_l^T \boldsymbol{x}_{l'}. \end{aligned}$$

Note that $E(A_n) = \mathrm{tr}(\boldsymbol{\Sigma}^2)$. However, T_{CZZ} is biased for high-dimensional data because $E\{\mathrm{tr}(\boldsymbol{S}_n)^2\} > \mathrm{tr}(\boldsymbol{\Sigma})^2$. Although the test by Chen et al. (2010) is asymptotically equivalent to the test by (1.7), the latter is much more applicable to the sequential analysis ensuring prespecified accuracy as seen in the next section.

1.3 Two-stage sampling scheme to control both size and power

We are interested in designing a test of (1.2) having size α and power no less than $1 - \beta$ when $\Delta \geq \Delta_L$, where $\alpha \in (0, 1/2)$, $\beta \in (0, 1/2)$ and Δ_L (> 0) are prespecified constants. We assume that $\Delta_L \to \infty$ and $\Delta_L = o(d)$ as $d \to \infty$.

From Theorem 1.1 we consider n satisfying

$$\frac{n\Delta}{2\mathrm{tr}(\boldsymbol{\Sigma}^2)} - z_\alpha \geq z_\beta \quad \text{when} \quad \Delta \geq \Delta_L.$$

Then, one finds the sample size as

$$n \geq \frac{2(z_\alpha + z_\beta)\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\Delta_L} \quad (= C, \text{ say}). \tag{1.9}$$

We note that $C \to \infty$ as $d \to \infty$ from the facts that $\mathrm{tr}(\boldsymbol{\Sigma}^2) \geq \sigma^2 d$ and $\Delta_L = o(d)$ as $d \to \infty$. Also, note that $C/d \to 0$ as $d \to \infty$ under H_1 in (1.2) from the fact that $\Delta_L \to \infty$ as $d \to \infty$. Then, from Theorem 1.1, we have the following result.

Theorem 1.2. *Assume (A-ii) and (A-v). For the test by (1.7) with $n \geq C$ given by (1.9), we have under (A-iii)*

$$\textit{Size} = \alpha + o(1), \quad \textit{and} \quad \textit{Power} \geq 1 - \beta + o(1) \quad \textit{when} \;\; \Delta \geq \Delta_L. \tag{1.10}$$

Since C includes unknown parameter $\mathrm{tr}(\boldsymbol{\Sigma}^2)$, it is necessary to estimate C with some pilot samples. We proceed with the following two steps:

1. Choose $m(\geq 4)$ satisfying

$$\frac{m}{C} \leq 1, \quad \frac{C}{m^2} \to 0 \;\text{ and }\; \frac{C}{m}\frac{\mathrm{tr}(\boldsymbol{\Sigma}^4)}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2} \to 0 \;\text{ as }\; d \to \infty \;\text{ under }\; \frac{\mathrm{tr}(\boldsymbol{\Sigma}^4)}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2} \to 0 \;\text{ as }\; d \to \infty. \tag{1.11}$$

Take pilot samples, $\boldsymbol{x}_j,\ j = 1, \ldots, m$, of size m. Then, calculate W_m according to (1.4). Define the total sample size by

$$N = \max\left\{m,\ \left\lceil \frac{2(z_\alpha + z_\beta)W_m}{\Delta_L} \right\rceil\right\}. \tag{1.12}$$

2. If $N = m$, do not take any additional samples and otherwise, that is if $N > m$, take additional samples, $\boldsymbol{x}_j,\ j = m+1, \ldots, N$, of size $N - m$. By combining the pilot samples and the additional samples, calculate U_N and T_N according to (1.5) and (1.5). Then, we propose a test procedure for (1.2) by

$$\text{rejecting } H_0 \Longleftrightarrow \frac{NT_N}{2U_N} > z_\alpha. \tag{1.13}$$

We have the following result.

Theorem 1.3. *Assume (A-ii). Assume also*

(A-vi) $\displaystyle\limsup_{d\to\infty} \left\{\frac{C\Delta}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right\} < \infty.$

For the test by (1.13), we have (1.10) as $d \to \infty$.

Remark 2. *Under (A-vi), the condition "$\mathrm{tr}(\boldsymbol{\Sigma}^4)/\mathrm{tr}(\boldsymbol{\Sigma}^2)^2 \to 0$ as $d \to \infty$" in (1.16) holds. From Lemma 1.2, under (A-i) and (1.12), we have that $W_m = \mathrm{tr}(\boldsymbol{\Sigma}^2)\{1 + o_P(C^{-1/2})\}$ as $d \to \infty$. Then, it holds that $N - C = o_P(C^{1/2})$ as $d \to \infty$.*

1.4 Detection of divergently spiked noise

In this section, we consider the detection of divergently spiked noise as an application of the sphericity test. Paul (2007) and Jung and Marron (2009) handled the following multicomponent covariance model:

$$\boldsymbol{x}_j = \boldsymbol{\mu} + \sum_{i=1}^{k} \boldsymbol{\rho}_i \xi_{ij} + \tau^{1/2}\boldsymbol{\varepsilon}_j \;\text{ for } j = 1, \ldots, n, \tag{1.14}$$

where $\tau \in (0, \infty)$ as $d \to \infty$, ξ_{ij}s are i.i.d. as $N(0,1)$, $\boldsymbol{\varepsilon}_j$s are i.i.d. as $N_d(\mathbf{0}, \boldsymbol{I}_d)$, and ξ_{ij}s and $\boldsymbol{\varepsilon}_j$s are mutually independent. Here, k is a fixed positive integer (not depending on p) and $\boldsymbol{\rho}_i$s are mutually orthogonal with

$$||\boldsymbol{\rho}_1||^2 \geq \cdots \geq ||\boldsymbol{\rho}_k||^2 > 0.$$

Note that (A-ii) is met under (1.14). We have that $\boldsymbol{\Sigma} = \sum_{i=1}^k \boldsymbol{\rho}_i \boldsymbol{\rho}_i^T + \tau \boldsymbol{I}_d$ and

$$\lambda_j = ||\boldsymbol{\rho}_j||^2 + \tau \quad \text{for } j = 1, \ldots, k, \quad \text{and} \quad \lambda_{k+1} = \cdots = \lambda_d = \tau. \tag{1.15}$$

In (1.15), the first k eigenvalues are spiked compared to the remaining. Johnstone (2001), Baik and Silverstein (2006), Paul (2007), and Jung and Marron (2009) considered the following spiked model:

$$\lambda_j \ (> \tau) \text{ is fixed (not depending on } d) \text{ for } j = 1, \ldots, k. \tag{1.16}$$

They studied asymptotic behaviors of the conventional principal component analysis (PCA) when $n/d \to c > 0$ under (1.16). However, high-dimensional eigenvalues naturally depend on d and it is probable that $\lambda_j \to \infty$ as $d \to \infty$ for the first several js. See Jung and Marron (2009), Yata and Aoshima (2009), Fan et al. (2013), Ishii et al. (2016), Shen et al. (2016), and Aoshima and Yata (2018) for the details. They considered the following spiked model in which the first k eigenvalues are divergently spiked:

$$\lambda_j = d^{\alpha_j} \quad \text{for } j = 1, \ldots, k. \tag{1.17}$$

Here, α_js are fixed positive constants (not depending on d) preserving the order that $\lambda_1 \geq \cdots \geq \lambda_k$. For such divergently spiked models, Yata and Aoshima (2012b,a) developed new PCA methods. They showed that the new PCAs can enjoy consistency properties both for the eigenvalues and PC directions when $\lambda_j \to \infty$ as $d \to \infty$.

One would be interested in testing the following hypotheses:

$$H_0 : (1.16) \text{ holds} \quad \text{vs.} \quad H_1 : (1.17) \text{ holds.} \tag{1.18}$$

From (1.15) we have that

$$\mathrm{tr}(\boldsymbol{\Sigma}^2) = (d-k)\tau^2 + \sum_{j=1}^k \lambda_j^2 \quad \text{and} \quad \sigma = \frac{(d-k)\tau}{d} + \frac{\sum_{j=1}^k \lambda_j}{d}.$$

Under (1.16), we have that $\Delta = O(1)$ as $d \to \infty$, so that from Lemma 3.2 it holds that

$$\frac{T_n}{2\mathrm{tr}(\boldsymbol{\Sigma}^2)/n} \Rightarrow N(0,1)$$

under (A-iii) since $n\Delta/\mathrm{tr}(\boldsymbol{\Sigma}^2) = O(n/d) \to 0$. Under (1.17), we have that

$$\Delta = \sum_{j=1}^k d^{2\alpha_j}\{1 + o(1)\} \to \infty \quad \text{as } d \to \infty.$$

Thus, for the test of (1.18), one can apply the test procedure (1.7) or (1.13).

Corollary 1.2. *The test procedure (1.7) for (1.18) has (1.8) under (A-iii) and (A-v).*

Corollary 1.3. *The test procedure (1.8) for (1.18) has (1.15) as $d \to \infty$ under (A-vi).*

We note that

$$\Delta \geq d^{2\alpha_1}\{1 + o(1)\} \tag{1.19}$$

as $d \to \infty$ under (1.17). Thus we can consider a lower bound of Δ as $\Delta > d^{2\alpha_1}$. Then, one may set $\Delta_L = d^{2\alpha_{1L}}$ with a prespecified constant $\alpha_{1L} \in (0, 1/2)$ in view of the assumptions that $\Delta_L \to \infty$ and $\Delta_L = o(d)$ as $d \to \infty$.

1.5 Performances

In this section, we consider the detection of divergently spiked noise as an application of the sphericity test. In order to investigate the performance of the test procedure (1.13) for (1.2) or (1.18), we used computer simulations.

1.5.1 Two-stage procedure for sphericity

In order to investigate the performance of the test procedure (1.8) for (1.2) or (1.18), we used computer simulations. We set $\Delta_L = d^{2/3}$. We considered constructing a test having size $\alpha = 0.01$ and power no less than $1 - \beta = 0.8$ when $\Delta \geq \Delta_L$. We set $d = 250,\ 500,\ 1000,\ 2000$ and 4000. We put $d_1 = \lceil (6/5)d^{2/3} \rceil$. The following two cases were considered:

$$(a)\ \boldsymbol{\Sigma} = \boldsymbol{I}_d \quad \text{and} \quad (b)\ \boldsymbol{\Sigma} = \boldsymbol{I}_d + \boldsymbol{G},$$

where $\boldsymbol{G} = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$ whose first d_1 elements are 1. Note that $\Delta = 3d_1 + d - (d_1 + d)^2/d = d_1 - d_1^2/d \geq \Delta_L$ when d is large for (b). Also, note that (A-vi) is met both for (a) and (b). We considered a non-Gaussian case by setting $q = d$, $\boldsymbol{\Gamma} = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}$ and $w_{rj} = (v_{rj} - 5)/10^{1/2}$ in (1.1), where v_{rj}s are i.i.d. as the chi-squared distribution with 5 degrees of freedom. Note that (A-ii) is met. We set $m = \lceil C/2 \rceil$. Note that (1.16) is met both for (a) and (b).

In Tables 1.1 and 1.2, we summarized the findings obtained by averaging the outcomes from 2000 (= $2R$, say) replications, where the first 1000 replications were generated for (a) and the last 1000 replications were generated for (b). Under a fixed scenario, suppose that the rth replication ends with $N = N_r$ observations given by (1.17) and the test result given by (1.8). We defined $P_r = 1$ (or 0) accordingly as H_0 was falsely rejected (or not) and H_1 was falsely rejected (or not). We defined $\overline{\alpha} = R^{-1} \sum_{r=1}^{R} P_r$ to estimate the size and $1 - \overline{\beta} = 1 - R^{-1} \sum_{r=R+1}^{2R} P_r$ to estimate the power when $\Delta \geq \Delta_L$, while their estimated standard errors, $s(\overline{\alpha})$ and $s(\overline{\beta})$, were given by $s^2(\overline{\alpha}) = R^{-1}\overline{\alpha}(1 - \overline{\alpha})$ and $s^2(\overline{\beta}) = R^{-1}\overline{\beta}(1 - \overline{\beta})$. For (a), we calculated $\overline{N} = R^{-1} \sum_{r=1}^{R} N_r$ and $\mathrm{MSE}(N/C) = R^{-1} \sum_{r=1}^{R} (N_r/C - 1)^2$. Similarly, we calculated $\overline{N}$ and $\mathrm{MSE}(N/C)$ for (b).

Table 1.1: Required sample size and average size by the test procedure (1.8) for (1.2) in case of (a).

d	m	C	$\overline{N}$	$\overline{N} - C$	$\mathrm{MSE}(N/C)$	$\overline{\alpha}$	$s(\overline{\alpha})$
250	20	39.91	39.91	-0.01	0.0173	0.017	0.0041
500	26	50.29	50.63	0.34	0.0094	0.014	0.0037
1000	32	63.36	63.65	0.29	0.0052	0.016	0.004
2000	40	79.83	79.95	0.12	0.0033	0.011	0.0033
4000	51	100.58	101.23	0.66	0.0018	0.014	0.0037

We observed that the test procedure (1.8) for (1.2) provides good performances especially when p is large.

1.5.2 Two-stage procedure for detection of divergently spiked noise

We set $\Delta_L = (5/6)d^{3/4}$. We considered constructing a test having size $\alpha = 0.05$ and power no less than $1 - \beta = 0.9$ when $\Delta \geq \Delta_L$. We set $p = 250,\ 500,\ 1000,\ 2000$ and 4000. We handled (1.14) with (1.15).

Table 1.2: Required sample size and average power by the test procedure (1.8) for (1.2) in case of (b).

d	m	C	$\overline{N}$	$\overline{N}-C$	MSE(N/C)	$1-\overline{\beta}$	$s(\overline{\beta})$
250	32	62.9	63.29	0.39	0.0079	0.771	0.0133
500	37	73.22	73.5	0.28	0.0048	0.809	0.0124
1000	44	86.17	86.43	0.26	0.0028	0.835	0.0117
2000	52	102.7	103.21	0.51	0.0021	0.871	0.0106
4000	62	123.43	124	0.56	0.0013	0.897	0.0096

The following two cases were considered:

$$(c)\ \boldsymbol{\Sigma} = \mathrm{diag}(2, 1, 0, \ldots, 0) + \boldsymbol{I}_d \quad \text{and} \quad (d)\ \boldsymbol{\Sigma} = \mathrm{diag}(d^{3/8}, d^{1/4}, 0, \ldots, 0) + \boldsymbol{I}_d$$

for (1.16) and (1.17), respectively. Note that (A-vi) is met both for (c) and (d), and $\Delta = d^{3/4} + d^{1/2} + o(1) \geq \Delta_L$ when d is large for (d). We set $m = \lceil C/2 \rceil$.

Similar to Tables 1.1 and 1.2, we calculated $\overline{\alpha}$, $1-\overline{\beta}$, $s(\overline{\alpha})$, $s(\overline{\beta})$, $\overline{N}$ and MSE(N/C), by 2000 replications. In Tables 1.3 and 1.4, we summarized the results. We observed that the test procedure (1.8) for (1.18) provides good performances especially when d is large.

Table 1.3: Required sample size and average size by the test procedure (1.8) for (1.18) in case of (c).

d	m	C	$\overline{N}$	$\overline{N}-C$	MSE(N/C)	$\overline{\alpha}$	$s(\overline{\alpha})$
250	15	29.16	29.82	0.66	0.031	0.115	0.0101
500	17	33.94	34.38	0.43	0.02	0.072	0.0082
1000	20	39.93	40.52	0.59	0.013	0.056	0.0073
2000	24	47.23	47.9	0.68	0.01	0.06	0.0075
4000	29	56.01	56.67	0.66	0.0062	0.058	0.0074

Throughout the simulations, we observed that the test procedure (1.8) meets the required accuracy successfully.

1.5.3 Real data analysis

We analyzed gene expression data for the test of (1.18). We handled microarray data of Naderi et al. (2007) with 47293 ($= d$) genetic probes. We used the data set of luminal group (84 samples). We set $\alpha = 0.05$ and $\beta = 0.1$. From (1.19) we set $\Delta_L = d^{4/5}$, that is, we designed the test of (1.18) to have size 0.05 and power no less than 0.9 when $\lambda_1 \geq d^{2/5}$. We set $m = 30$. We took the first 30 samples as a pilot sample. We calculated $W_m = 35079$ according to (1.3). From (1.17) the total sample size was calculated as

$$N = \max\left\{30,\ \left\lceil \frac{2(z_\alpha + z_\beta)W_m}{\Delta_L} \right\rceil\right\} = 38.$$

Table 1.4: Required sample size and average power by the test procedure (1.8) for (1.18) in case of (d).

d	m	C	$\overline{N}$	$\overline{N}-C$	MSE(N/C)	$1-\overline{\beta}$	$s(\overline{\beta})$
250	20	39.38	39.65	0.28	0.0385	0.888	0.001
500	22	43.71	44.35	0.63	0.026	0.924	0.0084
1000	25	49.27	49.77	0.51	0.0193	0.914	0.0089
2000	29	56.17	56.57	0.4	0.0119	0.917	0.0087
4000	33	64.61	65.4	0.79	0.0088	0.94	0.0075

Thus we took the next 8 $(= 38 - 30)$ samples. We calculated U_N and T_N according to (1.4) and (1.5). Then, it follows that

$$\frac{NT_N}{2U_N} > z_\alpha \ (= 1.64),$$

so that H_0 in (1.18) was rejected in terms of (1.15). We concluded that $\lambda_j \to \infty$ for the first several js and $\lambda_1^{-1} = O(d^{-2/5})$. Hence, we recommend to use new PCA methods given by Yata and Aoshima (2012b,a) because $\lambda_j \to \infty$ for the first several js.

For instance, Yata and Aoshima (2012a) developed a new PCA called the noise-reduction (NR) methodology. In the NR method, λ_js are estimated by

$$\tilde{\lambda}_j = \hat{\lambda}_j - \frac{\mathrm{tr}(\boldsymbol{S}_n) - \sum_{i=1}^{j} \hat{\lambda}_i}{n-1-j} \quad (j = 1, \ldots, n-2),$$

where $\hat{\lambda}_j$ is the j-th eigenvalue of $\boldsymbol{S}_n$. We note that $\tilde{\lambda}_j$ has a consistency property in the sense that

$$\tilde{\lambda}_j/\lambda_j = 1 + o_P(1) \ \text{ when } \lambda_j \to \infty \text{ as } d \to \infty$$

under some regularity conditions. On the other hand, the conventional estimator, $\hat{\lambda}_j$, includes a large bias in the sense that

$$\hat{\lambda}_j/\lambda_j = 1 + \kappa_j + o_P(1) \ \text{ when } \lambda_j \to \infty \text{ as } d \to \infty, \tag{1.20}$$

where $\kappa_j = \lambda_j^{-1} \sum_{i=k+1}^{d} \lambda_i/(n-1)$. See Yata and Aoshima (2012a) and Aoshima and Yata (2018) for the details. In Table 1.5, we estimated the first five eigenvalues for the data set (38 $(= n)$ samples) both by the NR method and the conventional PCA. We observed that $\hat{\lambda}_j$ is quite large compared with $\tilde{\lambda}_j$ for

Table 1.5: Estimates of the first five eigenvalues by the NR method and conventional PCA together with their ratios for the data set in Naderi et al. (2007).

j	1	2	3	4	5
$\tilde{\lambda}_j$	225.2	120.1	89.3	66	54.3
$\hat{\lambda}_j$	278.8	170.4	137	111.8	98.5
$\hat{\lambda}_j/\tilde{\lambda}_j$	1.238	1.419	1.535	1.694	1.813

all j. This is probably because the bias in (1.20) is quite large for each j. On the other hand, $\tilde{\lambda}_j$ does not depend on the bias under (1.17). Thus, we recommend to use the NR method (or the CDM method by Yata and Aoshima (2012b)) when H_0 in (1.18) is rejected.

Appendix 1

We put $K = 4\mathrm{tr}(\boldsymbol{\Sigma}^2)^2/n^2$ throughout this section. Let $\widehat{\boldsymbol{\Sigma}}_{ij(1)} = n_{(1)}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{(1)(i+j)})(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{(1)(i+j)})^T/(n_{(1)}-1)$, $\widehat{\boldsymbol{\Sigma}}_{ij(2)} = n_{(2)}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{(2)(i+j)})(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{(2)(i+j)})^T/(n_{(2)}-1)$, $A_{ij} = \mathrm{tr}\{(\widehat{\boldsymbol{\Sigma}}_{ij(1)} - \sigma \boldsymbol{I}_d)(\widehat{\boldsymbol{\Sigma}}_{ij(2)} - \sigma \boldsymbol{I}_d)\}$ and $B_{ij} = \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{ij(1)} - \sigma \boldsymbol{I}_d)\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{ij(2)} - \sigma \boldsymbol{I}_d)/d$ for all $i < j$.

Proofs of Lemmas 1.2 and 1.3. We write that

$$\begin{aligned}\varepsilon_{ij} =& A_{ij} + \sigma\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{ij(1)}) + \sigma\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{ij(2)}) - \sigma^2 d \quad \text{and} \\ \zeta_{ij} =& B_{ij} + \sigma\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{ij(1)}) + \sigma\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{ij(2)}) - \sigma^2 d\end{aligned}$$

for all $i < j$. Note that $W_n = 2\sum_{i<j}^n \varepsilon_{ij}/\{n(n-1)\}$ and $U_n = 2\sum_{i<j}^n \zeta_{ij}/\{n(n-1)\}$. Thus, it holds that

$$T_n = 2\sum_{i<j}^n \frac{A_{ij}}{n(n-1)} - 2\sum_{i<j}^n \frac{B_{ij}}{n(n-1)}. \tag{1.21}$$

Here, we can evaluate that

$$\begin{aligned}&\mathrm{Var}\Big(2\sum_{i<j}^n \frac{B_{ij}}{n(n-1)}\Big) = O\Big(\frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2}{d^2 n^2}\Big) = o(K) \quad \text{and} \\ &\mathrm{Var}\Big(2\sum_{i<j}^n \frac{\sigma\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{ij(1)}) + \sigma\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{ij(2)})}{n(n-1)}\Big) = O\Big(\frac{\sigma^2\mathrm{tr}(\boldsymbol{\Sigma}^2)}{n}\Big)\end{aligned} \tag{1.22}$$

under (A-i) and (A-iii). Thus, we conclude the result of Lemma 1.2. On the other hand, from Lemma 5.1 in Yata and Aoshima (2016), we have that

$$\begin{aligned}&\mathrm{Var}\Big(2\sum_{i<j}^n \frac{A_{ij}}{n(n-1)}\Big) \\ &= \Big\{\frac{8\mathrm{tr}\{(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_*)^2\} + 4\sum_{j=1}^q (M_j - 2)(\boldsymbol{\gamma}_j^T \boldsymbol{\Sigma}_* \boldsymbol{\gamma}_j)^2}{n} + K\Big\}\{1 + o(1)\} + O\Big(\frac{\mathrm{tr}(\boldsymbol{\Sigma}^4)}{n^2}\Big)\end{aligned} \tag{1.23}$$

under (A-i) and (A-iii). Then, by noting that $\sum_{j=1}^q (\boldsymbol{\gamma}_j^T \boldsymbol{\Sigma}_* \boldsymbol{\gamma}_j)^2 \le \sum_{j,j'=1}^q (\boldsymbol{\gamma}_j^T \boldsymbol{\Sigma}_* \boldsymbol{\gamma}_{j'})^2 = \mathrm{tr}\{(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_*)^2\}$ and

$$E\Big\{\Big(2\sum_{i<j}^n \frac{A_{ij}}{n(n-1)} - \Delta\Big)\Big(2\sum_{i<j}^n \frac{B_{ij}}{n(n-1)}\Big)\Big\} = o\Big\{\mathrm{Var}\Big(2\sum_{i<j}^n \frac{A_{ij}}{n(n-1)}\Big)^{1/2} K^{1/2}\Big\}$$

under (A-i) and (A-iii), from (1.22) and (1.23), we can conclude the result of Lemma 1.3. □

Proof of Lemma 1.4. Note that $\mathrm{tr}(\boldsymbol{\Sigma}^4) \le \mathrm{tr}(\boldsymbol{\Sigma}^2)^2$ and $\mathrm{tr}\{(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_*)^2\} \le \lambda_1 \mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}\boldsymbol{\Sigma}_*) \le \lambda_1^2 \Delta \le \mathrm{tr}(\boldsymbol{\Sigma}^2)\Delta$. Then, from Lemma 1.3, it holds that

$$\mathrm{Var}(T_n/\Delta) = O\Big\{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2/(n^2\Delta^2) + \mathrm{tr}\{(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_*)^2\}/(n\Delta^2)\Big\} \to 0$$

under (A-i), (A-iii) and (A-iv), so that $T_n/\Delta = 1 + o_P(1)$. It concludes the result. □

Proof of Lemma 1.5. If $\liminf_{d\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}^2)/(\sigma^2 d) > 1$, it holds that $\liminf_{d\to\infty} \Delta/\mathrm{tr}(\boldsymbol{\Sigma}^2) > 0$, so that (A-iv) holds. Thus under (A-v), it holds that as $d \to \infty$

$$\mathrm{tr}(\boldsymbol{\Sigma}^2)/(\sigma^2 d) \to 1. \tag{1.24}$$

If $\liminf_{d\to\infty} \lambda_1^2/(\sigma^2 d) > 0$, it holds that $\liminf_{d\to\infty} \Delta/(\sigma^2 d) > 0$ from the fact that $\sigma d = \sum_{j=1}^d \lambda_j$. Thus, under (A-v) it follows that $\lambda_1^2/\mathrm{tr}(\boldsymbol{\Sigma}^2) \to 0$ as $d \to \infty$, so that

$$\frac{\mathrm{tr}(\boldsymbol{\Sigma}^4)}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2} \leq \frac{\lambda_1^2 \mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2} \to 0 \tag{1.25}$$

as $d \to \infty$ under (A-v). Then, from Corollary 5.2 in Yata and Aoshima (2016), we have that

$$2\sum_{i<j}^n \frac{A_{ij}}{K^{1/2} n(n-1)} \Rightarrow N(0,1)$$

under (A-ii), (A-iii) and (A-v). Thus, from (1.21) and (1.22) we conclude the result. □

Proofs of Theorem 1.1 and Corollary 1.1. First, we consider Corollary 1.1. From Lemmas 1.2 and 1.4 we have that

$$\begin{aligned} P\left(\frac{nT_n}{2U_n} > z_\alpha\right) &= P\left(\frac{T_n}{\Delta} > z_\alpha \frac{2\sigma^2 d\{1+o_P(1)\}}{n\Delta}\right) \\ &= P\left\{1+o_P(1) > o_P(1)\right\} \to 1 \end{aligned}$$

under (A-i), (A-iii) and (A-iv) from the fact that $\sigma^2 d \leq \mathrm{tr}(\boldsymbol{\Sigma}^2)$. It concludes the result of Corollary 1.1.

Next, we consider Theorem 1.1. From Lemmas 1.1, 1.5 and (1.24) we have that

$$\begin{aligned} P\Big(\frac{nT_n}{2U_n} > z_\alpha\Big) &= P\Big(\frac{T_n - \Delta}{2\mathrm{tr}(\boldsymbol{\Sigma}^2)/n} > z_\alpha \frac{U_n}{\mathrm{tr}(\boldsymbol{\Sigma}^2)} - \frac{n\Delta}{2\mathrm{tr}(\boldsymbol{\Sigma}^2)}\Big) \\ &= \Phi\Big(\frac{n\Delta}{2\mathrm{tr}(\boldsymbol{\Sigma}^2)} - z_\alpha\Big) + o(1) \end{aligned} \tag{1.26}$$

under (A-ii), (A-iii) and (A-v). Hence, we conclude the result of Theorem 1.1. The proofs are completed. □

Proof of Theorem 1.2. From Theorem 1.1, the result of Theorem 1.3 is obtained straightforwardly. □

Proof of Theorem 1.3. We assume that $\boldsymbol{\mu} = \mathbf{0}$ without loss of generality. Let $C_L = \lfloor C - (\omega C)^{1/2} \rfloor$ and $C_U = \lceil C + (\omega C)^{1/2} \rceil$, where ω (> 0) is a variable such that $\omega \to 0$ as $d \to \infty$. Under (A-vi), (1.25) holds as $d \to \infty$. Then, from the proof of Theorem 5 in Aoshima and Yata (2014), it holds that under (A-i) and (A-vi)

$$\max\{m, C_L\} \leq N < C_U \tag{1.27}$$

as $d \to \infty$ with probability tending to 1. Let $A_{oij} = \mathrm{tr}\{(\boldsymbol{x}_i\boldsymbol{x}_i^T - \sigma\boldsymbol{I}_d)(\boldsymbol{x}_j\boldsymbol{x}_j^T - \sigma\boldsymbol{I}_d)\}$ and $B_{oij} = \mathrm{tr}(\boldsymbol{x}_i\boldsymbol{x}_i^T - \sigma\boldsymbol{I}_d)\mathrm{tr}(\boldsymbol{x}_j\boldsymbol{x}_j^T - \sigma\boldsymbol{I}_d)/d$ for all $i < j$. Now, we write that

$$T_{oN} = \sum_{i<j}^{C_L} \frac{2(A_{oij} - B_{oij})}{N(N-1)} + \sum_{j=C_L+1}^{N} \sum_{i=1}^{C_L} \frac{2(A_{oij} - B_{oij})}{N(N-1)} + \sum_{i\neq j(>C_L)}^{N} \frac{(A_{oij} - B_{oij})}{N(N-1)}. \tag{1.28}$$

Let $K_C = 4\mathrm{tr}(\mathbf{\Sigma}^2)^2/C_L^2$. By using Chebyshev's inequality and Schwarz's inequality, for any $\eta > 0$, from (1.27), we have that as $d \to \infty$

$$
\begin{aligned}
&P\Big(\Big|\sum_{j=C_L+1}^{N}\sum_{i=1}^{C_L}(A_{oij}-B_{oij}-\Delta)/C^2\Big| > \eta K_C^{1/2}\Big)\\
&\le P\Big(\sum_{j=C_L+1}^{C_U}\Big|\sum_{i=1}^{C_L}(A_{oij}-B_{oij}-\Delta)/C^2\Big| > \eta K_C^{1/2}\Big)\\
&= O\big\{\omega(\mathrm{tr}(\mathbf{\Sigma}^2)^2+\mathrm{tr}\{(\mathbf{\Sigma}\mathbf{\Sigma}_*)^2\})/(C^2K_C)\big\} \to 0 \ \text{ and}\\
&P\Big(\Big|\sum_{i\neq j(>C_L)}^{N}(A_{oij}-B_{oij}-\Delta)/C^2\Big| > \eta K_C^{1/2}\Big)\\
&\le P\Big(\sum_{i\neq j(>C_L)}^{C_U}\Big|(A_{oij}-B_{oij}-\Delta)/C^2\Big| > \eta K_C^{1/2}\Big) = O\big\{\omega^2(\mathrm{tr}(\mathbf{\Sigma}^2)^2+\mathrm{tr}\{(\mathbf{\Sigma}\mathbf{\Sigma}_*)^2\})/(C^2K_C)\big\} \to 0
\end{aligned}
$$

under (A-ii) and (A-vi) from the fact that $\mathrm{tr}\{(\mathbf{\Sigma}\mathbf{\Sigma}_*)^2\} = O\{\mathrm{tr}(\mathbf{\Sigma}^2)^2\}$. Thus, from (1.28) and Lemma 1.5, we have that

$$
\frac{T_{oN}-\Delta}{K_C^{1/2}} = \sum_{i<j}^{C_L}\frac{2(A_{oij}-B_{oij}-\Delta)}{K_C^{1/2}N(N-1)} + o_P(1) = \sum_{i<j}^{C_L}\frac{2(A_{oij}-B_{oij}-\Delta)}{K_C^{1/2}C_L(C_L-1)} + o_P(1) \Rightarrow N(0,1) \tag{1.29}
$$

under (A-ii) and (A-vi) from the fact that

$$
T_{C_L} - \Delta = \sum_{i<j}^{C_L}\frac{2(A_{oij}-B_{oij}-\Delta)}{C_L(C_L-1)} + o_P(K_C^{1/2}).
$$

Here, in a way similar to the proof of Lemma A.5 in Yata and Aoshima (2013), we have that

$$
T_N = T_{oN} + o_P(K_C^{1/2}) \tag{1.30}
$$

under (A-ii) and (A-vi). By combining (1.29) with (1.30), we conclude the result. □

Proofs of Corollaries 1.2 and 1.3. Under (1.14) and (1.16), it holds that $n\Delta/\mathrm{tr}(\mathbf{\Sigma}^2) \to 0$ under (A-iii) and $C\Delta/\mathrm{tr}(\mathbf{\Sigma}^2) = O(1/\Delta_L) \to 0$ as $d \to \infty$. Then, from Theorems 1.1 and 1.3, we conclude the results. □

Chapter 2

Hard-margin linear SVM in the HDLSS context

In this chapter, we consider the binary classification by the hmLSVM in the HDLSS context. This chapter is organized by Nakayama et al. (2017).

The classification problem for high-dimensional data has been studied in a wide range of fields. Aoshima and Yata (2019b) considered quadratic classifiers in general and discussed asymptotic properties and optimality of the classifies under high-dimension, non-sparse settings. According to them, linear classifiers give a preferable performance under the non-sparsity. Also, such non-sparse situations often occur in high-dimensional settings. See Aoshima and Yata (2019b) for the details. Hence, in this chapter, we focus on linear classifiers. In the field of machine learning, there are many studies about the classification in the context of supervised learning. The LSVM is a typical linear classifier and has versatility and effectiveness both for low-dimensional and high-dimensional data. Hall et al. (2005) and Qiao and Zhang (2015) investigated the LSVM for HDLSS data because the training data sets are linearly separable. In this chapter, we also consider asymptotic properties of the LSVM under non-sparsity of mean vectors.

In Section 2.2, we show that the LSVM holds a consistency property under certain severe conditions in the sense that misclassification rates go to 0. We show that the LSVM is very biased in HDLSS settings and its performance is affected by the bias directly.

In order to overcome such difficulties, we propose a bias-corrected LSVM (BC-LSVM) in Section 2.3. We show that the BC-LSVM improves the LSVM even when n_is or $\mathbf{\Sigma}_i$s are unbalanced.

In Section 2.4, we check the performance of the BC-LSVM by numerical simulations and use it in real data analyses.

In Section 2.5, we discuss multiclass LSVMs in HDLSS settings.

2.1 Introduction

Suppose we have independent and d-variate two populations, $\pi_i,\ i = 1, 2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i$ ($\geq \boldsymbol{O}$) for each i. In Section 2.5, we consider a multiclass cases. We assume that $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0, \infty)$ as $d \to \infty$ for $i = 1, 2$. Here, for a function, $f(\cdot)$, "$f(d) \in (0, \infty)$ as $d \to \infty$" implies $\liminf_{d\to\infty} f(d) > 0$ and $\limsup_{d\to\infty} f(d) < \infty$. Let $\Delta_\mu = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, where $\|\cdot\|$ denotes the Euclidean norm. We assume that $\limsup_{d\to\infty} \Delta_\mu/d < \infty$. We have independent and identically distributed (i.i.d.) observations, $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$, from each π_i. We assume $n_i \geq 2,\ i = 1, 2$. Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to one of the two populations. Let $n = n_1 + n_2$.

Now, let us use the following toy examples to see the performance of the hard-margin linear support vector machine (hmLSVM) given by (2.5). We set $n = 20$ and $d = 2^s,\ s = 5, \ldots, 11$. Independent pseudo random observations were generated from $\pi_i : N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. We set $\boldsymbol{\mu}_1 = \boldsymbol{0}$ and $\boldsymbol{\mu}_2 = (1/3, \ldots, 1/3)^T$, so that $\Delta_\mu = d/9$. We considered three cases:

(a) $(n_1, n_2) = (10, 10)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_d$;
(b) $(n_1, n_2) = (6, 14)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_d$; and
(c) $(n_1, n_2) = (10, 10)$, $\boldsymbol{\Sigma}_1 = 0.6\boldsymbol{I}_d$ and $\boldsymbol{\Sigma}_2 = 1.4\boldsymbol{I}_d$,

where $\boldsymbol{I}_d$ denotes the d-dimensional identity matrix. Note that $\Delta_\mu > |\mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 - \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2|$ for (a) to (c). Then, from Theorem 1 in Hall et al. (2005), the classifier holds

$$e(i) \to 0 \quad \text{as } d \to \infty \text{ for } i = 1, 2 \tag{2.1}$$

for (a) to (c), where $e(i)$ denotes the error rate of misclassifying an individual from π_i into the other class. We repeated 2000 times to confirm if the classifier does (or does not) classify $\boldsymbol{x}_0 \in \pi_i$ correctly and defined $P_{ir} = 0$ (or 1) accordingly for each π_i ($i = 1, 2$). We calculated the error rates, $\overline{e}(i) = \sum_{r=1}^{2000} P_{ir}/2000$, $i = 1, 2$. Also, we calculated the average error rate, $\overline{e} = \{\overline{e}(1) + \overline{e}(2)\}/2$. Their standard deviations are less than 0.011. In Figure 2.1, we plotted $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ for (a) to (c). We observed that the LSVM gives a good performance as d increases for (a). Contrary to expectations, it led undesirable performances both for (b) and (c). The error rates became small as d increases, however, $\overline{e}(1)$ and $\overline{e}(2)$ became quite unbalanced. We will discuss its theoretical reasons in Section 2.2.2.

2.2 Linear SVM in HDLSS settings

In this section, we give asymptotic properties of the LSVM when $d \to \infty$. Since HDLSS data are linearly separable by a hyperplane, we consider the LSVM.

2.2.1 Setup of linear SVM

First, we introduce the LSVM. We consider the following linear classifier:

$$y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b, \tag{2.2}$$

where $\boldsymbol{w}$ is a weight vector and b is an intercept term. Let us write that $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}, \boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2})$. Let $t_j = -1$ for $j = 1, \ldots, n_1$ and $t_j = 1$ for $j = n_1 + 1, \ldots, n$. The LSVM is defined by maximizing the smallest distance of all observations to the separating hyperplane. The optimization problem of the LSVM can be written as follows:

$$\underset{\boldsymbol{w},b}{\mathrm{argmin}}\, \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{subject to} \quad t_j(\boldsymbol{w}^T\boldsymbol{x}_j + b) \geq 1,\ j = 1, \ldots, n.$$

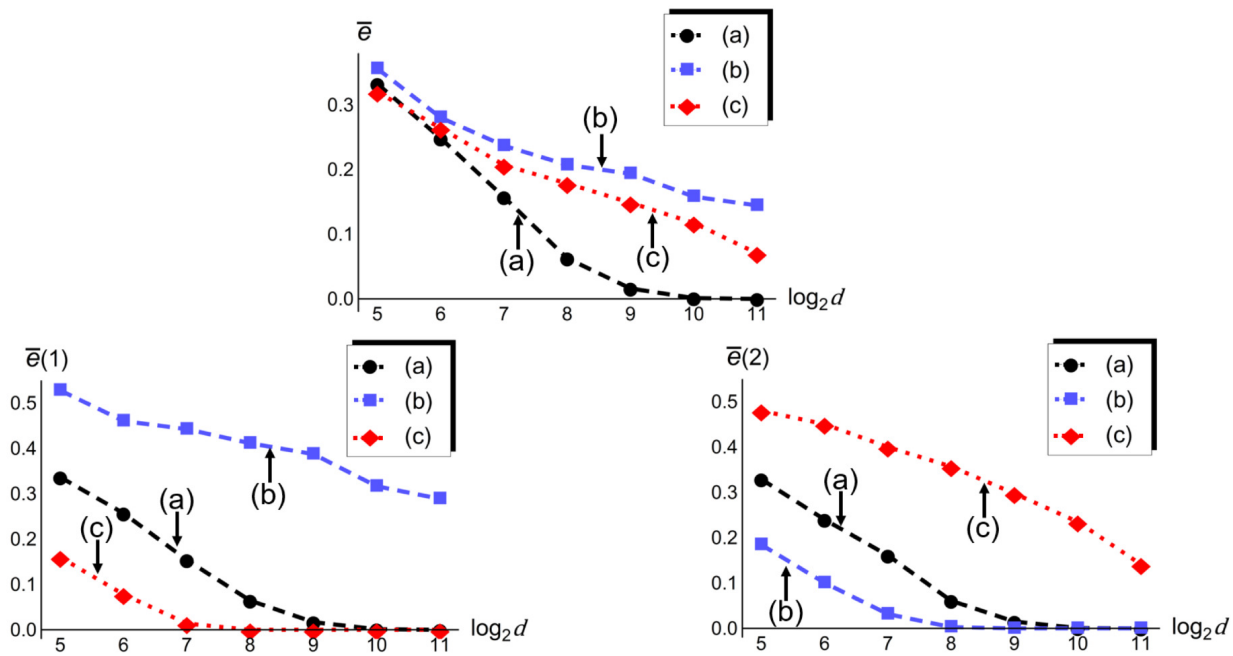


Figure 2.1: The performance of the LSVM given by (2.5) in HDLSS settings. The left panel displays $\overline{e}(1)$, the right panel displays $\overline{e}(2)$ and the top panel displays $\overline{e}$.

A Lagrangian formulation is given by

$$L(\boldsymbol{w}, b; \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{j=1}^{n} \alpha_j \{t_j(\boldsymbol{w}^T \boldsymbol{x}_j + b) - 1\},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ and α_js are Lagrange multipliers. By differentiating the Lagrangian formulation with respect to $\boldsymbol{w}$ and b, we obtain the following conditions:

$$\boldsymbol{w} = \sum_{j=1}^{n} \alpha_j t_j \boldsymbol{x}_j \quad \text{and} \quad \sum_{j=1}^{n} \alpha_j t_j = 0.$$

After substituting the two conditions into $L(\boldsymbol{w}, b; \boldsymbol{\alpha})$, we obtain the dual form:

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^{n} \alpha_j - \frac{1}{2} \sum_{j=1}^{n} \sum_{j'=1}^{n} \alpha_j \alpha_{j'} t_j t_{j'} \boldsymbol{x}_j^T \boldsymbol{x}_{j'}. \tag{2.3}$$

The optimization problem can be transformed into the following:

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L(\boldsymbol{\alpha})$$

subject to

$$\alpha_j \geq 0, \;\; j = 1, \ldots, n, \quad \text{and} \quad \sum_{j=1}^{n} \alpha_j t_j = 0. \tag{2.4}$$

Let us write that

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)^T = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L(\boldsymbol{\alpha}) \;\; \text{subject to (2.4)}.$$

There exist some $\boldsymbol{x}_j$s satisfying that $t_j y(\boldsymbol{x}_j) = 1$ (i.e., $\hat{\alpha}_j \neq 0$). Such $\boldsymbol{x}_j$s are called the support vector. Let $\hat{S} = \{j | \hat{\alpha}_j \neq 0,\ j = 1, \ldots, n\}$ and $N_{\hat{S}} = \#\hat{S}$, where $\#A$ denotes the number of elements in a set A. The intercept term is given by

$$\hat{b} = \frac{1}{n_{\hat{S}}} \sum_{j \in \hat{S}} \Big(t_j - \sum_{j' \in \hat{S}} \hat{\alpha}_{j'} t_{j'} \boldsymbol{x}_j^T \boldsymbol{x}_{j'} \Big).$$

Then, the linear classifier in (2.2) is defined by

$$\hat{y}(\boldsymbol{x}) = \sum_{j \in \hat{S}} \hat{\alpha}_j t_j \boldsymbol{x}_j^T \boldsymbol{x} + \hat{b}. \tag{2.5}$$

Finally, in the LSVM, one classifies $\boldsymbol{x}_0$ into π_1 if $\hat{y}(\boldsymbol{x}_0) < 0$ and into π_2 otherwise. See Vapnik (2000) for the details.

2.2.2 Asymptotic properties of linear SVM when d tends to infinity

In this section, we give asymptotic properties of the LSVM in the HDLSS context. We assume the following assumptions:

(A-i) $\dfrac{\mathrm{Var}(\|\boldsymbol{x}_{ik} - \boldsymbol{\mu}_i\|^2)}{\Delta_\mu^2} \to 0$ as $d \to \infty$ for $i = 1, 2$;

(A-ii) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{\Delta_\mu^2} \to 0$ as $d \to \infty$ for $i = 1, 2$.

Note that $\mathrm{Var}(\|\boldsymbol{x}_{ik} - \boldsymbol{\mu}_i\|^2) = 2\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ when π_i is Gaussian, so that (A-i) and (A-ii) are equivalent when π_is are Gaussian. We have the following result.

Lemma 2.1. *Assume (A-i) and (A-ii). Under (2.4), it holds that as $d \to \infty$*

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^n \alpha_j - \frac{\Delta_\mu}{8} \Big(\sum_{j=1}^n \alpha_j \Big)^2 \{1 + o_P(1)\} - \frac{1}{2} \Big(\mathrm{tr}(\boldsymbol{\Sigma}_1) \sum_{j=1}^{n_1} \alpha_j^2 + \mathrm{tr}(\boldsymbol{\Sigma}_2) \sum_{j=n_1+1}^n \alpha_j^2 \Big).$$

Let $\Delta_{\mu*} = \Delta_\mu + \mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2$. Under the constraint that $\sum_{j=1}^{n_1} \alpha_j = \sum_{j=n_1+1}^n \alpha_j$ ($= \alpha_\star$, say), we can claim that

$$\max_{\boldsymbol{\alpha}} \Big\{ -\frac{1}{2} \Big(\mathrm{tr}(\boldsymbol{\Sigma}_1) \sum_{j=1}^{n_1} \alpha_j^2 + \mathrm{tr}(\boldsymbol{\Sigma}_2) \sum_{j=n_1+1}^n \alpha_j^2 \Big) \Big\} = -\alpha_\star^2 \{ \mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2 \} \tag{2.6}$$

when $\alpha_j = \alpha_\star/(2n_1),\ j = 1, \ldots, n_1$ and $\alpha_j = \alpha_\star/(2n_2),\ j = n_1 + 1, \ldots, n$ under (2.4). Then, from Lemma 2.1 it holds that

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = -\frac{\Delta_{\mu*}}{2} \Big(\alpha_\star - \frac{2 + o_P(1)}{\Delta_{\mu*}} \Big)^2 \{1 + o_P(1)\} + \frac{2 + o_P(1)}{\Delta_{\mu*}} \tag{2.7}$$

for given $\alpha_\star (> 0)$. Let $\delta_\Sigma = \mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 - \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2$ Hence, by choosing $\alpha_\star \approx 2/\Delta_{\mu*}$, we have the maximum of $L(\boldsymbol{\alpha})$ asymptotically. Then, we have the following result.

Lemma 2.2. *Assume (A-i) and (A-ii). It holds that as $d \to \infty$*

$$\hat{\alpha}_j = \frac{2}{\Delta_{\mu*} n_1}\{1 + o_p(1)\} \quad \textit{for } j = 1, \ldots, n_1; \quad \textit{and}$$
$$\hat{\alpha}_j = \frac{2}{\Delta_{\mu*} n_2}\{1 + o_p(1)\} \quad \textit{for } j = n_1 + 1, \ldots, n.$$

Furthermore, it holds that as $d \to \infty$

$$\hat{y}(\boldsymbol{x}_0) = \frac{(-1)^i \Delta_\mu}{\Delta_{\mu*}}\{1 + o_P(1)\} + \frac{\delta_\Sigma}{\Delta_{\mu*}}, \quad \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2.$$

Remark 3. *From Lemma 2.2, all the data points are the support vectors under (A-i) and (A-ii) in the HDLSS context. Ahn and Marron (2010) called this phenomenon the "data piling". See Sections 1 and 2 in Ahn and Marron (2010) for the details.*

From Lemma 2.2, under (A-i) and (A-ii), it holds that as $d \to \infty$

$$\frac{\Delta_{\mu*}}{\Delta_\mu}\hat{y}(\boldsymbol{x}_0) = (-1)^i + \frac{\delta_\Sigma}{\Delta_\mu} + o_P(1) \tag{2.8}$$

when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$. Hence, "δ_Σ/Δ_μ" is the bias term of the (normalized) LSVM. We assume the following assumption:

(A-iii) $\displaystyle \limsup_{d\to\infty} \frac{|\delta_\Sigma|}{\Delta_\mu} < 1.$

Then, we have the following results.

Theorem 2.1. *Assume (A-i) to (A-iii). For the LSVM we have (2.1).*

Corollary 2.1. *Assume (A-i) and (A-ii). For the LSVM we have the following inconsistent properties:*

$$e(1) \to 1 \quad \textit{and} \quad e(2) \to 0 \quad \textit{as } d \to \infty \quad \textit{if} \quad \liminf_{d\to\infty} \frac{\delta_\Sigma}{\Delta_\mu} > 1; \quad \textit{and}$$
$$e(1) \to 0 \quad \textit{and} \quad e(2) \to 1 \quad \textit{as } d \to \infty \quad \textit{if} \quad \limsup_{d\to\infty} \frac{\delta_\Sigma}{\Delta_\mu} < -1.$$

Remark 4. *For the LSVM, Hall et al. (2005) and Qiao and Zhang (2015) also showed (2.1) and the inconsistent properties in Corollary 2.1 under different conditions. We emphasize that (A-i), (A-ii) and (A-iii) are milder than their conditions. Moreover, we derived the bias term in the LSVM.*

From (2.8), we conclude for sufficiently large d that $e(1)$ and $e(2)$ for the LSVM become small but unbalanced if $0 < |\delta_\Sigma|/\Delta_\mu < 1$. If $\delta_\Sigma/\Delta_\mu > 0$ (or $\delta_\Sigma/\Delta_\mu < 0$), $e(1)$ (or $e(2)$) is larger than $e(2)$ (or $e(1)$). Actually, in Figure 2.1, we observed that $\overline{e}(1)$ is larger than $\overline{e}(2)$ for each d in the case of (b) $(n_1, n_2) = (6, 14)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_d$ (i.e., $\delta_\Sigma/\Delta_\mu = 6/7$); on the other hand, $\overline{e}(2)$ is larger than $\overline{e}(1)$ for each d in the case of (c) $(n_1, n_2) = (10, 10)$, $\boldsymbol{\Sigma}_1 = 0.6\boldsymbol{I}_d$ and $\boldsymbol{\Sigma}_2 = 1.4\boldsymbol{I}_d$ (i.e., $\delta_\Sigma/\Delta_\mu = -18/25$). As for (a) $(n_1, n_2) = (10, 10)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_d$ (i.e., $\delta_\Sigma = 0$), the LSVM gave a preferable performance.

2.2.3 Asymptotic properties of linear SVM when both d and n tend to infinity

In this section, we give asymptotic properties of the LSVM when both $d, n \to \infty$ while $n/d \to 0$. One may consider $n = O(\log d)$ for example. We assume the following assumptions:

(A-i') $\dfrac{n\mathrm{Var}(\|\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i\|^2)}{\Delta_\mu^2} \to 0$ as $d, n \to \infty$ for $i = 1, 2$;

(A-ii') $\dfrac{n^2 \mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{\Delta_\mu^2} \to 0$ as $d, n \to \infty$ for $i = 1, 2$;

(A-iv) $\displaystyle\liminf_{d,n\to\infty} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i)}{\Delta_\mu n_i} > 0$ for $i = 1, 2$.

Note that $\Delta_\mu^2/\mathrm{tr}(\boldsymbol{\Sigma}_i^2) = O(d)$ from the facts that $\limsup_{d\to\infty} \Delta_\mu/d < \infty$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0, \infty)$ as $d \to \infty$ for $i = 1, 2$. Thus, $n = o(d^{1/2})$ when (A-ii') is met.

Lemma 2.3. *Under (A-i'), (A-ii') and (A-iv), it holds that as $d, n \to \infty$*

$$\hat{y}(\boldsymbol{x}_0) = \frac{(-1)^i \Delta_\mu}{\Delta_{\mu*}} + \frac{\delta_\Sigma}{\Delta_{\mu*}} + o_P\Big(\frac{\Delta_\mu}{\Delta_{\mu*}}\Big) \quad \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2.$$

Corollary 2.2. *Under (A-i'), (A-ii') and (A-iv), the LSVM holds the following properties:*

$$e(1) \to 0 \ \textit{ and } \ e(2) \to 0 \ \textit{ as } d, n \to \infty \ \textit{ if } \ \limsup_{d,n\to\infty} \frac{|\delta_\Sigma|}{\Delta_\mu} < 1;$$

$$e(1) \to 1 \ \textit{ and } \ e(2) \to 0 \ \textit{ as } d, n \to \infty \ \textit{ if } \ \liminf_{d,n\to\infty} \frac{\delta_\Sigma}{\Delta_\mu} > 1; \quad \textit{and}$$

$$e(1) \to 0 \ \textit{ and } \ e(2) \to 1 \ \textit{ as } d, n \to \infty \ \textit{ if } \ \limsup_{d,n\to\infty} \frac{\delta_\Sigma}{\Delta_\mu} < -1.$$

2.3 Bias-corrected linear SVM

From the argument in Section 2.2.2, if $\liminf_{d\to\infty} |\delta_\Sigma|/\Delta_\mu > 0$, the LSVM gives an undesirable performance. In addition, from Corollary 2.1, if $\liminf_{d\to\infty} |\delta_\Sigma|/\Delta_\mu > 1$, one should not use the LSVM. In order to overcome such difficulties, we propose a bias-corrected LSVM in this section.

We estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ by $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})^T/(n_i - 1)$. We estimate $\Delta_{\mu*}$ by $\widehat{\Delta}_{\mu*} = \|\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}\|^2$. Note that $E(\widehat{\Delta}_{\mu*}) = \Delta_{\mu*}$. Let $\hat{\delta}_\Sigma = \mathrm{tr}(\boldsymbol{S}_{1n_1})/n_1 - \mathrm{tr}(\boldsymbol{S}_{2n_2})/n_2$. Note that $E(\hat{\delta}_\Sigma) = \delta_\Sigma$. Then, we have the following result.

Lemma 2.4. *Assume (A-i) and (A-ii). It holds that as $d \to \infty$*

$$\frac{\hat{\delta}_\Sigma}{\widehat{\Delta}_{\mu*}} = \frac{\delta_\Sigma}{\Delta_{\mu*}} + o_P\Big(\frac{\Delta_\mu}{\Delta_{\mu*}}\Big).$$

Now, we define the bias-corrected LSVM (BC-LSVM) by

$$\hat{y}_{BC}(\boldsymbol{x}_0) = \hat{y}(\boldsymbol{x}_0) - \frac{\hat{\delta}_\kappa}{\widehat{\Delta}_{\mu*}}, \tag{2.9}$$

where $\hat{y}(\boldsymbol{x}_0)$ is given by (2.5). In the BC-LSVM, one classifies $\boldsymbol{x}_0$ into π_1 if $\hat{y}_{BC}(\boldsymbol{x}_0) < 0$ and into π_2 otherwise.

By combining (2.8) with Lemma 2.4, under (A-i) and (A-ii), it holds that as $d \to \infty$

$$\frac{\Delta_{\mu*}}{\Delta_\mu}\hat{y}_{BC}(\boldsymbol{x}_0) = (-1)^i + o_P(1) \tag{2.10}$$

when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$. Hence, we have the following result.

Theorem 2.2. *Assume (A-i) and (A-ii). For the BC-LSVM, we have (2.1).*

Remark 5. *The BC-LSVM has the consistency property without assuming (A-iii). Chan and Hall (2009) considered a different bias correction for the LSVM. They showed the consistency property under stricter conditions than (A-i) and (A-ii).*

Remark 6. *Aoshima and Yata (2014) considered the distance-based classifier as follows: One classifies an individual into π_1 if $y_{AY}(\boldsymbol{x}_0) < 0$ and into π_2 otherwise, where $y_{AY}(\boldsymbol{x}_0) = \{\boldsymbol{x}_0 - (\overline{\boldsymbol{x}}_{1n_1} + \overline{\boldsymbol{x}}_{2n_2})/2\}^T(\overline{\boldsymbol{x}}_{2n_2} - \overline{\boldsymbol{x}}_{1n_1}) - \mathrm{tr}(\boldsymbol{S}_{1n_1})/(2n_1) + \mathrm{tr}(\boldsymbol{S}_{2n_2})/(2n_2)$. Then, from Theorem 1 in Aoshima and Yata (2014), under (A-ii), it holds that as $d \to \infty$*

$$(2/\Delta_\mu)y_{AY}(\boldsymbol{x}_0) = (-1)^i + o_P(1)$$

when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$.

When both $d, n \to \infty$, we have the following result.

Corollary 2.3. *Under (A-i'), (A-ii') and (A-iv), it holds for the BC-LSVM that $e(i) \to 0$ as $d, n \to \infty$ for $i = 1, 2$.*

2.4 Performances

In this section, we check the performance of the BC-LSVM in numerical simulations and real data analyses.

2.4.1 Simulations

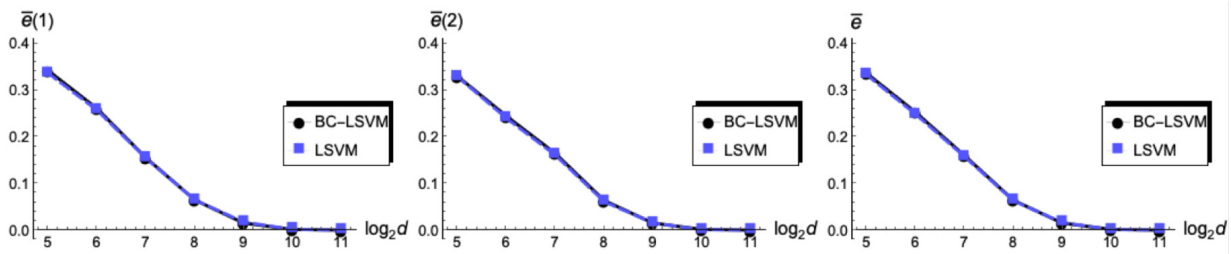
First, we checked the performance of the BC-LSVM by using the toy examples in Figure 2.1. Similar to Section 2.1, we calculated the error rates, $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$, by 2000 replications and plotted the results in Figure 2.2. We laid $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ for the LSVM by borrowing from Figure 2.1. As expected theoretically, we observed that the BC-LSVM gives preferable performances even for (b) and (c) in which $\liminf_{d\to\infty} |\delta_\Sigma|/\Delta_\mu > 0$.

Next, we compared the performance of the BC-LSVM with the LSVM in complex settings. We set $\boldsymbol{\mu}_1 = \boldsymbol{0}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{B}(0.4^{|i-j|^{1/3}})\boldsymbol{B}$, where

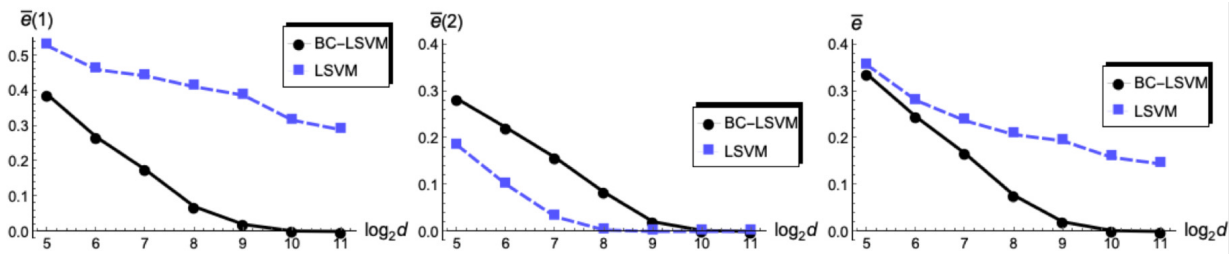
$$\boldsymbol{B} = \mathrm{diag}[\{0.5 + 1/(d+1)\}^{1/2}, \ldots, \{0.5 + d/(d+1)\}^{1/2}].$$

Note that $\mathrm{tr}(\boldsymbol{\Sigma}_1) = \mathrm{tr}(\boldsymbol{\Sigma}_2) = d$. We considered two cases:

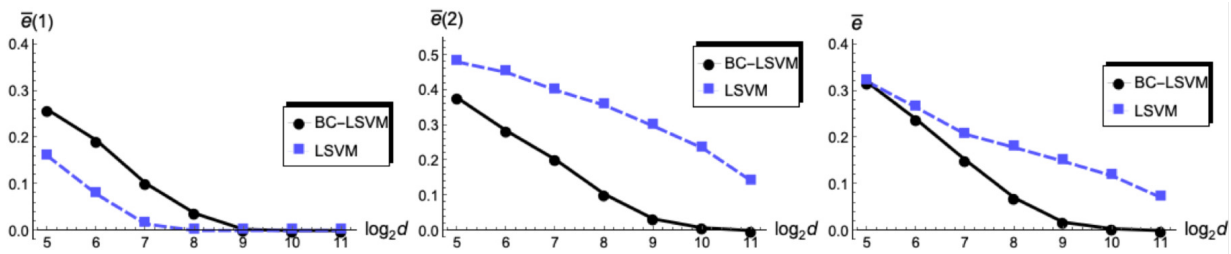
$\boldsymbol{\mu}_2 = (1, \ldots, 1, 0, \ldots, 0, -1, \ldots, -1)^T$ $(= \boldsymbol{\mu}_\alpha(t),$ say) whose first $t/2$ elements are 1 and last $t/2$ elements are -1 for a positive even number t; and
$\boldsymbol{\mu}_2 = (t^{1/2}/2, t^{1/2}/2, 0, \ldots, 0, -t^{1/2}/2, -t^{1/2}/2)^T$ $(= \boldsymbol{\mu}_\beta(t),$ say) whose first two elements are $t^{1/2}/2$ and last two elements are $-t^{1/2}/2$ for a positive number t.



(a) $(n_1, n_2) = (10, 10)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_d$ (i.e., $\delta_\Sigma = 0$)



(b) $(n_1, n_2) = (6, 14)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_d$ (i.e., $\delta_\Sigma / \Delta_\mu = 6/7$)



(c) $(n_1, n_2) = (10, 10)$, $\boldsymbol{\Sigma}_1 = 0.6\boldsymbol{I}_d$ and $\boldsymbol{\Sigma}_2 = 1.4\boldsymbol{I}_d$ (i.e., $\delta_\Sigma / \Delta_\mu = -18/25$)

Figure 2.2: The performance of the BC-LSVM in HDLSS settings. The error rates are denoted by the solid lines for (a), (b) and (c). The left panels display $\overline{e}(1)$, the middle panels display $\overline{e}(2)$ and the right panels display $\overline{e}$. The corresponding error rates by the LSVM are denoted by the dashed lines.

Note that $\Delta_\mu = t$ both for $\boldsymbol{\mu}_\alpha(t)$ and $\boldsymbol{\mu}_\beta(t)$. We generated $\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i$, $i = 1, 2;\ j = 1, 2, \ldots,$ independently either from (I) $N_d(\mathbf{0}, \boldsymbol{\Sigma}_i),\ i = 1, 2$, or (II) a d-variate t-distribution, $t_d(\boldsymbol{\Sigma}_i, 10),\ i = 1, 2$, with mean zero, covariance matrix $\boldsymbol{\Sigma}_i$ and degrees of freedom 10. Note that (A-i) holds under (A-ii) for (I). Let $d_* = 2\lceil d^{2/3}/2 \rceil$. We considered four cases:

(d) $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_\alpha(d_*)$, $(n_1, n_2) = (5, 25)$ and $d = 2^s,\ s = 6, \ldots, 12$, for (I);
(e) $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_\alpha(d_*)$, $d = 1000$ and $(n_1, n_2) = (4s, 8s),\ s = 1, \ldots, 7$, for (II);
(f) $d = 1000$, $(n_1, n_2) = (10, 20)$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_\alpha(2^s),\ s = 1, \ldots, 7$, for (II); and
(g) $d = 1000$, $(n_1, n_2) = (10, 20)$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_\beta(2^s),\ s = 1, \ldots, 7$, for (II).

Note that $\Delta_\mu = d_* = o(d)$ and (A-ii) holds for (d) and (e) from the fact that $\mathrm{tr}(\boldsymbol{\Sigma}_i^2) = O(d)$, $i = 1, 2$. Also, note that (A-i) holds for (d). However, (A-i) does not hold for (e) and (A-iii) does not hold both for (d) and (e). For (f) and (g), we note that $\Delta_\mu = 2^s,\ s = 1, \ldots, 7$. Especially, (g) is a sparse case such that the only four elements of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ are nonzero. Similar to Section 2.1, we calculated the error rates, $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$, by 2000 replications and plotted the results in Figure 2.3.

We observe that the LSVM gives quite bad performances for (d) in Figure 2.3. The main reason must be due to the bias term in the LSVM. Note that $\delta_\Sigma/\Delta_\mu \to \infty$ as $d \to \infty$ for (d). Thus $\overline{e}(1)$ becomes close to 1 as d increases. See Corollary 2.1 for the details. Also, the LSVM gives bad performances for (e) to (g) when n_is are small or Δ_μ is small. This is because δ_Σ/Δ_μ becomes large when n_is are small or Δ_μ is small. On the other hand, from Figures 2.2 and 2.3, the BC-LSVM gives adequate performances even when n_is and $\boldsymbol{\Sigma}_i$s are unbalanced. The BC-LSVM also gives a better performance than the LSVM even when Δ_μ is small (or sparse).

2.4.2 Real data analyses

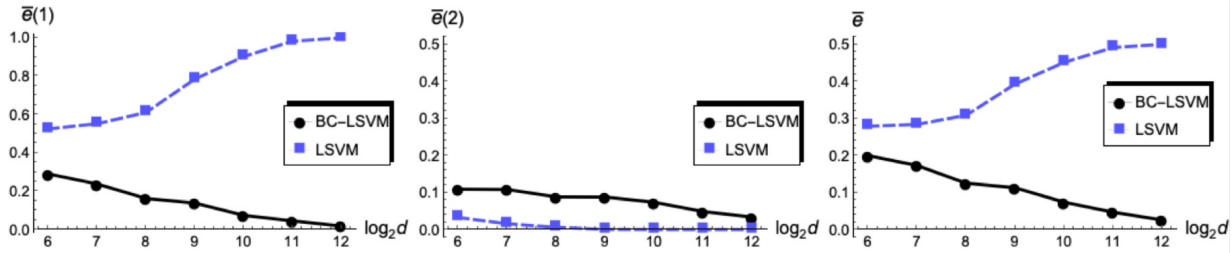
First, we used colon cancer data with 2000 ($= d$) genes given by Alon et al. (1999) which consists of π_1 : colon tumor (40 samples) and π_2 : normal colon (22 samples). We set $n_1 = n_2 = 10$. We randomly split the data sets from (π_1, π_2) into training data sets of sizes (n_1, n_2) and test data sets of sizes $(40-n_1, 22-n_2)$. We constructed the BC-LSVM and the LSVM by using the training data sets. We checked accuracy by using the test data set for each π_i and denoted the misclassification rates by $\widehat{e}(1)_r$ and $\widehat{e}(2)_r$. We repeated this procedure 100 times and obtained $\widehat{e}(1)_r$ and $\widehat{e}(2)_r$, $r = 1, \ldots, 100$, both for the BC-LSVM and the LSVM. We had the average misclassification rates as $\overline{e}(1)$ $(= \sum_{r=1}^{100} \widehat{e}(1)_r/100) = 0.16$, $\overline{e}(2)$ $(= \sum_{r=1}^{100} \widehat{e}(2)_r/100) = 0.166$ and $\overline{e}$ $(= \{\overline{e}(1) + \overline{e}(2)\}/2) = 0.163$ for the BC-LSVM, and $\overline{e}(1) = 0.158$, $\overline{e}(2) = 0.161$ and $\overline{e} = 0.159$ for the LSVM. By using all the samples, we considered estimating δ_Σ/Δ_μ. We set $m_1 = 40$ and $m_2 = 22$. From Section 3.1 in Aoshima and Yata (2011), an unbiased estimator of Δ_μ was given by $\widehat{\Delta}_{\mu(m)} = \|\overline{\boldsymbol{x}}_{1m_1} - \overline{\boldsymbol{x}}_{2m_2}\|^2 - \mathrm{tr}(\boldsymbol{S}_{1m_1})/m_1 - \mathrm{tr}(\boldsymbol{S}_{2m_2})/m_2$. We estimated δ_Σ/Δ_μ by

$$\hat{\delta}_\Sigma/\widehat{\Delta}_\mu = \{\mathrm{tr}(\boldsymbol{S}_{1m_1})/n_1 - \mathrm{tr}(\boldsymbol{S}_{2m_2})/n_2\}/\widehat{\Delta}_{\mu(m)}$$

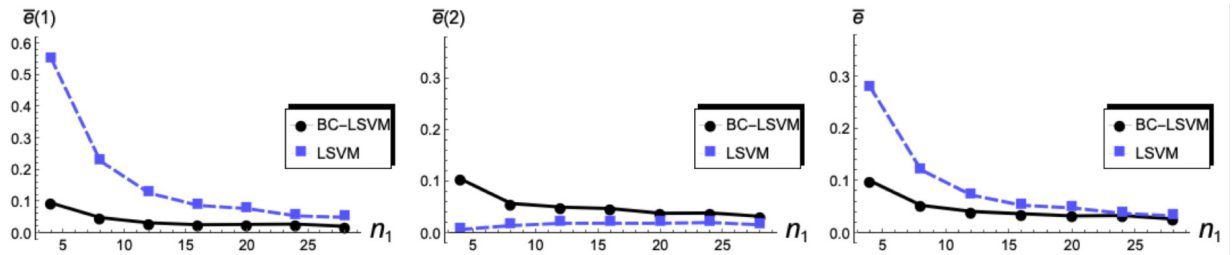
and had $\hat{\delta}_\Sigma/\widehat{\Delta}_\mu = 0.003$ for the 62 samples. In view of (2.9), we expect that the BC-LSVM is asymptotically equivalent to the LSVM in such cases. We estimated $(\mathrm{tr}(\boldsymbol{\Sigma}_1)/\Delta_\mu, \mathrm{tr}(\boldsymbol{\Sigma}_2)/\Delta_\mu)$ by $(\mathrm{tr}(\boldsymbol{S}_{1m_1})/\widehat{\Delta}_{\mu(m)}, \mathrm{tr}(\boldsymbol{S}_{2m_2})/\widehat{\Delta}_{\mu(m)}) = (3.99, 3.959)$. It is difficult to estimate the standard deviation of the average misclassification rate. However, by noting that $\mathrm{Var}\{\overline{e}(i)\}^{1/2} < \mathrm{Var}\{\widehat{e}(i)_r\}^{1/2} = [e(i)\{1 - e(i)\}/(m_i - n_i)]^{1/2}$, one may have an upper bound of the standard deviation for $\overline{e}(i)$ as

$$s_u(i) = [\overline{e}(i)\{1 - \overline{e}(i)\}/(m_i - n_i)]^{1/2},$$

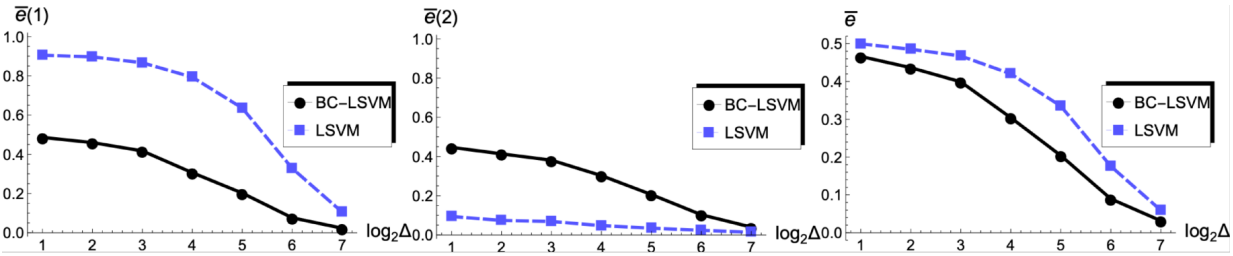
so that $\{\sum_{i=1}^2 s_u(i)^2/2\}^{1/2}$ ($= s_u$, say) for $\overline{e}$. For the BC-LSVM, $s_u(1) = 0.067$, $s_u(2) = 0.107$ and $s_u = 0.089$. We summarized the results for various n_is in Table 2.1.



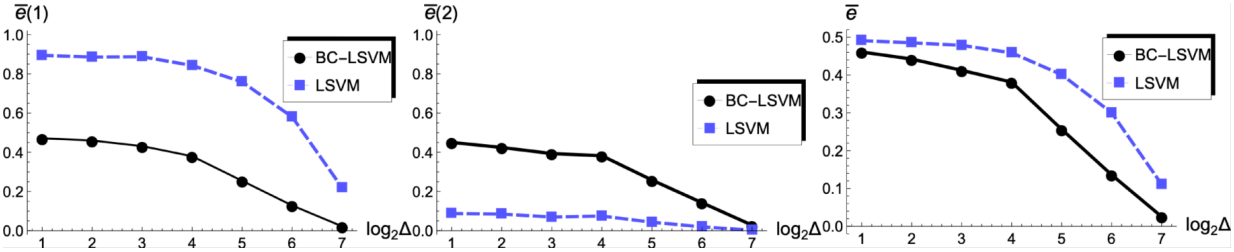
(d) $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_\alpha(d_*)$ $(\Delta_\mu \approx d^{2/3})$, $(n_1, n_2) = (5, 25)$ and $d = 2^s$, $s = 6, \ldots, 12$, for (I) $N_d(\mathbf{0}, \boldsymbol{\Sigma}_i)$



(e) $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_\alpha(d_*)$ $(\Delta_\mu \approx d^{2/3})$, $d = 1000$ and $(n_1, n_2) = (4s, 8s)$, $s = 1, \ldots, 7$, for (II) $t_d(\boldsymbol{\Sigma}_i, 10)$



(f) $d = 1000$, $(n_1, n_2) = (10, 20)$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_\alpha(2^s)$ $(\Delta_\mu = 2^s)$, $s = 1, \ldots, 7$, for (II) $t_d(\boldsymbol{\Sigma}_i, 10)$



(g) $d = 1000$, $(n_1, n_2) = (10, 20)$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_\beta(2^s)$ $(\Delta_\mu = 2^s)$, $s = 1, \ldots, 7$, for (II) $t_d(\boldsymbol{\Sigma}_i, 10)$

Figure 2.3: The error rates of the BC-LSVM and the LSVM are denoted by the solid lines and the dashed lines, respectively, for (d) to (g). The left panels display $\overline{e}(1)$, the middle panels display $\overline{e}(2)$ and the right panels display $\overline{e}$. Their standard deviations are less than 0.0112.

Table 2.1: Average misclassification rates of the BC-LSVM and the LSVM, together with $\hat{\delta}_\Sigma/\widehat{\Delta}_\mu$, for Alon et al. (1999)'s colon cancer data ($d = 2000$, $m_1 = 40$ and $m_2 = 22$). For each case, the standard deviations of $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ are less than $s_u(1)$, $s_u(2)$ and s_u, respectively.

	BC-LSVM			LSVM			
(n_1, n_2)	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}$	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}$	$\hat{\delta}_\Sigma/\widehat{\Delta}_\mu$
$(10, 5)$	0.188	0.209	0.198	0.122	0.309	0.215	-0.393
$(10, 10)$	0.16	0.166	0.163	0.158	0.161	0.159	0.003
$(10, 15)$	0.184	0.156	0.17	0.206	0.134	0.17	0.135
$(20, 5)$	0.164	0.249	0.206	0.082	0.475	0.278	-0.592
$(20, 10)$	0.141	0.177	0.159	0.116	0.23	0.173	-0.196
$(20, 15)$	0.142	0.167	0.154	0.133	0.181	0.157	-0.064
$(30, 5)$	0.144	0.302	0.223	0.083	0.566	0.324	-0.659
$(30, 10)$	0.12	0.236	0.178	0.108	0.318	0.213	-0.263
$(30, 15)$	0.115	0.203	0.159	0.1	0.263	0.181	-0.131

Next, we used leukemia data with 7129 ($= d$) genes given by Golub et al. (1999) which consists of π_1 : ALL (47 ($= m_1$) samples) and π_2 : AML (25 ($= m_2$) samples). We applied the BC-LSVM and the LSVM to the leukemia data and summarized the results in Table 2.2. When $n_1 \neq n_2$, $|\widehat{\Delta}_\mu|$ becomes large since $(\mathrm{tr}(\boldsymbol{S}_{1m_1})/\widehat{\Delta}_{\mu(m)}, \mathrm{tr}(\boldsymbol{S}_{2m_2})/\widehat{\Delta}_{\mu(m)}) = (2.693, 2.785)$. As expected theoretically, we observe that the BC-LSVM gives adequate performances compared to the LSVM when $|\widehat{\Delta}_\mu|$ is not small.

Finally, we used myeloma data with 12625 ($= d$) genes given by Tian et al. (2003) which consists of π_1 : patients without bone lesions (36 ($= m_1$) samples) and π_2 : patients with bone lesions (137 ($= m_2$) samples). We applied the BC-LSVM and the LSVM to the myeloma data and summarized the results in Table 2.3. When n_1 and n_2 are unbalanced, the LSVM gives a very bad performance. This is because Δ_μ in such cases is not sufficiently large since $(\mathrm{tr}(\boldsymbol{\Sigma}_1)/\Delta_\mu, \mathrm{tr}(\boldsymbol{\Sigma}_2)/\Delta_\mu) \approx (\mathrm{tr}(\boldsymbol{S}_{1m_1})/\widehat{\Delta}_{\mu(m)}, \mathrm{tr}(\boldsymbol{S}_{2m_2})/\widehat{\Delta}_{\mu(m)}) = (33.69, 33.53)$, so that δ_Σ/Δ_μ becomes too large when $n_1 \neq n_2$. Especially when $\hat{\delta}_\Sigma/\widehat{\Delta}_\mu > 1$, $\overline{e}(1)$ of the LSVM is too large. See Corollary 2.1 for the details. The BC-LSVM also does not give a low error rate for this data because Δ_μ is not sufficiently large. However, the BC-LSVM gives adequate performances compared to the LSVM especially when $\hat{\delta}_\Sigma/\widehat{\Delta}_\mu > 1$. Throughout Sections 2.3 and 2.4, we recommend to use the BC-LSVM rather than the LSVM for high-dimensional data.

2.5 Multiclass linear SVMs

In this section, we consider multiclass LSVMs in HDLSS settings. We have i.i.d. observations, $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$, from each π_i ($i = 1, \ldots, g$), where $g \geq 3$ and π_i has a d-dimensional distribution with an unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i$ ($\geq \boldsymbol{O}$). We assume $n_i \geq 2$, $i = 1, \ldots, g$. Let $\Delta_{\mu,ij} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2$ for $i, j = 1, \ldots, g$; $i \neq j$. We assume that $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0, \infty)$ as $d \to \infty$ for $i = 1, \ldots, g$, and $\limsup_{d\to\infty} \Delta_{\mu,ij}/d < \infty$ for $i, j = 1, \ldots, g$; $i \neq j$. We consider the one-versus-one approach (the max-wins rule). See Friedman (1996) and Bishop (2006) for the details. Let $n_g = \sum_{i=1}^g n_i$. First, we consider the case when $d \to \infty$ while n_g is fixed. We consider the following assumptions:

(B-i) $\dfrac{\max_{l=i,j} \mathrm{Var}(\|\boldsymbol{x}_{lk} - \boldsymbol{\mu}_l\|^2)}{\Delta_{\mu,ij}^2} \to 0$ as $d \to \infty$ for $i, j = 1, \ldots, g$; $i \neq j$;

Table 2.2: Average misclassification rates of the BC-LSVM and the LSVM, together with $\hat{\delta}_\Sigma/\widehat{\Delta}_\mu$, for Golub et al. (1999)'s leukemia data ($d = 7129$, $m_1 = 47$ and $m_2 = 25$). For each case, the standard deviations of $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ are less than $s_u(1)$, $s_u(2)$ and s_u, respectively.

	BC-LSVM			LSVM			
(n_1, n_2)	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}$	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}$	$\hat{\delta}_\Sigma/\widehat{\Delta}_\mu$
$(10, 5)$	0.044	0.077	0.06	0.012	0.148	0.08	-0.288
$(10, 10)$	0.036	0.043	0.04	0.036	0.046	0.041	-0.009
$(10, 20)$	0.044	0.034	0.039	0.074	0.026	0.05	0.13
$(20, 5)$	0.031	0.067	0.049	0.004	0.199	0.102	-0.422
$(20, 10)$	0.019	0.051	0.035	0.011	0.071	0.041	-0.144
$(20, 20)$	0.028	0.046	0.037	0.028	0.046	0.037	-0.005
$(40, 5)$	0.017	0.102	0.059	0.0	0.297	0.149	-0.49
$(40, 10)$	0.016	0.047	0.031	0.003	0.091	0.047	-0.211
$(40, 20)$	0.011	0.03	0.021	0.006	0.032	0.019	-0.072

(B-ii) $\dfrac{\max_{l=i,j} \mathrm{tr}(\boldsymbol{\Sigma}_l^2)}{\Delta_{\mu,ij}^2} \to 0$ as $d \to \infty$ for $i, j = 1, \ldots, g;\ i \neq j$.

Let $\delta_{\Sigma,ij} = \mathrm{tr}(\boldsymbol{\Sigma}_i)/n_i - \mathrm{tr}(\boldsymbol{\Sigma}_j)/n_j$ for $i, j = 1, \ldots, g;\ i \neq j$. We consider the following condition:

(B-iii) $\limsup_{d\to\infty} \dfrac{|\delta_{\Sigma,ij}|}{\Delta_{\mu,ij}} < 1$ for $i, j = 1, \ldots, g;\ i \neq j$.

From Theorem 2.1, for the one-versus-one approach by (2.5), we have the following result.

Corollary 2.4. *Under (B-i) to (B-iii), it holds for the multiclass LSVM that*

$$e(i) \to 0 \quad \textit{as } d \to \infty \textit{ for } i = 1, \ldots, g. \tag{2.11}$$

From Theorem 2.2, for the one-versus-one approach by (2.9), we have the following result.

Corollary 2.5. *Under (B-i) and (B-ii), the multiclass BC-LSVM holds (2.11).*

Note that the BC-LSVM satisfies the consistency property without (B-iii). Thus we recommend to use the BC-LSVM in multiclass HDLSS settings.

Next, we consider the case when both $d, n_g \to \infty$ while $n_g/d \to 0$. Similar to Section 2.2.3 and Corollary 2.2, the multiclass LSVMs have the consistency property under some regularity conditions.

We checked the performance of the multiclass LSVMs by using leukemia data with 12582 ($= d$) genes given by Armstrong et al. (2002) which consists of π_1 : ALL (24 ($= m_1$) samples), π_2 : MLL (20 ($= m_2$) samples) and π_3 : AML (28 ($= m_3$) samples). We applied the multiclass BC-LSVM and LSVM to the leukemia and summarized the results in Table 2.4. We had $(\mathrm{tr}(\boldsymbol{S}_{1m_1})/\widehat{\Delta}_{\mu,12(m)}, \mathrm{tr}(\boldsymbol{S}_{2m_2})/\widehat{\Delta}_{\mu,12(m)}) = (2.724, 3.213)$, $(\mathrm{tr}(\boldsymbol{S}_{1m_1})/\widehat{\Delta}_{\mu,13(m)}, \mathrm{tr}(\boldsymbol{S}_{3m_3})/\widehat{\Delta}_{\mu,13(m)}) = (0.738, 0.9)$ and $(\mathrm{tr}(\boldsymbol{S}_{2m_2})/\widehat{\Delta}_{\mu,23(m)}, \mathrm{tr}(\boldsymbol{S}_{3m_3})/\widehat{\Delta}_{\mu,23(m)}) = (1.533, 1.585)$, where $\widehat{\Delta}_{\mu,ij(m)} = \|\overline{\boldsymbol{x}}_{im_i} - \overline{\boldsymbol{x}}_{jm_j}\|^2 - \mathrm{tr}(\boldsymbol{S}_{im_i})/m_i - \mathrm{tr}(\boldsymbol{S}_{jm_j})/m_j$ that is an unbiased estimator of $\Delta_{\mu,ij}$. Thus $|\delta_{\Sigma,ij}/\Delta_{\mu,ij}|$ must become large when $n_i \neq n_j$. Actually, the multiclass BC-LSVM gives adequate performances for all the cases.

Table 2.3: Average misclassification rates of the BC-LSVM and the LSVM, together with $\hat{\delta}_\Sigma/\widehat{\Delta}_\mu$, for Tian et al. (2003)'s myeloma data ($d = 12625$, $m_1 = 36$ and $m_2 = 137$). For each case, the standard deviations of $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ are less than $s_u(1)$, $s_u(2)$ and s_u, respectively.

	BC-LSVM			LSVM			
(n_1, n_2)	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}$	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}$	$\hat{\delta}_\Sigma/\widehat{\Delta}_\mu$
$(10, 25)$	0.367	0.307	0.337	0.787	0.059	0.423	2.028
$(10, 50)$	0.407	0.265	0.336	0.936	0.013	0.475	2.698
$(10, 100)$	0.501	0.193	0.347	0.993	0.003	0.498	3.034
$(20, 25)$	0.311	0.288	0.299	0.401	0.214	0.308	0.343
$(20, 50)$	0.343	0.25	0.296	0.646	0.085	0.365	1.014
$(20, 100)$	0.436	0.175	0.306	0.872	0.026	0.449	1.349
$(30, 25)$	0.303	0.288	0.296	0.25	0.341	0.295	-0.218
$(30, 50)$	0.33	0.26	0.295	0.467	0.162	0.314	0.452
$(30, 100)$	0.382	0.195	0.288	0.713	0.068	0.391	0.788

Table 2.4: Average misclassification rates of the BC-LSVM and the LSVM for Armstrong et al. (2002)'s leukemia data ($d = 12582$, $m_1 = 24$, $m_2 = 20$ and $m_3 = 28$). For each case, the standard deviations of $\overline{e}(i)$, $i = 1, 2, 3$, and $\overline{e}$ are less than $s_u(i)$, $i = 1, 2, 3$, and $s_u = \{\sum_{i=1}^{3} s_u(i)^2/3\}^{1/2}$, respectively

	BC-LSVM				LSVM			
(n_1, n_2, n_3)	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}(3)$	$\overline{e}$	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}(3)$	$\overline{e}$
$(5, 5, 10)$	0.085	0.089	0.071	0.082	0.069	0.118	0.06	0.082
$(5, 5, 20)$	0.103	0.087	0.07	0.087	0.089	0.135	0.053	0.092
$(5, 10, 10)$	0.049	0.06	0.066	0.058	0.095	0.047	0.066	0.069
$(5, 10, 20)$	0.044	0.068	0.064	0.059	0.088	0.06	0.06	0.069
$(10, 5, 10)$	0.051	0.077	0.063	0.064	0.021	0.143	0.049	0.071
$(10, 5, 20)$	0.051	0.073	0.061	0.062	0.018	0.148	0.044	0.07
$(10, 10, 10)$	0.028	0.056	0.063	0.049	0.025	0.059	0.064	0.049
$(10, 10, 20)$	0.031	0.051	0.071	0.051	0.03	0.058	0.065	0.051

Appendix 2

Throughout, let $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_* = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$.

Proof of Lemma 2.1. Under (A-ii), we have that as $d \to \infty$

$$\boldsymbol{\mu}^T \boldsymbol{\Sigma}_i \boldsymbol{\mu}/\Delta_\mu^2 \leq \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}/\Delta_\mu = o(1), \quad i = 1, 2. \tag{2.12}$$

Then, by using Chebyshev's inequality, for any $\tau > 0$, under (A-ii), we have that

$$\begin{aligned}
&P(|(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*) - \Delta_\mu/4| \geq \tau\Delta_\mu) \\
&\leq (\tau\Delta)^{-2}E[\{(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*) - \Delta_\mu/4\}^2] \\
&= O\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2) + \boldsymbol{\mu}^T\boldsymbol{\Sigma}_1\boldsymbol{\mu}\}/\Delta_\mu^2 = o(1) \quad \text{for } 1 \leq j < j' \leq n_1; \\
&P(|(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*) - \Delta_\mu/4| \geq \tau\Delta_\mu) \\
&= O\{\mathrm{tr}(\boldsymbol{\Sigma}_2^2) + \boldsymbol{\mu}^T\boldsymbol{\Sigma}_2\boldsymbol{\mu}\}/\Delta_\mu^2 = o(1) \quad \text{for } n_1 + 1 \leq j < j' \leq n; \quad \text{and} \\
&P(|(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*) + \Delta_\mu/4| \geq \tau\Delta_\mu) \\
&= O\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) + \boldsymbol{\mu}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\boldsymbol{\mu}\}/\Delta_\mu^2 = o(1) \\
&\quad \text{for } 1 \leq j \leq n_1 \text{ and } n_1 + 1 \leq j' \leq n
\end{aligned} \tag{2.13}$$

from the fact that $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) \leq \{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_2^2)\}^{1/2}$. From (2.12), for any $\tau > 0$, we have that

$$\begin{aligned}
&P(|\|\boldsymbol{x}_j - \boldsymbol{\mu}_*\|^2 - \Delta_\mu/4 - \mathrm{tr}(\boldsymbol{\Sigma}_1)| \geq \tau\Delta_\mu) \\
&= O\{\mathrm{Var}(\|\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1\|^2) + \boldsymbol{\mu}^T\boldsymbol{\Sigma}_1\boldsymbol{\mu}\}/\Delta_\mu^2 = o(1) \quad \text{for } j = 1, \ldots, n_1; \quad \textit{and} \\
&P(|\|\boldsymbol{x}_j - \boldsymbol{\mu}_*\|^2 - \Delta_\mu/4 - \mathrm{tr}(\boldsymbol{\Sigma}_2)| \geq \tau\Delta_\mu) = o(1) \quad \text{for } j = n_1 + 1, \ldots, n
\end{aligned} \tag{2.14}$$

under (A-i) and (A-ii). Here, subject to (2.4), we can write for (2.3) that

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^n \alpha_j - \frac{1}{2}\sum_{j=1}^n\sum_{j'=1}^n \alpha_j\alpha_{j'}t_jt_{j'}(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*). \tag{2.15}$$

Then, by noting that $\alpha_j \geq 0$ for all j subject to (2.4), from (2.13) and (2.14), we have that

$$\begin{aligned}
L(\boldsymbol{\alpha}) = &\sum_{j=1}^n \alpha_j - \frac{\Delta_\mu}{8}\Big(\sum_{j=1}^n \alpha_j\Big)^2 - \frac{1}{2}\Big(\mathrm{tr}(\boldsymbol{\Sigma}_1)\sum_{j=1}^{n_1}\alpha_j^2 + \mathrm{tr}(\boldsymbol{\Sigma}_2)\sum_{j=n_1+1}^n \alpha_j^2\Big) \\
&+ o_P\Big\{\Delta_\mu\Big(\sum_{j=1}^n \alpha_j\Big)^2\Big\}
\end{aligned} \tag{2.16}$$

subject to (2.4) under (A-i) and (A-ii). It concludes the result. □

Proof of Lemma 2.2. By combining Lemma 2.6 with (2.7) and (3.9), we can claim the first result.

When $\hat{S} = \{1, \ldots, n\}$, by noting that $\sum_{j=1}^n \hat{\alpha}_j t_j = 0$, we have that

$$\begin{aligned}
\hat{y}(\boldsymbol{x}_0) = &\sum_{j=1}^n \hat{\alpha}_j t_j(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_0 - \boldsymbol{\mu}_*) + \sum_{j=1}^n \hat{\alpha}_j t_j(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T\boldsymbol{\mu}_* + \hat{b} \\
= &\sum_{j=1}^n \hat{\alpha}_j t_j(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_0 - \boldsymbol{\mu}_*) \\
&+ \frac{-n_1 + n_2}{n} - \frac{1}{n}\sum_{j=1}^n\sum_{j'=1}^n \hat{\alpha}_{j'}t_{j'}(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*).
\end{aligned} \tag{2.17}$$

From the first result of Lemma 2.2, (2.13) and (2.14), we have that as $d \to \infty$

$$\begin{aligned}
&\frac{-n_1+n_2}{n} - \frac{1}{n}\sum_{j=1}^{n}\sum_{j'=1}^{n} \hat{\alpha}_{j'} t_{j'} (\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T (\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*) \\
&= \frac{-n_1+n_2}{n} + \frac{(n_1-n_2)\Delta_\mu}{\Delta_{\mu*} n} + 2\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)}{\Delta_{\mu*} n} + o_P\Big(\frac{\Delta_\mu}{\Delta_{\mu*}}\Big) \\
&= \frac{-n_1+n_2}{n}\Big(\frac{\delta}{\Delta_{\mu*}}\Big) + 2\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)}{\Delta_{\mu*} n} + o_P\Big(\frac{\Delta_\mu}{\Delta_{\mu*}}\Big) \\
&= \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 - \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2}{\Delta_{\mu*}} + o_P\Big(\frac{\Delta_\mu}{\Delta_{\mu*}}\Big) \qquad (2.18)
\end{aligned}$$

under (A-i) and (A-ii). Similar to (2.13), under (A-ii), we obtain that $(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_0 - \boldsymbol{\mu}_*)/\Delta_\mu = (-1)^{i+1}/4 + o_P(1)$ for $j = 1, \ldots, n_1$, and $(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_0 - \boldsymbol{\mu}_*)/\Delta_\mu = (-1)^i/4 + o_P(1)$ for $j = n_1+1, \ldots, n$, when $\boldsymbol{x}_0 \in \pi_i$ $(i = 1, 2)$. Then, from the first result of Lemma 2.2, under (A-i) and (A-ii), it holds that

$$\sum_{j=1}^{n} \hat{\alpha}_j t_j (\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T (\boldsymbol{x}_0 - \boldsymbol{\mu}_*) = \frac{(-1)^i \Delta_\mu}{\Delta_{\mu*}} + o_P\Big(\frac{\Delta_\mu}{\Delta_{\mu*}}\Big) \qquad (2.19)$$

when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$. By combining (2.17) with (2.18) and (2.19), we can conclude the second result. □

Proofs of Theorem 2.1 and Corollary 2.1. By using (2.8), the results are obtained straightforwardly. □

Proof of Lemma 2.3. Similar to (2.13), under (A-ii'), from (2.12), we have that as $d, n \to \infty$

$$\begin{aligned}
&\sum_{1 \le j < j' \le n_1} P(|(\boldsymbol{x}_j - \boldsymbol{\mu}_1)^T (\boldsymbol{x}_{j'} - \boldsymbol{\mu}_1)| \ge \tau\Delta_\mu) = O\Big(\frac{n_1^2 \mathrm{tr}(\boldsymbol{\Sigma}_1^2)}{\Delta_\mu^2}\Big) = o(1); \\
&\sum_{n_1+1 \le j < j' \le n} P(|(\boldsymbol{x}_j - \boldsymbol{\mu}_2)^T (\boldsymbol{x}_{j'} - \boldsymbol{\mu}_2)| \ge \tau\Delta_\mu) = O\Big(\frac{n_2^2 \mathrm{tr}(\boldsymbol{\Sigma}_2^2)}{\Delta_\mu^2}\Big) = o(1); \\
&\sum_{j=1}^{n_1}\sum_{j'=n_1+1}^{n} P(|(\boldsymbol{x}_j - \boldsymbol{\mu}_1)^T (\boldsymbol{x}_{j'} - \boldsymbol{\mu}_2)| \ge \tau\Delta_\mu) = O\Big(\frac{n_1 n_2 \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)}{\Delta_\mu^2}\Big) = o(1); \\
&\sum_{j=1}^{n_1} P(|(\boldsymbol{x}_j - \boldsymbol{\mu}_1)^T \boldsymbol{\mu}| \ge \tau\Delta_\mu) = O\Big(\frac{n_1 \boldsymbol{\mu}^T \boldsymbol{\Sigma}_1 \boldsymbol{\mu}}{\Delta_\mu^2}\Big) = O\Big(\frac{n_1 \mathrm{tr}(\boldsymbol{\Sigma}_1^2)^{1/2}}{\Delta_\mu}\Big) = o(1); \\
&\text{and} \quad \sum_{j=n_1+1}^{n} P(|(\boldsymbol{x}_j - \boldsymbol{\mu}_2)^T \boldsymbol{\mu}| \ge \tau\Delta_\mu) = O\Big(\frac{n_2 \mathrm{tr}(\boldsymbol{\Sigma}_2^2)^{1/2}}{\Delta_\mu}\Big) = o(1)
\end{aligned}$$

for any $\tau > 0$. Then, under (A-ii'), we have that

$$\begin{aligned}
&(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T (\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*) = \Delta_\mu\{1 + o_P(1)\}/4 \quad \text{for all } 1 \le j < j' \le n_1; \\
&(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T (\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*) = \Delta_\mu\{1 + o_P(1)\}/4 \quad \text{for all } n_1+1 \le j < j' \le n; \quad \text{and} \\
&(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T (\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*) = -\Delta_\mu\{1 + o_P(1)\}/4 \\
&\text{for all } 1 \le j \le n_1 \text{ and } n_1+1 \le j' \le n. \qquad (2.20)
\end{aligned}$$

On the other hand, for any $\tau > 0$, we have that $\sum_{j=1}^{n_1} P(|\|\boldsymbol{x}_j - \boldsymbol{\mu}_*\|^2 - \Delta_\mu/4 - \mathrm{tr}(\boldsymbol{\Sigma}_1)| \ge \tau\Delta_\mu) = O\{n_1 \mathrm{Var}(\|\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1\|^2) + n_1 \boldsymbol{\mu}^T \boldsymbol{\Sigma}_1 \boldsymbol{\mu}\}/\Delta_\mu^2 = o(1)$ and $\sum_{j=n_1+1}^{n} P(|\|\boldsymbol{x}_j - \boldsymbol{\mu}_*\|^2 - \Delta_\mu/4 - \mathrm{tr}(\boldsymbol{\Sigma}_2)| \ge \tau\Delta_\mu) =$

$o(1)$ under (A-i') and (A-ii') as $d, n \to \infty$, so that

$$\begin{aligned} \|\boldsymbol{x}_j - \boldsymbol{\mu}_*\|^2 &= \Delta_\mu\{1 + o_P(1)\}/4 + \mathrm{tr}(\boldsymbol{\Sigma}_1) \quad \text{for all } 1 \le j \le n_1; \quad \text{and} \\ \|\boldsymbol{x}_j - \boldsymbol{\mu}_*\|^2 &= \Delta_\mu\{1 + o_P(1)\}/4 + \mathrm{tr}(\boldsymbol{\Sigma}_2) \quad \text{for all } n_1 + 1 \le j \le n. \end{aligned} \tag{2.21}$$

Then, by combining (2.15) with (2.20) and (2.21), we have (2.16) as $d, n \to \infty$, subject to (2.4) under (A-i') and (A-ii'). Similar to the proof of Lemma 2.2, by noting (A-iv), we can conclude the result. □

Proof of Lemma 2.4. We have that

$$\begin{aligned} \widehat{\Delta}_{\mu*} - \Delta_{\mu*} = &\sum_{i=1}^{2}\sum_{j=1}^{n_i} \frac{\|\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i\|^2 - \mathrm{tr}(\boldsymbol{\Sigma}_i)}{n_i^2} + \sum_{i=1}^{2}\sum_{j \neq j'}^{n_i} \frac{(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)^T(\boldsymbol{x}_{ij'} - \boldsymbol{\mu}_i)}{n_i^2} \\ &+ \sum_{i=1}^{2}(-1)^{i+1}\boldsymbol{\mu}^T(\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i) - 2(\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1)^T(\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2). \end{aligned} \tag{2.22}$$

Note that $E[\{\|\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i\|^2 - \mathrm{tr}(\boldsymbol{\Sigma}_i)\}^2] = o(\Delta_\mu^2)$ as $d \to \infty$ under (A-i) for all i, j. Also, note that $E[\{\boldsymbol{\mu}^T(\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i)\}^2] = \boldsymbol{\mu}^T\boldsymbol{\Sigma}_i\boldsymbol{\mu}/n_i \le \Delta_\mu \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}/n_i = o(\Delta_\mu^2/n_i)$ as $d \to \infty$ under (A-ii) for $i = 1, 2$. Then, from (2.22), we can claim that $E\{(\widehat{\Delta}_{\mu*} - \Delta_{\mu*})^2\} = o(\Delta_\mu^2)$ under (A-i) and (A-ii), so that $\widehat{\Delta}_{\mu*} = \Delta_{\mu*} + o_P(\Delta_\mu)$. On the other hand, we have that

$$\mathrm{tr}(\boldsymbol{S}_{in_i}) - \mathrm{tr}(\boldsymbol{\Sigma}_i) = \sum_{j=1}^{n_i} \frac{\|\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i\|^2 - \mathrm{tr}(\boldsymbol{\Sigma}_i)}{n_i} - \sum_{j \neq j'}^{n_i} \frac{(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)^T(\boldsymbol{x}_{ij'} - \boldsymbol{\mu}_i)}{n_i(n_i - 1)}.$$

Then, similar to $\widehat{\Delta}_{\mu*}$, we can claim that $\mathrm{tr}(\boldsymbol{S}_{in_i}) = \mathrm{tr}(\boldsymbol{\Sigma}_i) + o_P(\Delta_\mu)$ for $i = 1, 2$, under (A-i) and (A-ii), so that $\hat{\delta}_\Sigma = \delta_\Sigma + o_P(\Delta_\mu)$. Hence, by noting that $|\delta_\Sigma|/\Delta_\mu \le 1$, we can claim the result. □

Proof of Theorem 2.2. By using (2.10), the result is obtained straightforwardly. □

Proofs of Corollaries 2.2 and 2.3. From Lemma 2.4, we have (2.8) as $d, n \to \infty$ under (A-i'), (A-ii') and (A-iv). We note that Lemma 2.3 holds even when $d, n \to \infty$. Hence, from (2.8) and Lemma 2.3, we can claim the results. □

Proofs of Corollaries 2.4 and 2.5. By using Theorems 2.1 and 2.2, the results are obtained straightforwardly. □

Chapter 3

Soft-margin linear SVM in the HDLSS context

The hmLSVM shows high effective for high-dimensional data. However, it is not always classifiable for a test data due to overfitting. Therefore, further schemes are necessary to use the SVM for practical issues. First of all, it is possible to relax the constraint of the hard-margin to allow some discrepancies. The SVM defined in this formulation is called the soft-margin. In this chapter, we consider asymptotic properties of the smLSVM. This chapter is organized by Nakayama (2019).

In the HDLSS context, the smLSVM was investigated by Qiao and Zhang (2015) and Carmichael and Marron (2017). Qiao and Zhang (2015) proposed a unified family of classification machines, the FLexible Assortment MachinE (FLAME), which includes the SVM and distance weighted discrimination. Carmichael and Marron (2017) investigated behaviors of the SVM in the large and small tuning parameters using the Karush-Kuhn-Tucker conditions. They explored how the characteristics of the training data sets affect behaviors of the smSVM in many cases, the balance of classes, dimension and separability of data.

In this chapter, we investigate asymptotic properties of the LSVM both for the hard-margin and soft-margin. In Section 3.2, we give asymptotic properties of the hard-margin and soft-margin LSVM (naive SVMs) and show that their performances are affected by imbalance. In order to overcome such disadvantages, we propose a robust SVM (RSVM) to imbalanced data.

In Section 3.3, we check the performance of RSVM by numerical simulations.

In Section 3.4, we use RSVM in real data analyses.

3.1 Introduction

Suppose we have independent and d-variate two populations, $\pi_i,\ i = 1, 2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i$ ($\geq \boldsymbol{O}$). We assume that $\limsup_{d\to\infty} \|\boldsymbol{\mu}_i\|^2/d < \infty$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0, \infty)$ as $d \to \infty$ for $i = 1, 2$. Here, for a function, $f(\cdot)$, "$f(d) \in (0, \infty)$ as $d \to \infty$" implies $\liminf_{d\to\infty} f(d) > 0$ and $\limsup_{d\to\infty} f(d) < \infty$. Let $\Delta_\mu = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, where $\|\cdot\|$ denotes the Euclidean norm. We assume that $\limsup_{d\to\infty} \Delta_\mu/d < \infty$. We have independent and identically distributed (i.i.d.) observations, $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$, from each π_i. Let us write that

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}, \boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2}).$$

Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to one of the two populations. We assume $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent. We assume $n_i \geq 2,\ i = 1, 2$. Let $m = \min\{d, n_1, n_2\}$. Note that the divergence condition "$d \to \infty$, $n_1 \to \infty$ and $n_2 \to \infty$" is equivalent to "$m \to \infty$". Let $n = n_1 + n_2$ and $n_1 \leq n_2$. Let $t_j = -1$ for $j = 1, \ldots, n_1$ and $t_j = 1$ for $j = n_1 + 1, \ldots, n$.

In Chapter 2, we consider the hmLSVM for high-dimensional data. In this chapter, we consider asymptotic properties of the soft-margin LSVM (smLSVM). By borrowing the notation in Section 2.2.1, the discriminant function of the hmLSVM is given by

$$\hat{y}(\boldsymbol{x}) = \sum_{j\in\hat{S}} \hat{\alpha}_j t_j \boldsymbol{x}_j^T \boldsymbol{x} + \hat{b}. \tag{3.1}$$

Here $\hat{\alpha}_j$s are obtained by maximization $\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L(\boldsymbol{\alpha})$ subject to

$$\alpha_j \geq 0,\ j = 1, \ldots, n, \quad \text{and} \quad \sum_{j=1}^{n} \alpha_j t_j = 0. \tag{3.2}$$

Let $\hat{S} = \{j | \hat{\alpha}_j \neq 0,\ j = 1, \ldots, n\}$ and $n_S = \#S$, where $\#S$ denotes the number of elements in a set S. The intercept term is given by

$$\hat{b} = \frac{1}{n_{\hat{S}}} \sum_{j\in\hat{S}} \Big(t_j - \sum_{k\in\hat{S}} \hat{\alpha}_k t_k \boldsymbol{x}_j^T \boldsymbol{x}_k \Big).$$

the hmLSVM classifies $\boldsymbol{x}_0$ into π_1 if $\hat{y}(\boldsymbol{x}_0) < 0$ and into π_2 otherwise.

If the dataset is not linearly separable, one should use the smLSVM. The smLSVM is defined by the optimization problem $\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L(\boldsymbol{\alpha})$ in Section 2.2 under

$$0 \leq \alpha_j \leq C,\ j = 1, \ldots, n, \quad \text{and} \quad \sum_{j=1}^{n} \alpha_j t_j = 0. \tag{3.3}$$

where $C > 0$ is a regularization parameter. Let $\tilde{S} = \{j | \tilde{\alpha}_j \neq 0,\ j = 1, \ldots, n\}$. The intercept term of the smLSVM is given by

$$\hat{b}_{(S)} = \frac{1}{n_{\tilde{S}}} \sum_{j\in\tilde{S}} \Big(t_j - \sum_{j'\in\tilde{S}} \tilde{\alpha}_{j'} t_{j'} \boldsymbol{x}_j^T \boldsymbol{x}_{j'} \Big).$$

Then, we define the smLSVM is defined by

$$\hat{y}_{(S)}(\boldsymbol{x}) = \sum_{j\in\tilde{S}} \tilde{\alpha}_j t_j \boldsymbol{x}_j^T \boldsymbol{x} + \hat{b}_{(S)}. \tag{3.4}$$

The discriminant rule of the smLSVM is the same as the hmLSVM. See Vapnik (2000) for the details.

3.2 Asymptotic properties of linear SVM in HDLSS settings

In this section, we give asymptotic properties of the naive SVMs (hmLSVM and smLSVM) when $d, n \to \infty$ while $n/d \to 0$. Similar to Bai and Saranadasa (1996) and Aoshima and Yata (2014), we assume the following assumption for π_is as necessary:

(A-i) Let $\boldsymbol{z}_{ij},\ j = 1, \ldots, n_i$, be i.i.d. random p_i-vectors having $E(\boldsymbol{z}_{ij}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{z}_{ij}) = \boldsymbol{I}_{p_i}$ for each $i\ (= 1, 2)$ and some p_i. Let $\boldsymbol{z}_{ij} = (z_{i1j}, \ldots, z_{ip_ij})^T$ whose components satisfy that $\limsup_{d\to\infty} E(z_{irj}^4) < \infty$ for all r and

$$E(z_{irj}^2 z_{isj}^2) = E(z_{irj}^2)E(z_{isj}^2) = 1 \quad \text{and} \quad E(z_{irj}z_{isj}z_{itj}z_{iuj}) = 0$$

for all $r \neq s, t, u$. Then, the observations, $\boldsymbol{x}_{ij}$s, from each $\pi_i\ (i = 1, 2)$ are given by $\boldsymbol{x}_{ij} = \boldsymbol{\Gamma}_i \boldsymbol{z}_{ij} + \boldsymbol{\mu}_i,\ j = 1, \ldots, n_i$, where $\boldsymbol{\Gamma}_i$ is a $d \times p_i$ matrix such that $\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^T = \boldsymbol{\Sigma}_i$.

Note that z_{irj}s are i.i.d. as the standard normal distribution when the π_is are Gaussian and $\boldsymbol{\Gamma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i^{1/2}$, where $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i(1)}, \ldots, \lambda_{i(d)})$ is a diagonal matrix of eigenvalues, $\lambda_{i(1)} \geq \cdots \geq \lambda_{i(d)} \geq 0$, and $\boldsymbol{H}_i$ is an orthogonal matrix of the corresponding eigenvectors. Thus, (A-i) naturally holds when the π_is are Gaussian.

We assume the following assumption:

(A-ii) $\dfrac{n_i \mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{\Delta_\mu^2} \to 0$ as $m \to \infty$ for $i = 1, 2$.

Note that $\Delta_\mu^2/\mathrm{tr}(\boldsymbol{\Sigma}_i^2) = O(d)$ from the facts that $\limsup_{d\to\infty} \Delta_\mu/d < \infty$, $\mathrm{tr}(\boldsymbol{\Sigma}_i^2) \geq \mathrm{tr}(\boldsymbol{\Sigma}_i)^2/d$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0, \infty)$ as $d \to \infty$ for $i = 1, 2$. Thus, $n_i = o(d)$ when (A-ii) is met.

Note that $\sum_{j=1}^{n_1} \alpha_{1j} = \sum_{j=1}^{n_2} \alpha_{2j} (= \alpha_\star,$ say) from (3.2) and (3.3).

Lemma 3.1. *Assume (A-i) and (A-ii). Then, it holds that as $d \to \infty$*

$$\sum_{j=1}^{n} t_j \alpha_j \boldsymbol{x}_j^T \boldsymbol{x}_0 - \frac{1}{n}\sum_{j=1}^{n}\sum_{j'=1}^{n} \alpha_{j'} t_{j'} \boldsymbol{x}_j^T \boldsymbol{x}_{j'} = \frac{(-1)^i n_{i'} \alpha_\star \Delta_\mu}{n} + \frac{\alpha_\star\{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}}{n} + o_P(\alpha_\star \Delta_\mu)$$

when $\boldsymbol{x}_0 \in \pi_i,\ i \neq i'$.

The behaviors of (3.1) and (3.4) are different for a fixed dimension. We note that asymptotic properties of (3.1) and (3.4) are determined by evaluating $\alpha_\star$ from Lemma 3.1. In Section 3.2.1 and 3.2.2, we evaluate behaviors of $\hat{\alpha}_{ij}$s and $\tilde{\alpha}_{ij}$s.

3.2.1 Asymptotic properties of hard-margin SVM

In this section, we summarize asymptotic properties of the hmLSVM in HDLSS settings. Similar to Lemma 2.1 in Chapter 2 we evaluated the dual problem as follows:

Lemma 3.2. *Assume (A-i) and (A-ii). Under (3.2), it holds that as $m \to \infty$*

$$L(\boldsymbol{\alpha}) = 2\alpha_\star - \frac{\Delta_\mu}{2}\alpha_\star^2\{1 + o_P(1)\} - \frac{1}{2}\Big(\mathrm{tr}(\boldsymbol{\Sigma}_1)\sum_{j=1}^{n_1}\alpha_{1j}^2 + \mathrm{tr}(\boldsymbol{\Sigma}_2)\sum_{j=1}^{n_2}\alpha_{2j}^2\Big). \tag{3.5}$$

Then we consider the following two cases:

$$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i)}{n_i \Delta_\mu} \in (0, \infty); \tag{3.6}$$

$$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i)}{n_i \Delta_\mu} = o(1) \text{ as } m \to \infty. \tag{3.7}$$

We note that naive SVMs lose the discrimination performance under

$$\lim_{m\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_i)/\{n_i\Delta_\mu\} = \infty. \tag{3.8}$$

See Lemma 3.3 and Remark 7 for the details. Hence, we do not consider (3.8). First we evaluate (3.5) under (3.6). Let us write that

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n) = (\hat{\alpha}_{11}, \ldots, \hat{\alpha}_{1n_1}, \hat{\alpha}_{21}, \ldots, \hat{\alpha}_{2n_2})^T = \underset{\boldsymbol{\alpha}}{\mathrm{argmax}}\, L(\boldsymbol{\alpha}) \quad \text{subject to (3.2)}.$$

For the third term of (3.5), we have that

$$\min_{\sum_{j=1}^{n_1}\alpha_{1j}=\sum_{j=1}^{n_2}\alpha_{2j}=\alpha_\star} \mathrm{tr}(\boldsymbol{\Sigma}_1)\sum_{j=1}^{n_1}\alpha_{1j}^2 + \mathrm{tr}(\boldsymbol{\Sigma}_2)\sum_{j=1}^{n_2}\alpha_{2j}^2 = \alpha_\star^2\Big(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1)}{n_1} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_2)}{n_2}\Big)$$

when $\alpha_{ij} = \alpha_\star/n_i,\ j = 1, \ldots, n_i,\ i = 1, 2$ under (3.2). Let $\Delta_{\mu*} = \Delta_\mu + \mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2$. Under (3.6), it holds that

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = -\frac{\Delta_{\mu*}}{2}\Big(\alpha_\star - \frac{2+o_P(1)}{\Delta_{\mu*}}\Big)^2\{1+o_P(1)\} + \frac{2+o_P(1)}{\Delta_{\mu*}} \tag{3.9}$$

for $\alpha_\star$. Hence, by choosing $\alpha_\star \approx 2/\Delta_{\mu*}$, we have the maximum of $L(\boldsymbol{\alpha})$ asymptotically. Let $\delta_\Sigma = \mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 - \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2$. We have the following result.

Lemma 3.3. *Assume (A-i) and (A-ii). Under (3.6), it holds that as $m \to \infty$*

$$\hat{\alpha}_{ij} = \frac{2}{\Delta_{\mu*}n_i}\{1+o_P(1)\} \quad \textit{for } j = 1, \ldots, n_i;\ i = 1, 2.$$

Furthermore, it holds that as $m \to \infty$

$$\hat{y}(\boldsymbol{x}_0) = \frac{(-1)^i\Delta_\mu}{\Delta_{\mu*}} + \frac{\delta_\Sigma}{\Delta_{\mu*}} + o_P\Big(\frac{\Delta_\mu}{\Delta_{\mu*}}\Big) \quad \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2.$$

Remark 7. *From the second result of Lemma 3.3, the discrimination performance of the hmLSVM depends on the first term of $\hat{y}(\boldsymbol{x}_0)$ and we can observe that the hmLSVM needs large Δ_μ for the classification. Under (3.8), it holds that $\hat{y}(\boldsymbol{x}_0) = o(1) + \delta_\Sigma/\Delta_{\mu*} + o_P(1)$ as $m \to \infty$, so that we cannot use the hmLSVM.*

Nakayama et al. (2017) gave the result similar to Lemma 3.3 when $d \to \infty$ while n_is are fixed.

Next similar to the case of (3.6), we evaluate (3.5) under (3.7). By noting that $\alpha_\star^2 \geq \sum_{j=1}^{n_1}\alpha_j^2$ and $\alpha_\star^2 \geq \sum_{j=n_1+1}^{n}\alpha_j^2$ under (3.2), from Lemma 3.2, it holds that

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = -\frac{\Delta_\mu}{2}\Big(\alpha_\star - \frac{2+o_P(1)}{\Delta_\mu}\Big)^2\{1+o_P(1)\} + \frac{2+o_P(1)}{\Delta_\mu}. \tag{3.10}$$

for $\alpha_\star$. Hence, by choosing $\alpha_\star \approx 2/\Delta_\mu$, we have the maximum of $L(\boldsymbol{\alpha})$ asymptotically. Under (3.7), we have the following result.

Lemma 3.4. *Assume (A-i) and (A-ii). Under (3.7), it holds that as $m \to \infty$*

$$\alpha_\star = \frac{2}{\Delta_\mu}\{1+o_P(1)\}.$$

Furthermore, it holds that as $m \to \infty$

$$\hat{y}(\boldsymbol{x}_0) = (-1)^i + o_P(1) \quad \textit{when} \quad \boldsymbol{x}_0 \in \pi_i \quad \textit{for} \quad i = 1, 2.$$

We note that $\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha})$ is determined by $\alpha_\star$ from (3.9) and (3.10). By restricting the results of the hmLSVM to $0 \leq \alpha_j \leq C$, we consider asymptotic properties of the smLSVM.

3.2.2 Asymptotic properties of soft-margin SVM

In this section, we give asymptotic properties of the smLSVM in HDLSS settings. The properties of the smLSVM are affected by the regularization parameter C. For $C > 0$, we consider the following conditions:

$$\limsup_{m\to\infty} \frac{2}{n_1 \Delta_{\mu*} C} \leq 1; \textit{and} \tag{3.11}$$

$$\liminf_{m\to\infty} \frac{2}{n_1 \Delta_{\mu*} C} > 1. \tag{3.12}$$

We consider asymptotic properties of the smLSVM under the following four cases:

(I) The conditions (3.6) and (3.11); (II) The conditions (3.6) and (3.12);
(III) The conditions (3.7) and (3.11); (IV) The conditions (3.7) and (3.12).

First, we consider the smLSVM under (3.6). By noting that

$$0 \leq \alpha_\star \leq n_1 C \tag{3.13}$$

under (3.3), it holds that $\alpha_\star \leq 2/\Delta_{\mu*} \leq n_1 C$ under (3.11). Hence by choosing $\alpha_\star \approx 2/\Delta_{\mu*}$, we have the maximum of $L(\boldsymbol{\alpha})$ asymptotically. Similar to the hmLSVM, we denote

$$\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n) = (\tilde{\alpha}_{11}, \ldots, \tilde{\alpha}_{1n_1}, \tilde{\alpha}_{21}, \ldots, \tilde{\alpha}_{2n_2})^T = \operatorname*{argmax}_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) \quad \text{subject to (3.3)}.$$

We have the following results under (I) as $m \to \infty$:

$$\tilde{\alpha}_{ij} = \frac{2}{\Delta_{\mu*} n_i}\{1 + o_P(1)\} \quad \text{for } j = 1, \ldots, n_i;\ i = 1, 2.$$

From the fact that $\tilde{\boldsymbol{\alpha}} \approx \hat{\boldsymbol{\alpha}}$, it holds that $\hat{y}_{(S)}(\boldsymbol{x}_0) \approx \hat{y}(\boldsymbol{x}_0)$.

Remark 8. *By assuming*

$$\liminf_{m\to\infty} \frac{|\delta_\Sigma|}{\Delta_\mu} < 1, \tag{3.14}$$

The smLSVM holds

$$e(i) \to 0 \quad \textit{as } d \to \infty \textit{ for } i = 1, 2 \tag{3.15}$$

under (I). The smLSVM holds the following inconsistency properties:

$$e(1) = 1 + o(1) \quad \textit{and} \quad e(2) = o(1) \quad \textit{as } m \to \infty \tag{3.16}$$

$$\textit{if} \quad \liminf_{m\to\infty} \frac{\delta_\Sigma}{\Delta_\mu} > 1; \quad \textit{and}$$

$$e(1) = o(1) \quad \textit{and} \quad e(2) = 1 + o(1) \quad \textit{as } m \to \infty \tag{3.17}$$

$$\textit{if} \quad \limsup_{m\to\infty} \frac{\delta_\Sigma}{\Delta_\mu} < -1,$$

under (A-i) and (A-ii) with (I). For the hmLSVM, Hall et al. (2005), Qiao and Zhang (2015) and Nakayama et al. (2017) also showed the consistency property (3.15) and the inconsistent properties (3.16) and (3.17).

Under (3.12), it holds that $\alpha_\star \leq n_1 C \leq 2/\Delta_{\mu*}$. Hence, by choosing $\alpha_\star \approx n_1 C$, we have the maximum of $L(\boldsymbol{\alpha})$ asymptotically. $\alpha_\star = n_1 C$ is equivalent to the fact that $\alpha_{1j} = C,\ j = 1, \ldots, n_1$ under (3.3). We have the following results.

Lemma 3.5. *Assume (A-i) and (A-ii). Under (II), it holds that as $m \to \infty$*

$$\tilde{\alpha}_{1j} = C\{1 + o_P(1)\} \quad \textit{for } j = 1, \ldots, n_1; \textit{ and}$$
$$\tilde{\alpha}_{2j} = \frac{n_1 C}{n_2}\{1 + o_P(1)\} \quad \textit{for } j = 1, \ldots, n_2.$$

Furthermore, assume also

$$\liminf_{m \to \infty} \frac{n_1^2 C \Delta_\mu}{n} > 0. \tag{3.18}$$

Then, it holds that as $m \to \infty$

$$\hat{y}_{(S)}(\boldsymbol{x}_0) = \frac{n_1 C}{n}\Big((-1)^i n_{i'} \Delta_\mu + \mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\Big)\{1 + o_P(1)\} + \frac{n_2 - n_1}{n}$$
$$\textit{when} \quad \boldsymbol{x}_0 \in \pi_i \quad \textit{for} \quad i \neq i'.$$

Remark 9. *Lemma 24 in Carmichael and Marron (2017) showed that if $C < 1/\{2n_1 D^2\}$, all data points in the smaller class π_1 are the support vector, where $D = \max \|\boldsymbol{x}_{1j} - \boldsymbol{x}_{2j'}\|$.*

Note that the smLSVM loses the discrimination ability when (3.18) in Lemma 3.5 is not met.

Theorem 3.1. *Assume (A-i) and (A-ii). Assume also*

$$\limsup_{m \to \infty} \Big(\frac{n_{i'} - n_i}{n_1 n_i \Delta_\mu C} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i) - \mathrm{tr}(\boldsymbol{\Sigma}_{i'})}{n_i \Delta_\mu}\Big) < 1, \quad i \neq i'. \tag{3.19}$$

Under (II) and (3.18), the smLSVM holds (3.15).

In the balanced case ($n_1 = n_2$), (3.19) is equivalent to the condition (3.14). If one cannot assume (3.19), we have the following result.

Corollary 3.1. *Assume (A-i) and (A-ii). Under (II) and (3.18), (3.4) holds the following properties:*

$$(3.16) \textit{ if } \liminf_{m \to \infty} \Big(\frac{n_2 - n_1}{n_1^2 \Delta C} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)}{n_1 C}\Big) > 1; \quad \textit{and}$$
$$(3.17) \textit{ if } \limsup_{m \to \infty} \Big(\frac{n_2 - n_1}{n_1 n_2 \Delta C} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)}{n_2 C}\Big) < -1.$$

Next, we consider the smLSVM under (3.7). It holds that $\Delta_{\mu*} = \Delta_\mu\{1 + o(1)\}$ under (3.7), so that one may replace $2/\{n_1 \Delta_{\mu*} C\}$ in (3.11) and (3.12) with $2/\{n_1 \Delta_\mu C\}$. Under (3.11), we have that $\alpha_\star = 2/\Delta_\mu\{1 + o_P(1)\}$ from Lemma 3.4. Hence, we have that as $m \to \infty$

$$\hat{y}_{(S)}(\boldsymbol{x}_0) = (-1)^i + \frac{2\{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}}{n \Delta_\mu} + o_P(1) \quad \text{when} \quad \boldsymbol{x}_0 \in \pi_i \quad \text{for} \quad i = 1, 2.$$

By noting $\mathrm{tr}(\boldsymbol{\Sigma}_i)/\{n\Delta_\mu\} \to 0$ under (3.7), it holds that as $m \to \infty$

$$\hat{y}_{(S)}(\boldsymbol{x}_0) = (-1)^i + o_P(1) \quad \text{when} \quad \boldsymbol{x}_0 \in \pi_i \quad \text{for} \quad i = 1, 2.$$

Hence the smLSVM holds (3.15) under (III). On the other hand, by choosing $\alpha_\star \approx n_1 C$, we have the maximum of $L(\boldsymbol{\alpha})$ asymptotically under (IV). Let $\tilde{S}_i = \{j | \tilde{\alpha}_{ij} \neq 0, \ j = 1, \ldots, n_i\}$ for $i = 1, 2$. We have the following result.

Lemma 3.6. *Assume (A-i) and (A-ii). Under (IV), it holds that as $m \to \infty$*

$$\tilde{\alpha}_{1j} = C\{1 + o_P(1)\} \quad \textit{for } j = 1, \ldots, n_1; \textit{and}$$
$$\alpha_\star = n_1 C\{1 + o_P(1)\}$$

Furthermore, under (3.18), it holds that as $m \to \infty$

$$\hat{y}_{(S)}(\boldsymbol{x}_0) = \frac{n_1 C}{n_{\tilde{S}}}\Big((-1)^i n_{\tilde{S}_{i'}} \Delta_\mu - \mathrm{tr}(\boldsymbol{\Sigma}_2)\Big)\{1 + o_P(1)\} + \frac{n_{\tilde{S}_2} - n_1}{n_{\tilde{S}}}$$
$$\textit{when} \quad \boldsymbol{x}_0 \in \pi_i \quad \textit{for} \quad i \neq i'.$$

Note that although all the data points in the smaller class π_1 are the support vector, some data points in the larger class π_2 are. It holds that $n_{\tilde{S}_1} = n_1$ from Lemma 3.6. Then, we have the following theorem.

Theorem 3.2. *Assume (A-i) and (A-ii). Under (IV) and (3.18), the smLSVM holds that*

$$(3.15) \ \textit{if} \ \liminf_{m\to\infty} \frac{n_1 C\{n_1 \Delta_\mu - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}}{n_{\tilde{S}_2} - n_1} > 1.$$

Corollary 3.2. *Assume (A-i) and (A-ii). Under (IV) and (3.18), the smLSVM holds that*

$$(3.16) \ \textit{if} \ \limsup_{m\to\infty} \frac{n_1 n_{\tilde{S}_2} C \Delta_\mu}{n_{\tilde{S}_2} - n_1} < 1; \quad \textit{and}$$
$$(3.17) \ \textit{if} \ \limsup_{m\to\infty} \frac{n_1 C\{n_1 \Delta_\mu - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}}{n_{\tilde{S}_2} - n_1} < 1.$$

The smLSVM is robust to the choice of the regularization parameter C under (I) and (III). However the smLSVM holds inconsistency properties (3.16) and (3.17) under (II) and (IV) from Corollary 3.1 and 3.2. The main reason is that $\hat{b}_{(S)}$ depends on the imbalance of populations highly. In order to overcome the difficulty, we give a robust intercept term in Section 3.2.3.

3.2.3 Robust SVM for high-dimensional imbalanced data

From Section 3.2.1 and 3.2.2, the intercept term of naive SVMs is affected by imbalanced data. For such data, one should not use naive SVMs. In this section, we define a robust intercept to imbalanced data. For $\boldsymbol{w} = (w_{11}, \ldots, w_{1n_1}, w_{21}, \ldots, w_{2n_2})^T = (w_1, \ldots, w_n)^T$ whose components satisfy that $w_{ij} \geq 0$ and $W = \sum_{j=1}^{n_1} w_{1j} = \sum_{j=1}^{n_2} w_{2j}$, let

$$b_r(\boldsymbol{w}) = \sum_{j \neq j'}^{n_{S_1}} w_{1j} \boldsymbol{x}_{1j}^T \boldsymbol{x}_{1j'} / \{2(n_{S_1} - 1)\} - \sum_{j \neq j'}^{n_{S_2}} w_{2j} \boldsymbol{x}_{2j}^T \boldsymbol{x}_{2j'} / \{2(n_{S_2} - 1)\},$$

where $S_i = \{j | w_{ij} \neq 0, \ j = 1, \ldots, n_i\}$ for $i = 1, 2$. We define a weighted classifier:

$$y_r(\boldsymbol{x}_0; \boldsymbol{w}) = \sum_{j \in S} w_j t_j \boldsymbol{x}_j^T \boldsymbol{x}_0 + b_r(\boldsymbol{w}), \tag{3.20}$$

where $S = S_1 \cup S_2$.

We note that if we set $w_{1j} = 1/n_1$ and $w_{2j} = 1/n_2$ in (3.20), the classifier by (3.20) is equivalent to the distance-based discriminant analysis (DBDA) by Aoshima and Yata (2014). The classifier is the distance-based classifier. Note that $E[y_r(\boldsymbol{x}_0; \boldsymbol{w})] = (-1)^i W \Delta / 2$ for $\boldsymbol{x}_0 \in \pi_i$ without any assumptions. We assume the following assumptions:

(A-iii) $\liminf_{m\to\infty} W\Delta_\mu > 0.$

We note that $W = 1$ for the distance-based classifier, so that (A-iii) holds. We have the following theorem.

Theorem 3.3. *Assume (A-i) to (A-iii). Then, it holds that as $m \to \infty$*

$$y_r(\boldsymbol{x}_0) = \frac{(-1)^i W\Delta_\mu}{2}\{1 + o_P(1)\}.$$

Hence (3.20) holds (3.15).

By substituting $\hat{\alpha}_j$ or $\tilde{\alpha}_j$ for w_j in the classifier given by (3.20), we propose robust SVMs (RSVM):

$$y_{rh}(\boldsymbol{x}_0; \hat{\boldsymbol{\alpha}}) = \sum_{j\in\hat{S}} \hat{\alpha}_j t_j \boldsymbol{x}_j^T \boldsymbol{x}_0 + b_r(\hat{\boldsymbol{\alpha}}); \textit{and} \tag{3.21}$$

$$y_{rs}(\boldsymbol{x}_0; \tilde{\boldsymbol{\alpha}}) = \sum_{j\in\tilde{S}} \tilde{\alpha}_j t_j \boldsymbol{x}_j^T \boldsymbol{x}_0 + b_r(\tilde{\boldsymbol{\alpha}}) \tag{3.22}$$

From Theorem 3.3, the performance of (3.21) and (3.22) is determined by $\alpha_\star\Delta_\mu$. From Section 3.2.1 and 3.2.2, $\alpha_\star$ is asymptotically one of $2/\Delta_{\mu*}$ $2/\Delta_\mu$ and n_1C. We have that $\Delta_\mu/\Delta_{\mu*} = 1 + o(1)$ under (3.7). It holds that $\liminf_{m\to\infty} \Delta_\mu/\Delta_{\mu*} > 0$ in either case of (3.6) and (3.7). In the first two cases, (A-iii) is met.

Corollary 3.3. *Assume (A-i) and (A-ii). Then, RSVMs under (3.18) hold (3.15).*

Remark 10. *We consider asymptotic properties of RSVMs in the case where n_is are fixed and $n_i \to \infty$. Note that $\mathrm{tr}(\boldsymbol{\Sigma}_i)/\Delta_\mu = o(1)$ does not hold from the assumptions that $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0, \infty)$ and $\limsup_{d\to\infty} \|\boldsymbol{\mu}_i\|^2/d < \infty$. Hence we consider (3.6) or (I) and (II) when n_is are fixed. Under (3.6), it holds that $y_{rh}(\boldsymbol{x}_0; \hat{\boldsymbol{\alpha}}) = (-1)^i\Delta_\mu/\Delta_{\mu*}\{1 + o_P(1)\}$ for $\boldsymbol{x}_0 \in \pi_i,\ i = 1, 2$ from Lemma 3.3. Similarly, $y_{rs}(\boldsymbol{x}_0; \tilde{\boldsymbol{\alpha}}) = (-1)^i\Delta_\mu/\Delta_{\mu*}\{1 + o_P(1)\}$ holds under (I). Hence we can obtain (3.15) for RSVMs under (3.6) or (I) without assuming (3.14). Under (II), we have that $y_{rs}(\boldsymbol{x}_0; \tilde{\boldsymbol{\alpha}}) = (-1)^i n_1 n_{i'} C\Delta_\mu/n\{1+o_P(1)\}$ for $\boldsymbol{x}_0 \in \pi_i,\ i \neq i'$ from Lemma 3.5, so that the RSVM (3.22) has (3.15) without assuming (3.19). In the case where n_is are fixed, RSVMs give (3.15) under (3.6), (I) and (II). On the other hand, we consider (3.7), (III) and (IV) as $n_i \to \infty$. We have $y_{rh}(\boldsymbol{x}_0; \hat{\boldsymbol{\alpha}}) \approx \hat{y}(\boldsymbol{x}_0)$ and $y_{rs}(\boldsymbol{x}_0; \tilde{\boldsymbol{\alpha}}) \approx \hat{y}_{(S)}(\boldsymbol{x}_0)$ under (3.7) or (III). Similar to (II), $y_{rs}(\boldsymbol{x}_0; \tilde{\boldsymbol{\alpha}}) = (-1)^i n_1 n_{\tilde{S}_{i'}} C\Delta_\mu/n_{\tilde{S}}\{1 + o_P(1)\}$ holds for $\boldsymbol{x}_0 \in \pi_i,\ i \neq i'$ under (IV), so that we can obtain (3.15) without assuming $\liminf_{m\to\infty} n_1C\{n_1\Delta_\mu - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}/(n_{\tilde{S}_2} - n_1) > 1$. Hence we have (3.15) for RSVMs under (3.7), (III) and (IV) as $n_i \to \infty$.*

Remark 11. *From Section 3.2.1 and 3.2.2, for C satisfying (3.11), the smLSVM become the hmLSVM. From the condition (3.12), we consider the candidate $C_0 = 1/\{n_1\Delta_{\mu*}\}$. We have for C_0 that $\liminf_{d\to\infty} 2/\{n_1\Delta_{\mu*}C_0\} > 1$. Let $\widehat{\Delta}_{\mu*} = \|\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{n_2}\|^2$, where $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i,\ i = 1, 2$. It holds as $m \to \infty$ that $\widehat{\Delta}_{\mu*} = \Delta_{\mu*}\{1 + o_P(1)\}$ under (A-i) and (A-ii), so that we recommend $\widehat{C}_0 = 1/\{n_1\widehat{\Delta}_\mu\}$.*

The distance-based classifier is asymptotically equivalent to RSVMs in the sense that they classify by the difference of mean vectors. The distance-based classifier is defined by all the data with the same weights $1/n_i$. On the other hand, $\hat{S}$ and $\tilde{S}$ are sparse sets so that RSVMs use some data points. Nakayama et al. (2017) considered a bias correction for the hmLSVM (BC-LSVM). See Section 2 for the details. It holds for the BC-LSVM that

$$y_{bc}(\boldsymbol{x}_0) = (-1)^i \frac{\Delta_\mu}{\Delta_{\mu*}}\{1 + o_P(1)\} \quad \text{when } \boldsymbol{x}_0 \in \pi_i \text{ for } i = 1, 2$$

as $d \to \infty$ under (A-i) and (A-ii) when n_is are fixed. From Remark 10, The BC-LSVM is asymptotically equivalent to (3.21) and (3.22) under (3.6) and (I) when n_is are fixed.

3.3 Simulations

In this section, we check the performance of the RSVM in numerical simulations and real data analyses. We compare six linear classifiers, hmLSVM, smSVM, RSVM (3.22), BC-LSVM, DBDA and the diagonal linear discriminant analysis (DLDA). DLDA is given by Dudoit et al. (2002) and Bickel and Levina (2004). The rule of DLDA is given for $\boldsymbol{x}_0 \in \pi_1$ (or π_2) by

$$\{\boldsymbol{x}_0 - (\overline{\boldsymbol{x}}_{1n_1} + \overline{\boldsymbol{x}}_{2n_2})/2\}^T \boldsymbol{S}_d^{-1}(\overline{\boldsymbol{x}}_{2n_2} - \overline{\boldsymbol{x}}_{1n_1}) < 0 \text{ (or } \geq 0)$$

where $\boldsymbol{S}_d = \mathrm{diag}(s_{1n}, \ldots, s_{dn}), s_{jn} = \sum_{i=1}^{2}\sum_{l=1}^{n_i}(x_{ijl} - \overline{x}_{ijn_i})^2/(n_1+n_2-2)$, $\boldsymbol{x}_{ij} = (x_{i1j}, \ldots, x_{idj})$ and $\overline{x}_{ijn_i} = \sum_{l=1}^{n_i} x_{ijl}/n_i$. We set $n_1 = 40,\ n_2 = 20$ and $d = 2^s,\ s = 3, \ldots, 10$. Independent pseudo random observations were generated from $\pi_i : N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),\ i = 1, 2$. We set $\boldsymbol{\mu}_1 = \boldsymbol{0}$, $\boldsymbol{\Sigma}_1 = 0.9(0.3^{|i-j|^{1/3}})$ and $\boldsymbol{\Sigma}_2 = 1.1(0.4^{|i-j|^{1/3}})$. For the mean vector $\boldsymbol{\mu}_2$, we considered two cases:

(a) $\boldsymbol{\mu}_2 = (1/5, \ldots, 1/5)$; and
(b) $\boldsymbol{\mu}_2 = (1, \ldots, 1, 0, \ldots, 0)$ whose the first $\lceil d^{1/2} \rceil$ elements are 1.

Here, $\lceil x \rceil$ denotes the smallest integer $\geq x$. For $\boldsymbol{x}_0 \in \pi_i$ $(i = 1, 2)$ we repeated 2000 times to confirm if six classifiers do (or do not) classify $\boldsymbol{x}_0 \in \pi_i$ correctly and defined $P_{ir} = 0$ or 1 accordingly for each π_i $(i = 1, 2)$. We use $C = \widehat{C}_0$ for the RSVM from Remark 11. We calculated the error rates, $\overline{e}(i) = \sum_{r=1}^{2000} P_{ir}/2000,\ i = 1, 2$. Also, we calculated the average error rate, $\overline{e} = \{\overline{e}(1) + \overline{e}(2)\}/2$. Their standard deviations are less than 0.011. In Figure 3.1, we plotted $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ for (a) and (b). We observe that the hmLSVM and smLSVM give imbalanced results in Figure 3.1. The main reason is probably that the intercept term is affected by the bias of imbalance. Thus $\overline{e}(1)$ becomes close to 1 though $\overline{e}(2)$ approaches to 0 as d increases. See Remark 8, Corollary 3.1 and 3.2. DLDA became bad as the dimension increases. As expected theoretically, we observed that the RSVM gives preferable performances. Also, three classifiers, RSVM, BC-LSVM, and DBDA, gave equivalent and good results.

Next, we set $d = 2^3$ for (a) and (b). Then we see behaviors of the smLSVM and RSVM (3.22) for the regularization parameter $C = 8^{-s+2},\ s = 1, \ldots, 8$. We examine similar experiments for also 2^6 and 2^9. In Figure 3.2, we plotted $\overline{e}$ of the smLSVM, RSVM and RSVM with $C = \widehat{C}_0$ for $d = 2^3,\ 2^6$ and $d = 2^9$. We observed that the smLSVM tends to become bad for small C from Figure 3.2. On the other hand, the RSVM does not depend on the choice of C and gives lower error rates than those of the smSVM. Also, $\widehat{C}_0$ shows good behaviors as the dimension increase.

Next, we compared the performance of the RSVM in complex settings. We set $\boldsymbol{\mu}_1 = \boldsymbol{0}$, $\boldsymbol{\mu}_2 = (-1, \ldots, -1, 0, \ldots, 0)$ whose the first $\lceil d^{1/2} \rceil$ elements are -1, $\boldsymbol{\Sigma}_1 = 0.9\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_2 = 1.1\boldsymbol{B}(0.4^{|i-j|^{1/3}})\boldsymbol{B}$, where

$$B = \mathrm{diag}[\{0.5 + 1/(d+1)\}^{1/2}, \ldots, \{0.5 + d/(d+1)\}^{1/2}].$$

We generated $\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i,\ i = 1, 2;\ j = 1, 2, \ldots$, independently either from (D-I) $\boldsymbol{H}_i\boldsymbol{\Lambda}_i^{1/2}\boldsymbol{z}_{ij}$, where $z_{irj} = (y_{irj} - 1)/\sqrt{2}$ $(r = 1, \ldots, d)$ in which y_{irj} s are i.i.d. as the chi-squared distribution with 1 degree of freedom and $\boldsymbol{\Lambda}_i = \boldsymbol{H}_i^T\boldsymbol{\Sigma}_i\boldsymbol{H}_i$, and (D-II) a d-variate t-distribution, $t_d(\boldsymbol{0}, \boldsymbol{\Sigma}_i, \nu), i = 1, 2$, with mean zero, covariance matrix $\boldsymbol{\Sigma}_1$ and degrees of freedom $\nu = 10$. Note that (A-i) is met in (D-I). We considered four cases:

(c) $n_1 = n_2 = 5s,\ d = 2^s,\ s = 3, \ldots, 10$ for (D-I);
(d) $n_1 = 2\log_2 d,\ n_2 = 30,\ d = 2^s,\ s = 3, \ldots, 10$ for (D-I);
(e) $n_1 = 5s,\ n_2 = 10s,\ s = 1, \ldots, 7$ when $d = 2^3,\ 2^6$ and 2^9 for (D-II); and
(f) $n_1 = 10s,\ n_2 = 50,\ s = 1, \ldots, 7$ when $d = 2^3,\ 2^6$ and 2^9 for (D-II);

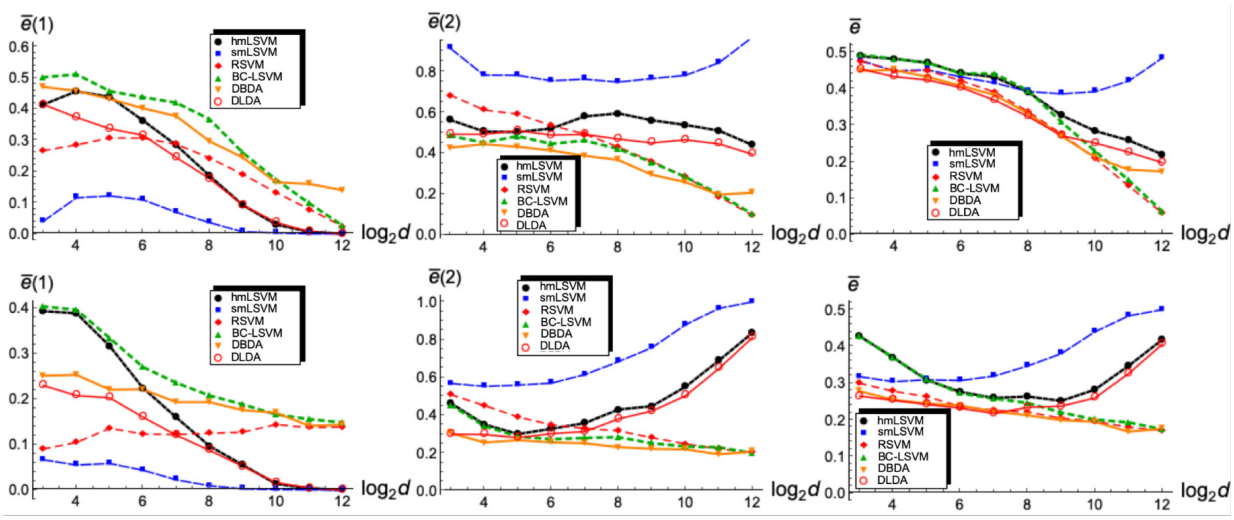


Figure 3.1: The performance of six classifiers, hmLSVM, smLSVM, RSVM (3.22), BC-LSVM, DBDA and DLDA. The left panel displays $\overline{e}(1)$, the middle panel displays $\overline{e}(2)$ and the right panel displays $\overline{e}$. The upper three panels show the case of (a) and the lower three panels show the case of (b).

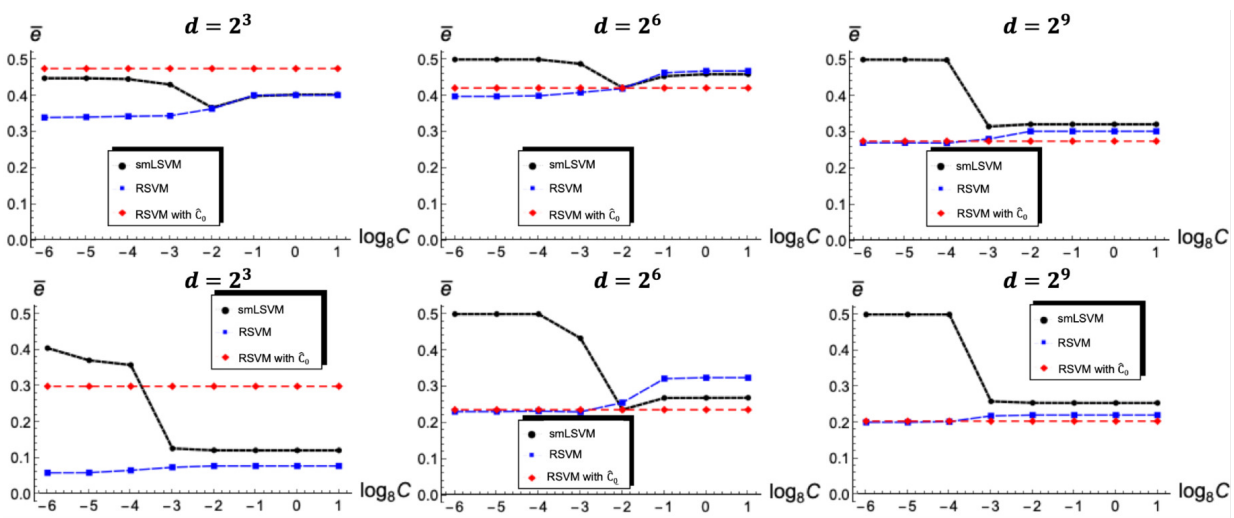


Figure 3.2: The performance of smLSVM and RSVM for C and RSVM (3.22) with $\widehat{C}_0$. The left panels display $d = 2^3$, the middle panels display $d = 2^6$ and the right panels display $d = 2^9$. The upper three panels show the case of (a) and the lower three panels show the case of (b).

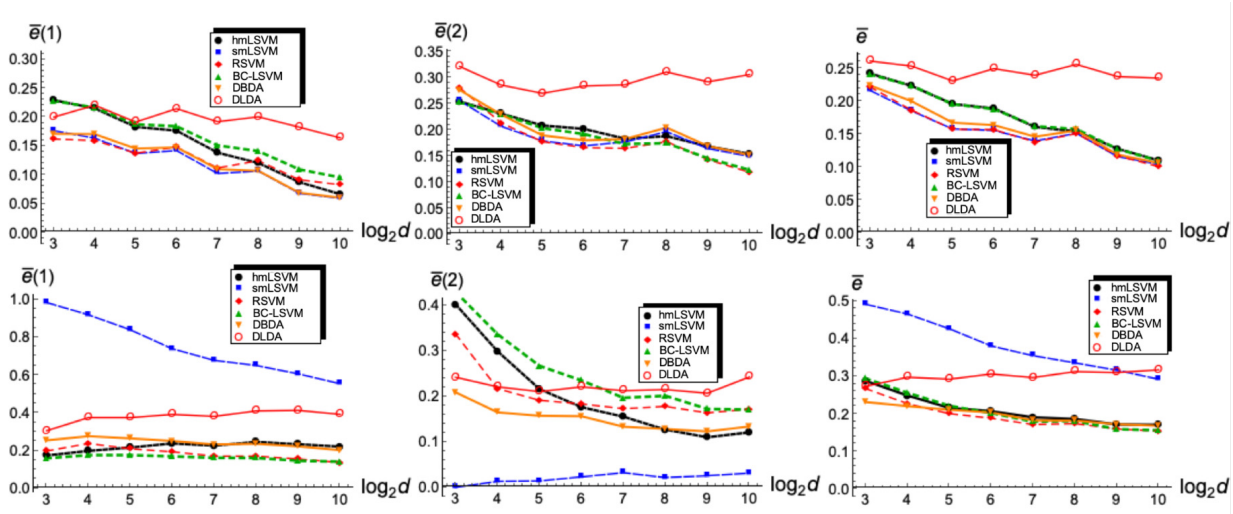


Figure 3.3: The performance of six classifiers, hmLSVM, smLSVM, RSVM (3.22), BC-LSVM, DBDA and DLDA. The left panel displays $\overline{e}(1)$, the middle panel displays $\overline{e}(2)$ and the right panel displays $\overline{e}$. The upper three panels show the case of (c) and the lower three panels show the case of (d).

Similar to (a) and (b), we calculated the error rates, $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$, by 2000 replications and plotted the results for (c) and (d) in Figure 3.3 and only $\overline{e}$ for (e) and (f) in Figure 3.4. In the balanced case (c), all classifiers except DLDA gave good performances. Among them, DLDA only gave the bad performance. We observed that the smLSVM gives quite bad performances for imbalanced settings (d), (e) and (f). The smLSVM was extremely bad when n_1 is small from Figure 3.4. The smLSVM is probably affected by the imbalance of n_is. Thus we do not recommend to use the smLSVM. Also, DLDA gave bad performances for (c) to (f) when n_is are small. On the other hand, The hmLSVM, RSVM, BC-LSVM and DBDA gave adequate performances for all cases (c) to (f). The RSVM and DBDA gave better performances than the hmLSVM even when sample size n_is are small from Figure 3.4. The hmLSVM and BC-LSVM gave the bad performance in the low dimensional case. We observed that the RSVM gives good performances both for low-dimensional and high-dimensional data.

3.4 Real data analysis

In this section, we apply the RSVM to microarray data set by same experiments as Section 2.4.2 in Chapter 2. We used colon cancer data with 2000 ($= d$) genes given by Alon et al. (1999) which consists of π_1 : colon tumor (40 samples) and π_2 : normal colon (22 samples). We randomly split the data sets from (π_1, π_2) into training data sets of sizes (n_1, n_2) and test data sets of sizes $(40 - n_1, 22 - n_2)$. We constructed the hmSVM, smSVM and RSVM (3.22) by using the training data sets. We used $C = \widehat{C}_0$. We checked accuracy by using the test data set for each π_i and denoted the misclassification rates by $\widehat{e}(1)_r$ and $\widehat{e}(2)_r$. We repeated this procedure 100 times and obtained $\widehat{e}(1)_r$ and $\widehat{e}(2)_r$, $r = 1, \ldots, 100$ and we got the average misclassification rates as $\overline{e}(1)$ $(= \sum_{r=1}^{100} \widehat{e}(1)_r/100)$, $\overline{e}(2)$ $(= \sum_{r=1}^{100} \widehat{e}(2)_r/100)$ and $\overline{e}$ $(= \{\overline{e}(1) + \overline{e}(2)\}/2)$ for three classifiers. We summarized the results for various n_is in Table 3.4. In balanced cases, we observed that three classifiers equivalent results. Naive SVMs gave bad performances for imbalanced data. RSVM seems to give good results in such cases.

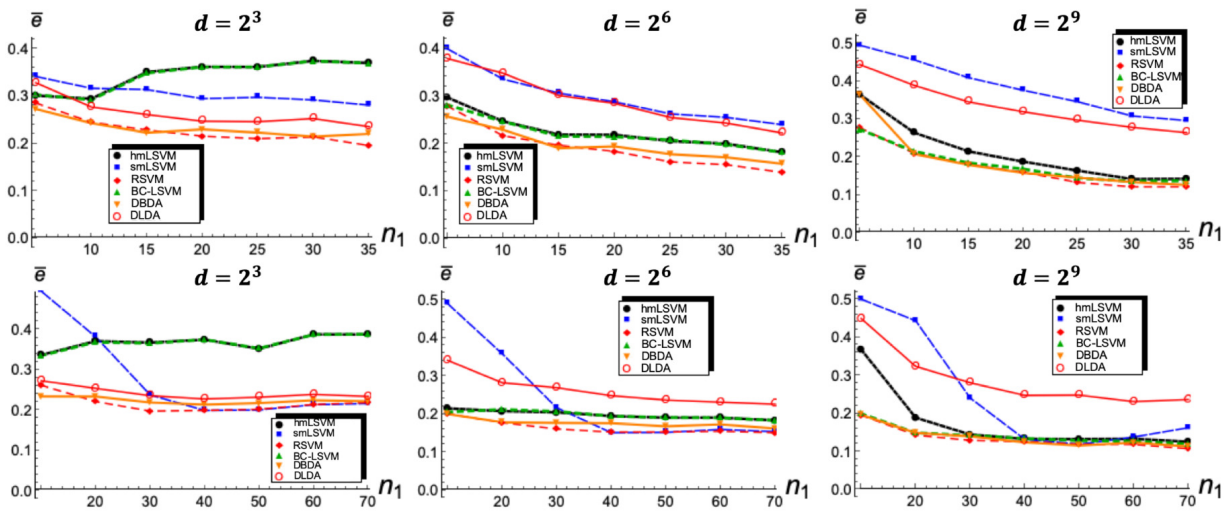


Figure 3.4: The performance of six classifiers, hmLSVM, smLSVM, RSVM (3.22), BC-LSVM, DBDA and DLDA. The left panel displays $d = 2^3$, the middle panel displays $d = 2^6$ and the right panel displays $d = 2^9$. The upper three panels show the case of (e) and the lower three panels show the case of (f).

Appendix 3

Proof of Lemma 3.1. Let $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_* = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Then, by using Chebyshev's inequality, for any $\tau > 0$, under (A-i) and (A-ii), we have that

$$\sum_{j=1}^{n_i} P\Big(\big|\|\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i\|^2 - \mathrm{tr}(\boldsymbol{\Sigma}_i)\big| \geq \tau\Delta_\mu\Big) = O\Big(n_i \mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_\mu^2\Big) \to 0 \tag{3.23}$$

for $i = 1, 2$. Note that $n_i(\boldsymbol{\mu}^T\boldsymbol{\Sigma}_i\boldsymbol{\mu})^2/\Delta_\mu^4 \leq n_i\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_\mu^2 = o(1)$, $i = 1, 2$ under (A-ii). We have that $\mathrm{tr}(\boldsymbol{\Sigma}_i^4) \leq \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^2$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) \leq \{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}^2$. Then, for any $\tau > 0$, under (A-i) and (A-ii), we have that

$$\begin{aligned}
&\sum_{1<j<j'<n_i}^{n_i} P(|(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)^\top(\boldsymbol{x}_{ij'} - \boldsymbol{\mu}_i)| \geq \tau\Delta_\mu) \\
\leq &\sum_{1<j<j'<n_i}^{n_i} (\tau\Delta_\mu)^{-4} E[\{(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)^\top(\boldsymbol{x}_{ij'} - \boldsymbol{\mu}_i)\}^4] = O\Big(n_i^2\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^2 + \mathrm{tr}(\boldsymbol{\Sigma}_i^4)\}/\Delta_\mu^4\Big) \to 0;
\end{aligned} \tag{3.24}$$

$$\begin{aligned}
&\sum_{j=1}^{n_1}\sum_{j'=1}^{n_2} P(|(\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1)^\top(\boldsymbol{x}_{2j'} - \boldsymbol{\mu}_2)| \geq \tau\Delta_\mu) \\
=&O\Big(n_1n_2\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}/\Delta_\mu^4\Big) \to 0; \quad \textit{and}
\end{aligned} \tag{3.25}$$

$$\begin{aligned}
&\sum_{j=1}^{n_i} P(|\boldsymbol{\mu}^T(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)| \geq \tau\Delta_\mu) \\
\leq &n_i(\tau\Delta_\mu)^{-4}E[\{\boldsymbol{\mu}^T(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)\}^4] = O\Big\{n_i\Big((\boldsymbol{\mu}^T\boldsymbol{\Sigma}_i\boldsymbol{\mu})^2 + \sum_{r=1}^{p_i}(\boldsymbol{\gamma}_{ir}^T\boldsymbol{\mu})^4\Big)/\Delta_\mu^4\Big\} \to 0
\end{aligned} \tag{3.26}$$

Table 3.1: Average misclassification rates of the hmLSVM, smLSVM and RSVM.

	hmLSVM			smLSVM			RSVM		
(n_1, n_2)	$\overline{e}$	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}$	$\overline{e}(1)$	$\overline{e}(2)$	$\overline{e}$	$\overline{e}(1)$	$\overline{e}(2)$
$(5, 5)$	0.205	0.193	0.216	0.204	0.202	0.205	0.209	0.202	0.215
$(10, 5)$	0.211	0.137	0.286	0.331	0.062	0.6	0.208	0.176	0.241
$(20, 5)$	0.242	0.111	0.374	0.479	0.013	0.945	0.214	0.17	0.259
$(30, 5)$	0.251	0.093	0.409	0.494	0.002	0.986	0.191	0.166	0.216
$(5, 10)$	0.197	0.266	0.128	0.296	0.55	0.042	0.181	0.211	0.15
$(10, 10)$	0.151	0.156	0.147	0.158	0.168	0.148	0.158	0.166	0.151
$(20, 10)$	0.172	0.133	0.211	0.228	0.078	0.378	0.143	0.129	0.158
$(30, 10)$	0.175	0.119	0.232	0.355	0.044	0.666	0.145	0.146	0.144
$(5, 15)$	0.209	0.302	0.116	0.394	0.777	0.011	0.195	0.229	0.161
$(10, 15)$	0.173	0.179	0.166	0.184	0.247	0.12	0.154	0.146	0.161
$(20, 15)$	0.15	0.14	0.16	0.139	0.109	0.169	0.136	0.133	0.139
$(30, 15)$	0.145	0.152	0.139	0.165	0.099	0.23	0.127	0.139	0.114

from the fact that $\sum_{r=1}^{p_i}(\boldsymbol{\gamma}_r^T\boldsymbol{\mu})^4 \leq (\boldsymbol{\mu}^T\boldsymbol{\Sigma}_i\boldsymbol{\mu})^2$, where $\boldsymbol{\Gamma}_i = [\boldsymbol{\gamma}_{i1}, \ldots, \boldsymbol{\gamma}_{ip_i}]$ for $i = 1, 2$. Similarly, under (A-ii), we obtain that

$$(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)^T(\boldsymbol{x}_0 - \boldsymbol{\mu}_{i'})/\Delta_\mu = (-1)^{i+i'}/4 + o_P(1) \tag{3.27}$$

when $\boldsymbol{x}_0 \in \pi_{i'}$ for $j = 1, \ldots, n_i$. Note that

$$\begin{aligned}(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_{i'j'} - \boldsymbol{\mu}_*) &= (\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)^T(\boldsymbol{x}_{i'j'} - \boldsymbol{\mu}_{i'}) + (-1)^{i+i'}\Delta_\mu/4 \\ &\quad + (-1)^{i'+1}\boldsymbol{\mu}^T(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)/2 + (-1)^{i+1}\boldsymbol{\mu}^T(\boldsymbol{x}_{i'j'} - \boldsymbol{\mu}_{i'})/2\end{aligned} \tag{3.28}$$

for $j = 1, \ldots, n_i,\ j' = 1, \ldots, n_{i'}$ and $i, i' = 1, 2$. From (3.23) to (3.27), we have that

$$\begin{aligned}&\sum_{j=1}^n t_j\alpha_j\boldsymbol{x}_j^T\boldsymbol{x}_0 - \frac{1}{n}\sum_{j=1}^n\sum_{j'=1}^n \alpha_j t_j \boldsymbol{x}_j^T\boldsymbol{x}_{j'} \\ =&\sum_{j=1}^n t_j\alpha_j(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_0 - \boldsymbol{\mu}_*) - \frac{1}{n}\sum_{j=1}^n\sum_{j'=1}^n \alpha_j t_j(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_{j'} - \boldsymbol{\mu}_*) \\ =&\alpha_\star\Big((-1)^i\Delta_\mu/2 + \frac{(n_1 - n_2)\Delta_\mu}{2n} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)}{n}\Big) + o_P(\alpha_\star\Delta_\mu) \\ =&\frac{(-1)^i n_{i'}\alpha_\star\Delta_\mu}{n} + \frac{\alpha_\star\{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}}{n} + o_P(\alpha_\star\Delta_\mu)\end{aligned}$$

under (3.2) for $i \neq i'$. Hence, we conclude the result. □

Proof of Lemma 3.2. Similar to Lemma 2.1 of Chapter 2, we can write for (2.3) that

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^n \alpha_j - \frac{1}{2}\sum_{j=1}^n\sum_{j'=1}^n \alpha_j\alpha_{j'}t_jt_{j'}(\boldsymbol{x}_j - \boldsymbol{\mu}_*)^T(\boldsymbol{x}_k - \boldsymbol{\mu}_*).$$

By noting that $\alpha_j \geq 0$ subject to (3.2), from (3.23) to (3.26) and (3.28), we can conclude the result. □

Proof of Lemma 3.3. From the fact that (3.9) holds when $\alpha_{ij} = \alpha_\star/n_i,\ j = 1, \ldots, n_i,\ i = 1, 2$ and $\alpha_\star \approx 2/\Delta_{\mu*}$ maximizes (3.9), we can obtain the first result. By combining Lemma 3.1 with $\alpha_\star = (2/\Delta_{\mu*})\{1 + o_P(1)\}$, we can conclude the second result. □

Proof of Lemma 3.4. From (3.10), the first result is obtained straightforwardly. By combining Lemma 3.1 with $\alpha_\star = (2/\Delta_\mu)\{1 + o_P(1)\}$, we can conclude the second result. □

Proof of Lemma 3.5. From (3.9) and (3.13), we can maximize $L(\boldsymbol{\alpha})$ by choosing $\alpha_\star \approx n_1 A$ under (3.12). Hence, we conclude the first result.

Note that $\tilde{S} = \{1, \ldots, n\}$. Then we have that

$$\hat{y}_{(S)}(\boldsymbol{x}_0) = \sum_{j=1}^{n} \tilde{\alpha}_j t_j \boldsymbol{x}_j^T \boldsymbol{x}_0 - \frac{1}{n} \sum_{j=1}^{n} \sum_{j'=1}^{n} \tilde{\alpha}_{j'} t_{j'} \boldsymbol{x}_j^T \boldsymbol{x}_{j'} + \frac{n_2 - n_1}{n} \tag{3.29}$$

By combining Lemma 3.1 with (3.29) and $\alpha_\star = n_1 C\{1 + o_P(1)\}$, we can conclude the second result. □

Proof of Theorem 3.1 and Corollary 3.1. By using Lemma 3.5, the results are obtained straightforwardly. □

Proof of Lemma 3.6. Similar to the proof of Lemma 3.5, by choosing $\alpha_\star \approx n_1 C$, we have the maximum of $L(\boldsymbol{\alpha})$ asymptotically under z(3.12) from (3.10) and (3.13). It holds that $\alpha_{1j} = C\{1+o_P(1)\}$, $j = 1, \ldots, n_1$ from (3.3) and $\alpha_\star = n_1 C\{1 + o_P(1)\}$. Hence, we conclude the first result.

By combining Lemma 3.1 with $\alpha_\star = n_1 C\{1 + o_P(1)\}$, we have that

$$\hat{y}_{(S)}(\boldsymbol{x}_0) = \frac{n_1 C}{n_{\tilde{S}}} \Big((-1)^i n_{\tilde{S}_{i'}} \Delta_\mu + \mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2) \Big) \{1 + o_P(1)\} + \frac{n_{\tilde{S}_2} - n_1}{n_{\tilde{S}}}$$

when $\boldsymbol{x}_0 \in \pi_i$ for $i \neq i'$. By noting that $\mathrm{tr}(\boldsymbol{\Sigma}_1)/\{n_i \Delta_\mu\} \leq \mathrm{tr}(\boldsymbol{\Sigma}_1)/\{n_1 \Delta_\mu\} = o(1)$ under (3.7), we can obtain the second result. □

Proof of Theorem 3.2 Corollary 3.2. We can obtain the results from Lemma 3.6. □

Proof of Theorem 3.3 and Corollary 3.3. We can write that

$$\begin{aligned}
& y_r(\boldsymbol{x}_0) \\
&= -\sum_{j=1}^{n_1} \alpha_{1j} (\boldsymbol{x}_{1j} - \boldsymbol{\mu})^T (\boldsymbol{x}_0 - \boldsymbol{\mu}) + \sum_{j=1}^{n_2} \alpha_{2j} (\boldsymbol{x}_{2j} - \boldsymbol{\mu})^T (\boldsymbol{x}_0 - \boldsymbol{\mu}) \\
&+ \sum_{j \neq j'} \alpha_{1j} (\boldsymbol{x}_{1j} - \boldsymbol{\mu})^T (\boldsymbol{x}_{1j'} - \boldsymbol{\mu}) / \{2(n_1 - 1)\} - \sum_{j \neq j'} \alpha_{2j} (\boldsymbol{x}_{2j} - \boldsymbol{\mu})^T (\boldsymbol{x}_{2j'} - \boldsymbol{\mu}) / \{2(n_2 - 1)\} \\
&- \sum_{j=1}^{n_1} \alpha_{1j} (\boldsymbol{x}_{1j} - \boldsymbol{\mu})^T \boldsymbol{\mu}/2 + \sum_{j \neq j'} \alpha_{1j} (\boldsymbol{x}_{1j'} - \boldsymbol{\mu})^T \boldsymbol{\mu} / \{2(n_1 - 1)\} \\
&+ \sum_{j=1}^{n_2} \alpha_{2j} (\boldsymbol{x}_{2j} - \boldsymbol{\mu})^T \boldsymbol{\mu}/2 - \sum_{j \neq j'} \alpha_{2j} (\boldsymbol{x}_{2j'} - \boldsymbol{\mu})^T \boldsymbol{\mu} / \{2(n_2 - 1)\}.
\end{aligned}$$

Then, by combining (3.24) to (3.27), under (A-i) to (A-iii), we can obtain the results of Theorem 3.3. By combining Lemma 3.3 to 3.6 with Theorem 3.3, we conclude Corollary 3.3. □

Chapter 4

Nonlinear SVM by kernel functions in the HDLSS context

Even if the smSVM is used, it is not always possible to construct a classifier having excellent performance for a nonlinear classification task. As a method for dealing with it, a method called "kernel trick" is known. It is that an original vector is subjected to nonlinear transformation to perform linear discrimination in a feature space. By using this method, the performance of the SVM has been dramatically improved. There are many studies on the LSVM for high-dimensional data. As long as we know, asymptotic properties of nonlinear SVMs with kernel functions seem not to have been sufficiently studied in the HDLSS context. In this chapter, we consider nonlinear SVMs with kernel functions in the HDLSS context. This chapter is organized by Nakayama et al. (2019).

In Section 4.2, we consider nonlinear SVMs in a general framework and study their asymptotic properties in the HDLSS context. We show that nonlinear SVMs are heavily biased in the HDLSS context, especially for imbalanced data. In order to overcome such difficulties, we propose a bias-corrected SVM (BC-SVM).

In Section 4.3, we give asymptotic properties of the BC-SVM both for the linear and Gaussian kernels. We show that the BC-SVM with the Gaussian kernel draws information about heteroscedasticity thorough the geometric representation of expanding two spheres having different radii, $\mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}$s.

In Section 4.4, we show that the performance of the BC-SVM is influenced by the scale parameter involved in the Gaussian kernel. We discuss a choice of the scale parameter yielding a high performance.

In Section 4.5, we examine the performance of the BC-SVM with the Gaussian kernel for several choices of the scale parameter by numerical simulations and real data analyses.

In Section 4.6, we give the performance of the BC-SVM even in the soft-margin.

4.1 Introduction

Suppose we have two independent populations, $\pi_i,\ i=1,2$, having a d-variate distribution with unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i$. We do not specify any distributional function for π_i. We have independent and identically distributed (i.i.d.) observations, $\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{in_i}$, from each π_i. We assume $n_i \geq 2$. Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to one of the π_is. We assume $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent. Let $n=n_1+n_2$. We consider the HDLSS context in which $d\to\infty$ while n is fixed or $n/d\to 0$ as $d,n\to\infty$.

We introduce a high-dimensional geometric representation. Let us consider the following condition for $\boldsymbol{\Sigma}_i,\ i=1,2$:

$$\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\mathrm{tr}(\boldsymbol{\Sigma}_i)^2 \to 0 \text{ as } d\to\infty. \tag{4.1}$$

We note that the ratio, $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\mathrm{tr}(\boldsymbol{\Sigma}_i)^2$, is a measure of sphericity and (4.1) is equivalent to "$\lambda_{\max}(\boldsymbol{\Sigma}_i)/\mathrm{tr}(\boldsymbol{\Sigma}_i)\to 0$ as $d\to\infty$", where $\lambda_{\max}(\boldsymbol{\Sigma}_i)$ denotes the largest eigenvalue of $\boldsymbol{\Sigma}_i$. See Ahn et al. (2007) and Aoshima and Yata (2019a). If we assume (4.1) and (A-ii) given in Section 4.3, we have that

$$\|\boldsymbol{x}_0-\boldsymbol{\mu}_i\| = \mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}\{1+o_P(1)\} \text{ as } d\to\infty \text{ when } \boldsymbol{x}_0\in\pi_i$$

from the fact that $\mathrm{Var}(\|\boldsymbol{x}_0-\boldsymbol{\mu}_i\|^2)=O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}$ when $\boldsymbol{x}_0\in\pi_i$, where $\|\cdot\|$ denotes the Euclidean norm. Thus, the centroid data concentrate near on the surface of an expanding sphere with radius, $\mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}$, when the dimension is large. See Hall et al. (2005) for the details of the geometric representation. We consider a toy example to see the geometric representation. We set $\pi_i : N_d(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i),\ i=1,2$, having $\boldsymbol{\Sigma}_1=\boldsymbol{I}_d$ and $\boldsymbol{\Sigma}_2=2\boldsymbol{I}_d$, where $\boldsymbol{I}_d$ denotes the d-dimensional identity matrix. Note that (4.1) and (A-ii) are met. Thus, for a large d, we expect that $\|\boldsymbol{x}_0-\boldsymbol{\mu}_1\|/d^{1/2}\approx 1$ when $\boldsymbol{x}_0\in\pi_1$ and $\|\boldsymbol{x}_0-\boldsymbol{\mu}_2\|/d^{1/2}\approx 2^{1/2}$ when $\boldsymbol{x}_0\in\pi_2$. Independent pseudorandom 2000 observations of $\|\boldsymbol{x}_0-\boldsymbol{\mu}_i\|/d^{1/2}$ were generated when $\boldsymbol{x}_0\in\pi_i$ for $i=1,2$. In Figure 4.1, we gave histograms of $\|\boldsymbol{x}_0-\boldsymbol{\mu}_i\|/d^{1/2}$ for $\boldsymbol{x}_0\in\pi_i,\ i=1,2$, when $d=16,80,400$ and 2000. We observed that $\|\boldsymbol{x}_0-\boldsymbol{\mu}_i\|/d^{1/2}$s converge to $\mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}/d^{1/2}$ for each case as d increases. In other words, $\boldsymbol{x}_0$ concentrates on the surface of the d-dimensional sphere with centre $\boldsymbol{\mu}_i$ and radius $\mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}$ as in Figure 4.2. In this chapter, we focus on the geometric representation for high-dimensional classification.

Let $e(i)$ denote the error rate of misclassifying an individual from π_i into the other class for $i=1,2$. We claim that a classifier has the consistency if

$$e(i)\to 0 \ \text{ as } d\to\infty \text{ for } i=1,2. \tag{4.2}$$

In this paper, we mainly investigate the following typical kernels.

(I) The linear kernel: $k(\boldsymbol{x}_j,\boldsymbol{x}_{j'})=\boldsymbol{x}_j^T\boldsymbol{x}_{j'}$ and
(II) The Gaussian kernel: $k(\boldsymbol{x}_j,\boldsymbol{x}_{j'})=\exp(-\|\boldsymbol{x}_j-\boldsymbol{x}_{j'}\|^2/\gamma)$,

where $\gamma(>0)$ is a scale parameter. In addition, we discuss a choice of γ in Section 4.4. We examine the following kernels numerically.

(III) The polynomial kernel: $k(\boldsymbol{x}_j,\boldsymbol{x}_{j'})=(\zeta+\boldsymbol{x}_j^T\boldsymbol{x}_{j'})^r$ and
(IV) The Laplace kernel: $k(\boldsymbol{x}_j,\boldsymbol{x}_{j'})=\exp(-\|\boldsymbol{x}_j-\boldsymbol{x}_{j'}\|_1/\xi)$,

where $\zeta\geq 0$, $\xi>0$, $r\in\mathbb{N}$ and $\|\cdot\|_1$ denotes the L_1-norm.

4.2 Nonlinear SVM in HDLSS settings

In this section, we consider the SVM in a general framework. We give asymptotic properties of the SVM under the following divergence condition:

$$d\to\infty \text{ either when } n\to\infty \text{ as } d\to\infty \text{ or } n \text{ is fixed.}$$

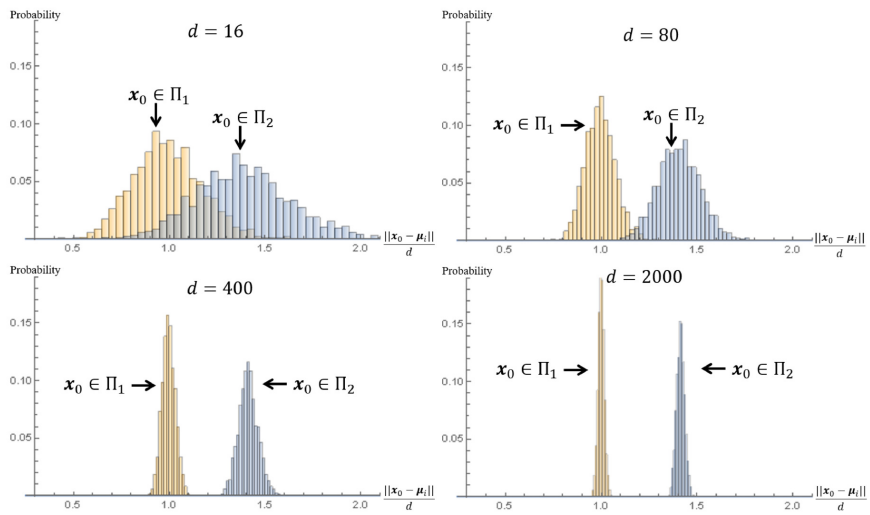


Figure 4.1: The histograms of $\|\boldsymbol{x}_0 - \boldsymbol{\mu}_i\|/d^{1/2}$ for $\boldsymbol{x}_0 \in \pi_i$, $i = 1, 2$, when $d = 16, 80, 400$ and 2000.

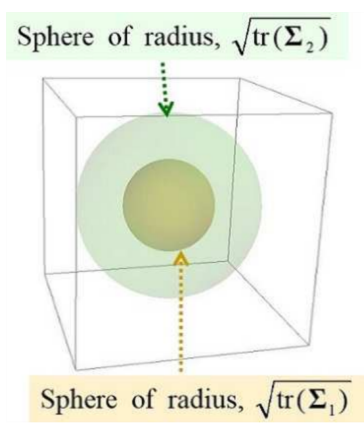


Figure 4.2: The geometric representation of expanding two spheres having different radii, $\mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}$s.

4.2.1 Setup of nonlinear SVM

Since HDLSS data are mostly separable by a hyperplane, we first consider the hard-margin SVM (hmSVM):

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + b, \tag{4.3}$$

where $\phi(\cdot)$ is a feature map, $\boldsymbol{w}$ is a weight vector and b is an intercept term. Let us write that $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}, \boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2})$. Let $t_j = -1$ for $j = 1, \ldots, n_1$ and $t_j = 1$ for $j = n_1 + 1, \ldots, n$. By differentiating the Lagrangian formulation with respect to $\boldsymbol{w}$ and b, we obtain the following dual form:

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^{n} \alpha_j - \frac{1}{2} \sum_{j=1}^{n} \sum_{j'=1}^{n} \alpha_j \alpha_{j'} t_j t_{j'} k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}), \tag{4.4}$$

where $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \phi(\boldsymbol{x}_j)^T \phi(\boldsymbol{x}_{j'})$ is a kernel function, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ and α_js are Lagrange multipliers such as $\boldsymbol{w} = \sum_{j=1}^{n} \alpha_j t_j \phi(\boldsymbol{x}_j)$. The optimization problem can be transformed into the following: $\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L(\boldsymbol{\alpha})$ subject to

$$\alpha_j \geq 0, \ j = 1, \ldots, n, \quad \text{and} \quad \sum_{j=1}^{n} \alpha_j t_j = 0 \tag{4.5}$$

Let us write that

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)^T = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L(\boldsymbol{\alpha}) \ \text{ subject to (4.5).}$$

Note that $\sum_{j=1}^{n_1} \hat{\alpha}_j = \sum_{j=n_1+1}^{n} \hat{\alpha}_j$. There exist some $\boldsymbol{x}_j$s satisfying that $t_j y(\boldsymbol{x}_j) = 1$ (i.e., $\hat{\alpha}_j \neq 0$). Such $\boldsymbol{x}_j$s are called the support vector. Let $\hat{S} = \{j | \hat{\alpha}_j \neq 0, \ j = 1, \ldots, n\}$ and $n_{\hat{S}} = \#\hat{S}$, where $\#A$ denotes the number of elements in a set A. The intercept term is given by $\hat{b} = n_{\hat{S}}^{-1} \sum_{j \in \hat{S}} \{t_j - \sum_{j' \in \hat{S}} \hat{\alpha}_{j'} t_{j'} k(\boldsymbol{x}_j, \boldsymbol{x}_{j'})\}$. Then, the classifier in (4.3) is given by

$$\hat{y}(\boldsymbol{x}) = \sum_{j \in \hat{S}} \hat{\alpha}_j t_j k(\boldsymbol{x}, \boldsymbol{x}_j) + \hat{b}. \tag{4.6}$$

One classifies $\boldsymbol{x}_0$ into π_1 if $\hat{y}(\boldsymbol{x}_0) < 0$ and into π_2 otherwise. See Vapnik (2000) for the details.

4.2.2 Asymptotic properties of nonlinear SVM

Let $\boldsymbol{K}$ be an $n \times n$ gram matrix with the (j, j') element $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'})$. First, we assume the following assumption for $d \to \infty$ either when n is fixed or $n \to \infty$:

(A-i) $k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j'}) = \kappa_1 + o_P(\Delta_\kappa)$ for all $1 \leq j < j' \leq n_1$,

$k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j}) = \kappa_2 + o_P(\Delta_\kappa)$ for all $1 \leq j \leq n_1$,

$k(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j'}) = \kappa_3 + o_P(\Delta_\kappa)$ for all $1 \leq j < j' \leq n_2$,

$k(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j}) = \kappa_4 + o_P(\Delta_\kappa)$ for all $1 \leq j \leq n_2$,

and $k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j'}) = \kappa_5 + o_P(\Delta_\kappa)$ for all $1 \leq j \leq n_1$ and $1 \leq j' \leq n_2$,

where $\Delta_\kappa = \kappa_1 + \kappa_3 - 2\kappa_5$ and κ_ls are variables (which may depend on d) such that $\Delta > 0$, $\kappa_2 \geq \kappa_1$ and $\kappa_4 \geq \kappa_3$.

Note that (A-i) is regarded as a convergence condition for the gram matrix and Δ_κ is a distance between the two populations. Also, note that κ_is are characteristic variables for each kernel in high-dimensional settings. They are naturally obtained by high-dimensional asymptotics. For example, $\Delta_\kappa = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$,

$\kappa_1 = \|\boldsymbol{\mu}_1\|^2$, $\kappa_2 = \|\boldsymbol{\mu}_1\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1)$, $\kappa_3 = \|\boldsymbol{\mu}_2\|^2$, $\kappa_4 = \|\boldsymbol{\mu}_2\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_2)$ and $\kappa_5 = \boldsymbol{\mu}_1^T \boldsymbol{\mu}_2$ when $k(\cdot,\cdot)$ is the linear kernel. See Section 4.3.1. Also, see Sections 4.3.2 and 4.3.4 for the Gaussian and polynomial kernels, respectively.

Let $\eta_1 = \kappa_2 - \kappa_1$ and $\eta_2 = \kappa_4 - \kappa_3$. We note that $k(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}) = k(\boldsymbol{x}_{ij'}, \boldsymbol{x}_{ij})$ for all $j \neq j'$ $(i = 1, 2)$. Then, under (A-i), we write that

$$\boldsymbol{K}/\Delta_\kappa \approx \begin{pmatrix} \kappa_1 \boldsymbol{J}_{n_1,n_1} + \eta_1 \boldsymbol{I}_{n_1} & \kappa_5 \boldsymbol{J}_{n_1,n_2} \\ \kappa_5 \boldsymbol{J}_{n_2,n_1} & \kappa_3 \boldsymbol{J}_{n_2,n_2} + \eta_2 \boldsymbol{I}_{n_2} \end{pmatrix} / \Delta_\kappa \quad (= \boldsymbol{K}_0/\Delta_\kappa, \text{ say}),$$

where $\boldsymbol{J}_{n_1,n_2}$ denotes the $n_1 \times n_2$ matrix with all the elements 1. Let $\acute{\boldsymbol{\alpha}} = (-\alpha_1, \ldots, -\alpha_{n_1}, \alpha_{n_1+1}, \ldots, \alpha_n)^T$. We note that $\sum_{j=1}^{n_1} \alpha_j = \sum_{j=n_1+1}^{n} \alpha_j$ $(= \alpha_\star,$ say) under (4.5). Then, it holds that

$$\acute{\boldsymbol{\alpha}}^T \boldsymbol{K}_0 \acute{\boldsymbol{\alpha}} = \Delta_\kappa \alpha_\star^2 + \eta_1 \sum_{j=1}^{n_1} \alpha_j^2 + \eta_2 \sum_{j=n_1+1}^{n} \alpha_j^2. \tag{4.7}$$

The second and third terms in (4.4) are regarded as a bias part. See Proposition 4.1. We have that $L(\boldsymbol{\alpha}) = 2\alpha_\star - \acute{\boldsymbol{\alpha}}^T \boldsymbol{K} \acute{\boldsymbol{\alpha}}/2$ under (4.5). Then, from (4.7) we claim the following lemma.

Lemma 4.1. *Under (4.5) and (A-i), it holds that*

$$L(\boldsymbol{\alpha}) = 2\alpha_\star - \frac{\Delta_\kappa}{2}\alpha_\star^2 - \frac{1}{2}\Big(\eta_1 \sum_{j=1}^{n_1} \alpha_j^2 + \eta_2 \sum_{j=n_1+1}^{n} \alpha_j^2\Big) + o_P(\Delta_\kappa \alpha_\star^2).$$

Note that

$$\min_{\boldsymbol{\alpha}} \eta_1 \sum_{j=1}^{n_1} \alpha_j^2 = \alpha_\star^2 \eta_1/n_1 \quad \text{and} \quad \min_{\boldsymbol{\alpha}} \eta_2 \sum_{j=n_1+1}^{n} \alpha_j^2 = \alpha_\star^2 \eta_2/n_2$$

when $\alpha_1 = \cdots = \alpha_{n_1} = \alpha_\star/n_1$ and $\alpha_{n_1+1} = \cdots = \alpha_n = \alpha_\star/n_2$ under (4.5). We first consider the following condition:

$$\liminf_{d\to\infty} \frac{\eta_i}{n_i \Delta_\kappa} > 0 \quad \text{for } i = 1, 2. \tag{4.8}$$

Let $\Delta_{\kappa*} = \Delta_\kappa + \eta_1/n_1 + \eta_2/n_2$. Note that $2\alpha_\star - \Delta_{\kappa*}\alpha_\star^2/2 = -\Delta_{\kappa*}(\alpha_\star - 2/\Delta_{\kappa*})^2/2 + 2/\Delta_{\kappa*}$. Then, in a way similar to Section 2.2 in Chapter 2, it follows from Lemma 4.1 that

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = -\frac{\Delta_{\kappa*}}{2}\Big(\alpha_\star - \frac{2 + o_P(1)}{\Delta_{\kappa*}}\Big)^2 \{1 + o_P(1)\} + \frac{2 + o_P(1)}{\Delta_{\kappa*}}$$

under (4.5), (4.8) and (A-i), so that $\alpha_\star \approx 2/\Delta_{\kappa*}$. Let $\hat{\alpha}_\star = \sum_{j=1}^{n_1} \hat{\alpha}_j$. Note that $\sum_{j=n_1+1}^{n} \hat{\alpha}_j = \hat{\alpha}_\star$.

Proposition 4.1. *Assume (A-i) and (4.8). It holds that*

$$\begin{aligned} &\hat{\alpha}_\star = (2/\Delta_{\kappa*})\{1 + o_P(1)\}, \\ &\sum_{j=1}^{n_1} \hat{\alpha}_j^2 = \frac{4}{\Delta_{\kappa*}^2 n_1}\{1 + o_P(1)\} \text{ and } \sum_{j=n_1+1}^{n} \hat{\alpha}_j^2 = \frac{4}{\Delta_{\kappa*}^2 n_2}\{1 + o_P(1)\}. \end{aligned} \tag{4.9}$$

We also assume

(A-i') $k(\boldsymbol{x}_0, \boldsymbol{x}_{ij}) = \kappa_{2i-1} + o_P(\Delta_\kappa)$ *for all* $1 \le j \le n_i$ *and* $k(\boldsymbol{x}_0, \boldsymbol{x}_{i'j}) = \kappa_5 + o_P(\Delta_\kappa)$ *for all* $1 \le j \le n_{i'}$ *when* $\boldsymbol{x}_0 \in \pi_i$ *for* $i = 1, 2;$ $i' \neq i$.

It holds that

$$\hat{y}(\boldsymbol{x}_0) = \frac{\Delta_\kappa}{\Delta_{\kappa*}}\left((-1)^i + \frac{\delta}{\Delta_\kappa} + o_P(1)\right) \quad \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2, \tag{4.10}$$

where $\delta = \eta_1/n_1 - \eta_2/n_2$.

We note that "δ/Δ_κ" is a (normalized) bias term of the SVM. From Proposition 4.1, under (A-i) and (4.8), it holds that $\sum_{j=1}^{n_1}(\hat{\alpha}_j - \hat{\alpha}_\star/n_1)^2 = o_P\{(n_1\Delta_{\kappa*}^2)^{-1}\}$ and $\sum_{j=n_1+1}^{n}(\hat{\alpha}_j - \hat{\alpha}_\star/n_2)^2 = o_P\{(n_2\Delta_{\kappa*}^2)^{-1}\}$, so that

$$\begin{aligned} \hat{\alpha}_j &= \frac{2}{\Delta_{\kappa*}n_1}\{1 + o_P(1)\} \quad \text{for all } j = 1, \ldots, n_1; \quad \text{and} \\ \hat{\alpha}_j &= \frac{2}{\Delta_{\kappa*}n_2}\{1 + o_P(1)\} \quad \text{for all } j = n_1 + 1, \ldots, n \end{aligned} \tag{4.11}$$

when $d \to \infty$ while n is fixed. It should be noted that all the data points are support vectors under (A-i) and (4.8) in the HDLSS context. Ahn and Marron (2010) called this phenomenon the "data piling".

Next, we consider the following condition instead of (4.8) under $d \to \infty$ either when $n \to \infty$:

$$\frac{\eta_i}{n_i\Delta_\kappa} = o(1) \quad \text{for } i = 1, 2. \tag{4.12}$$

It follows from Lemma 4.1 that

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = -\frac{\Delta_\kappa}{2}\left(\alpha_\star - \frac{2 + o_P(1)}{\Delta_\kappa}\right)^2\{1 + o_P(1)\} + \frac{2 + o_P(1)}{\Delta_\kappa} \tag{4.13}$$

under (4.5), (4.12) and (A-i), so that $\alpha_\star \approx 2/\Delta_\kappa$.

Proposition 4.2. *Assume (A-i) and (4.12). It holds that* $\hat{\alpha}_\star = (2/\Delta)\{1 + o_P(1)\}$. *Furthermore, we assume (A-i'). It holds that*

$$\hat{y}(\boldsymbol{x}_0) = (-1)^i + o_P(1) \quad \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2. \tag{4.14}$$

It should be noted that the data piling does not occur under (4.12). However, $\hat{y}(\boldsymbol{x}_0)$ has the consistency in the sense of (4.14). We consider the following condition:

(C-i) $\limsup_{d\to\infty} \frac{|\delta|}{\Delta_\kappa} < 1.$

Note that (C-i) is met under (4.12). From Proposition 4.1, "δ/Δ_κ" is a normalized bias term of the SVM. From (4.10), if (C-i) is met, it holds that $P\{(-1)^i\hat{y}(\boldsymbol{x}_0) > 0\} \to 1$ when $\boldsymbol{x}_0 \in \pi_i$ under (A-i) and (A-i'). Thus we have the following result.

Theorem 4.1. *Under (A-i), (A-i') and (C-i), the SVM holds the consistency (4.2).*

However, without (C-i), we have the following results.

Corollary 4.1. *Under (A-i) and (A-i'), the SVM holds the following properties:*

$$e(1) = 1 + o(1) \textit{ and } e(2) = o(1) \textit{ as } d \to \infty \tag{4.15}$$

$$\textit{if } \liminf_{d\to\infty} \frac{\delta}{\Delta_\kappa} > 1, \textit{ and}$$

$$e(1) = o(1) \textit{ and } e(2) = 1 + o(1) \textit{ as } d \to \infty \tag{4.16}$$

$$\textit{if } \limsup_{d\to\infty} \frac{\delta}{\Delta_\kappa} < -1.$$

Remark 12. *For the LSVM, Hall et al. (2005), Qiao and Zhang (2015) and Nakayama et al. (2017) showed the consistency (4.2) and the results in Corollary 4.1.*

From Corollary 4.1, if $|\delta|$ is larger than Δ_κ, the SVM would give a bad performance. When $n_i/n_{i'} \to 0$ for some i ($\neq i'$), $|\delta|$ tends to become large. Such an imbalanced data is called the "extremely imbalanced data". In such cases, the SVM brings the strong inconsistency property as "$e(1) = 1 + o(1)$" when $\eta_1 = \eta_2$, $\Delta_\kappa/\eta_i = o(1)$ and n_1 is fixed but $n_2 \to \infty$. In order to overcome such difficulties, we propose a bias-corrected SVM under (A-i) and (A-i').

4.2.3 Bias-corrected nonlinear SVM

Let

$$\hat{\eta}_i = \sum_{j=1}^{n_i} \frac{k(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij})}{n_i - 1} - \sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \frac{k(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'})}{n_i(n_i - 1)} \quad \text{for } i = 1, 2; \text{ and}$$

$$\widehat{\Delta}_{\kappa *} = \sum_{i=1}^{2} \left(\sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \frac{k(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'})}{n_i^2} \right) - 2 \sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} \frac{k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j'})}{n_1 n_2}.$$

We consider estimating Δ_κ and δ as $\widehat{\Delta}_\kappa = \widehat{\Delta}_{\kappa *} - \hat{\eta}_1/n_1 - \hat{\eta}_2/n_2$ and $\hat{\delta} = \hat{\eta}_1/n_1 - \hat{\eta}_2/n_2$. We have the following lemma.

Lemma 4.2. *Under (A-i) it holds that*

$$\widehat{\Delta}_\kappa/\Delta_\kappa = 1 + o_P(1) \quad \textit{and} \quad \hat{\delta}/\widehat{\Delta}_{\kappa *} = \delta/\Delta_{\kappa *} + o_P\big(\Delta_\kappa/\Delta_{\kappa *}\big).$$

From Proposition 4.1 and Lemma 4.2, we give a bias-corrected SVM (BC-SVM) as follows:

$$\hat{y}_{BC}(\boldsymbol{x}_0) = \hat{y}(\boldsymbol{x}_0) - \frac{\hat{\delta}}{\widehat{\Delta}_{\kappa *}}. \tag{4.17}$$

One classifies $\boldsymbol{x}_0$ into π_1 if $\hat{y}_{BC}(\boldsymbol{x}_0) < 0$ and into π_2 otherwise. We have the following result.

Theorem 4.2. *Under (A-i) and (A-i'), the BC-SVM holds the consistency (4.2).*

It should be noted that the BC-SVM (4.17) claims the consistency without (C-i) even when $|\delta/\Delta_\kappa| \to \infty$.

For imbalanced cases, Benjamin and Nathalie (2010) proposed the boosting SVM. There are several studies on SVMs in imbalanced cases. See He and Garcia (2009) for the review. However, it should be noted that they are algorithmic methods. On the other hand, the BC-SVM (4.17) can theoretically ensure the accuracy and have the consistency property at a low computational cost even for extremely imbalanced data.

Remark 13. *In Chapter 2, we gave the BC-LSVM. In this chapter, we generalize the concept of the BC-LSVM to nonlinear kernels.*

4.2.4 Performances of bias-corrected nonlinear SVM

We set $\pi_i : N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$, having $\boldsymbol{\mu}_2 = \boldsymbol{0}$, $\boldsymbol{\Sigma}_1 = c_1 \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_2 = c_2 \boldsymbol{B}(0.4^{|i-j|^{1/3}})\boldsymbol{B}$, where $\boldsymbol{B} = \mathrm{diag}[\{0.5 + 1/(d+1)\}^{1/2}, \ldots, \{0.5 + d/(d+1)\}^{1/2}]$. Note that $\mathrm{tr}(\boldsymbol{\Sigma}_i) = c_i d$ for $i = 1, 2$. We considered

$$\boldsymbol{\mu}_1 = (-1/5, 1/5, -1/5, \ldots, -1/5, 1/5)^T \ (= \boldsymbol{\mu}_\alpha, \text{ say}),$$

where the r-element is $(-1)^r/5$ for $r = 1, \ldots, d$.

First, we considered LSVM and the Gaussian kernel SVM (GSVM). We compared the performance of the BC-LSVM and bias-corrected GSVM (BC-GSVM) with the above ones. See (4.3.1) and (4.3.2) for the BC-LSVM and BC-GSVM. We set $(n_1, n_2) = (20, 10)$, $d = 2^s$, $s = 5, \ldots, 12$, and $\gamma = d/4$ in the Gaussian kernel (II). We considered three cases:

(a) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_\alpha$ and $(c_1, c_2) = (1, 1)$,
(b) $\boldsymbol{\mu}_1 = \mathbf{0}$ and $(c_1, c_2) = (0.9, 1.1)$, and
(c) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_\alpha$ and $(c_1, c_2) = (0.9, 1.1)$.

Note that $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 = d/25$ for (a) and (c), $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 = 0$ for (b), $|\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)| = 0$ for (a), and $|\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)| = 0.2d$ for (b) and (c). We repeated 2000 times to confirm if the classifier does (or does not) classify $\boldsymbol{x}_0 \in \pi_i$ correctly and defined $P_{ir} = 0$ (or 1) accordingly for each π_i $(i = 1, 2)$. We calculated the error rates, $\overline{e}(i) = \sum_{r=1}^{2000} P_{ir}/2000$, $i = 1, 2$. Also, we calculated the average error rate, $\overline{e} = \{\overline{e}(1) + \overline{e}(2)\}/2$. Their standard deviations are less than 0.0112 from the fact that $\mathrm{Var}\{\overline{e}(i)\} = e(i)\{1 - e(i)\}/2000 \leq 1/8000$. In Figure 4.3, we plotted $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ for $d = 2^s$, $s = 5, \ldots, 12$.

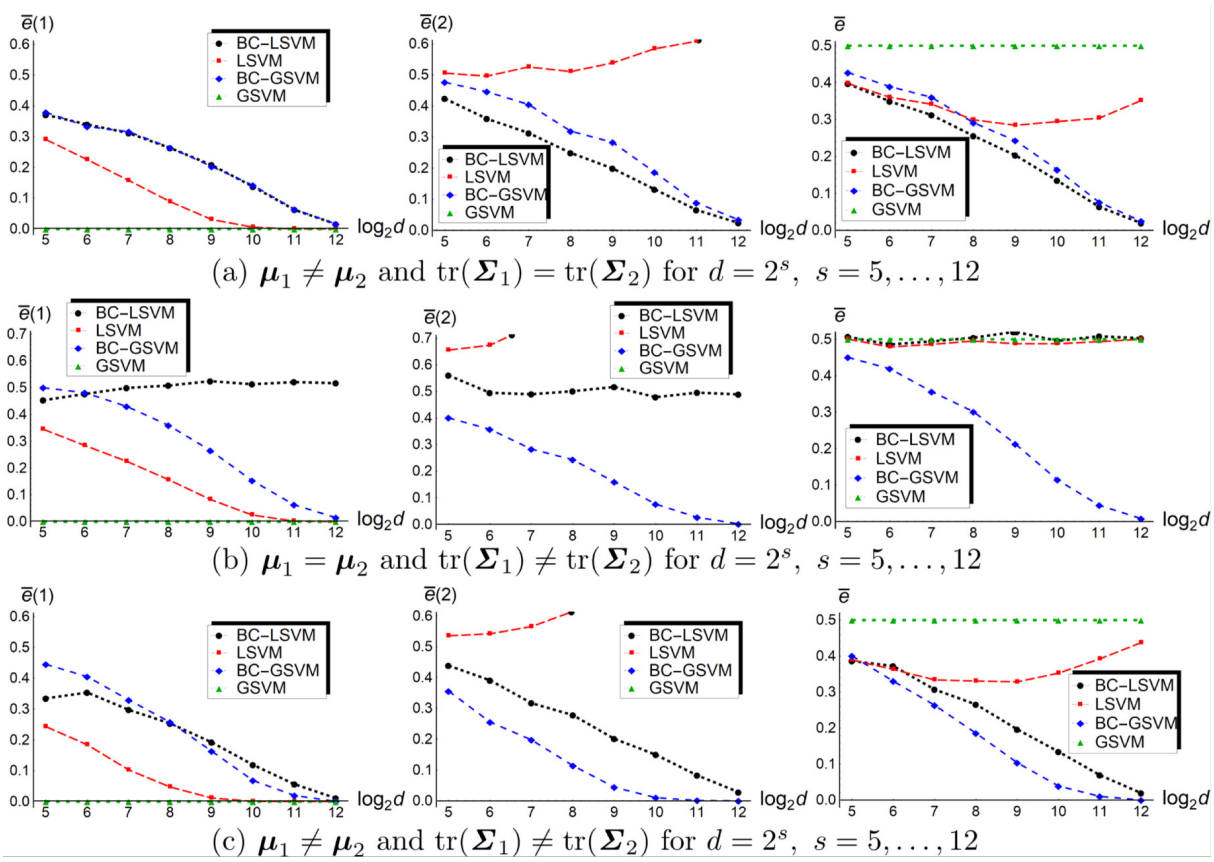


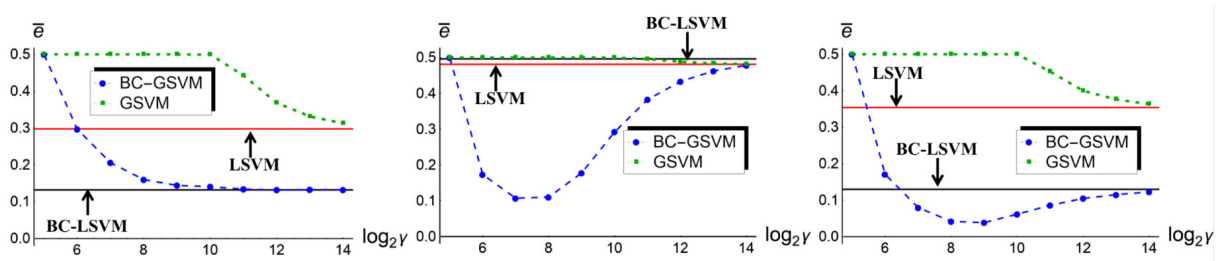
Figure 4.3: The error rates of the BC-LSVM, LSVM, BC-GSVM and GSVM for (a) to (c). The left panels display $\overline{e}(1)$, the middle panels display $\overline{e}(2)$ and the right panels display $\overline{e}$ for $d = 2^s$, $s = 5, \ldots, 12$. For the LSVM and GSVM, $\overline{e}(2)$ was too high to describe.

We observed that the BC-SVMs give good performances as d increases for (a) and (c). However, for (b), the error rate of the BC-LSVM is close to 0.5 because $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = 0$. On the other hand, the BC-GSVM gave good performances as d increases by drawing information about heteroscedasticity thorough the geometric representation as in Figures 4.1 and 4.2. For the LSVM and GSVM, $\overline{e}(1)$ and

$\overline{e}(2)$ became quite unbalanced as d increases. In particular, the strong inconsistency (4.16) occured for the GSVM. This is because of the bias in the GSVM. We give their theoretical backgrounds in Section 4.3.2.

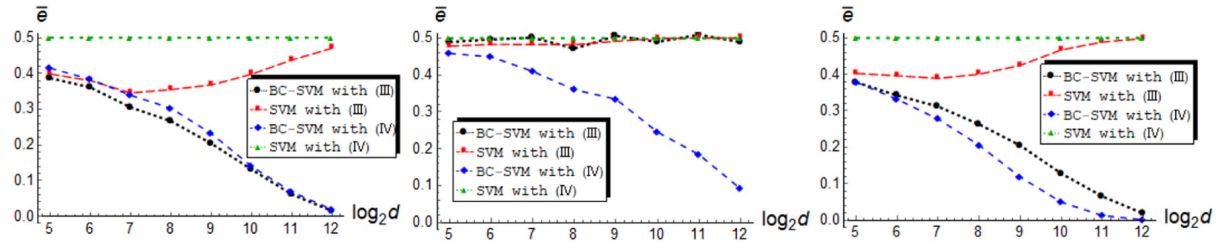
Next, we considered (a) to (c) for $(n_1, n_2) = (20, 10)$, $d = 1024$ $(= 2^{10})$ and $\gamma = 2^s,\ s = 5, \ldots, 14$ in (II). Similar to Figure 4.1, we calculated the average error rate $\overline{e}$ by 2000 replications and plotted the results in Figure 4.2. We observed that the BC-GSVM and GSVM are close to the BC-LSVM and LSVM, respectively, as γ increases for (a) and (c). We give their theoretical backgrounds in Section 4.3.3. For (b) and (c), the BC-GSVM gave better performances than the other SVMs for several settings of γ. We note that the performance of the BC-GSVM (or GSVM) heavily depends on γ. We discuss a choice of γ in Section 4.4.

Finally, we compared the performance of the BC-SVM with SVM for kernel functions (III) and (IV). We set $(\zeta, r) = (d, 2)$ in (III) and $\xi = d/4$ in (IV). We considered (a) to (c) for $(n_1, n_2) = (20, 10)$ and $d = 2^s,\ s = 5, \ldots, 12$. Similar to Figure 4.1, we calculated the average error rate $\overline{e}$ by 2000 replications and plotted the results in Figure 4.3. We observed that the BC-SVM with (III) or (IV) gives good performances compared to the SVMs for (a) and (c). On the other hand, for (b) the BC-SVM with (IV) gave good performances as d increases. This is probably because the kernel function (IV) can draw information about heteroscedasticity via the difference of $\mathbf{\Sigma}_i$s. We investigated their performances in other high-dimensional settings as well. In most cases, the BC-SVM with (III) or (IV) gave better performances than the SVMs. We investigate asymptotic properties of the BC-SVM with (III) in Section 4.3.4.



(a) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\mathbf{\Sigma}_1) = \mathrm{tr}(\mathbf{\Sigma}_2)$ (b) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\mathrm{tr}(\mathbf{\Sigma}_1) \neq \mathrm{tr}(\mathbf{\Sigma}_2)$ (c) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\mathbf{\Sigma}_1) \neq \mathrm{tr}(\mathbf{\Sigma}_2)$

Figure 4.4: The average error rate, $\overline{e}$, of the BC-GSVM and GSVM for (a) to (c) when $d = 1024$ and $\gamma = 2^s,\ s = 5, \ldots, 14$. The average error rates of the BC-LSVM and LSVM are described by the lines.



(a) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\mathbf{\Sigma}_1) = \mathrm{tr}(\mathbf{\Sigma}_2)$ (b) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\mathrm{tr}(\mathbf{\Sigma}_1) \neq \mathrm{tr}(\mathbf{\Sigma}_2)$ (c) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\mathbf{\Sigma}_1) \neq \mathrm{tr}(\mathbf{\Sigma}_2)$

Figure 4.5: The average error rates of the BC-SVM and SVM for (III) and (IV) in cases of (a) to (c), where $(\zeta, r) = (d, 2)$ in (III) and $\xi = d/4$ in (IV). The panels display the error rates for $d = 2^s,\ s = 5, \ldots, 12$.

4.3 Asymptotic properties by kernel functions

In this section, we investigate asymptotic properties of the nonlinear SVM brought by kernel functions. We assume that $\limsup_{d\to\infty} \|\boldsymbol{\mu}_i\|^2/d < \infty$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0,\infty)$ as $d \to \infty$ for $i = 1, 2$. Here, for a function, $f(\cdot)$, "$f(d) \in (0,\infty)$ as $d \to \infty$" implies $\liminf_{d\to\infty} f(d) > 0$ and $\limsup_{d\to\infty} f(d) < \infty$. Similar to Bai and Saranadasa (1996) and Aoshima and Yata (2014), we assume the following assumption for π_is as necessary:

(A-ii) Let $\boldsymbol{z}_{ij},\ j = 1, \ldots, n_i$, be i.i.d. random p_i-vectors having $E(\boldsymbol{z}_{ij}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{z}_{ij}) = \boldsymbol{I}_{p_i}$ for each i $(= 1, 2)$ and some p_i. Let $\boldsymbol{z}_{ij} = (z_{i1j}, \ldots, z_{ip_ij})^T$ whose components satisfy that $\limsup_{d\to\infty} E(z_{irj}^4) < \infty$ for all r and

$$E(z_{irj}^2 z_{isj}^2) = E(z_{irj}^2)E(z_{isj}^2) = 1 \quad \text{and} \quad E(z_{irj}z_{isj}z_{itj}z_{iuj}) = 0$$

for all $r \neq s, t, u$. Then, the observations, $\boldsymbol{x}_{ij}$s, from each π_i $(i = 1, 2)$ are given by $\boldsymbol{x}_{ij} = \boldsymbol{\Gamma}_i \boldsymbol{z}_{ij} + \boldsymbol{\mu}_i,\ j = 1, \ldots, n_i$, where $\boldsymbol{\Gamma}_i$ is a $d \times p_i$ matrix such that $\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^T = \boldsymbol{\Sigma}_i$.

Note that z_{irj}s are i.i.d. as the standard normal distribution when the π_is are Gaussian and $\boldsymbol{\Gamma}_i = \boldsymbol{\Sigma}_i^{1/2}$. Thus, (A-ii) naturally holds when the π_is are Gaussian. Another example satisfying (A-ii) is the case when the π_is have a skew normal distribution. See Remark S4.1 in Aoshima and Yata (2019b) for the details.

4.3.1 Linear kernel

We consider the LSVM, that is, the classifier (4.6) has the kernel function (I). We set $\kappa_1 = \|\boldsymbol{\mu}_1\|^2$, $\kappa_2 = \|\boldsymbol{\mu}_1\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1)$, $\kappa_3 = \|\boldsymbol{\mu}_2\|^2$, $\kappa_4 = \|\boldsymbol{\mu}_2\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_2)$ and $\kappa_5 = \boldsymbol{\mu}_1^T \boldsymbol{\mu}_2$, so that

$$\Delta_\kappa = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \ (= \Delta_{(I)},\ \text{say}) \ \text{ and } \ \eta_i = \mathrm{tr}(\boldsymbol{\Sigma}_i) \ (= \eta_{i(I)},\ \text{say}) \ \text{ for } i = 1, 2.$$

We note that the LSVM is invariant to linear transformations on the data set. Thus, in Section 4.3.1, we assume $\boldsymbol{\mu}_2 = \boldsymbol{0}$ without loss of generality, so that $\kappa_3 = \kappa_5 = 0$, $\kappa_4 = \eta_{2(I)}$ and $\Delta_{(I)} = \|\boldsymbol{\mu}_1\|^2$. In addition, we assume the following condition:

(C-ii) $\dfrac{n_i \mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{\Delta_{(I)}^2} = o(1)$ for $i = 1, 2$.

Note that $\Delta_{(I)}^2/\mathrm{tr}(\boldsymbol{\Sigma}_i^2) = O(d)$ from the facts that $\limsup_{d\to\infty} \Delta_{(I)}/d < \infty$, $\mathrm{tr}(\boldsymbol{\Sigma}_i^2) \geq \mathrm{tr}(\boldsymbol{\Sigma}_i)^2/d$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0,\infty)$ as $d \to \infty$ for $i = 1, 2$. Thus, $n_i = o(d)$ when (C-ii) is met. Under (4.1), (C-ii) holds when $\liminf_{d\to\infty} \Delta_{(I)}/d > 0$ and n_is are fixed. We have the following result.

Lemma 4.3. *Assume (A-ii) and (C-ii). Then, the assumptions (A-i) and (A-i') are met for the kernel function (I).*

By combining Lemma 4.3 with Theorem 4.1 and Corollary 4.1, we have the following results.

Corollary 4.2. *For the LSVM, one can claim that*

$$(4.2) \textit{ holds if } \limsup_{d\to\infty} \frac{|\delta_{(I)}|}{\Delta_{(I)}} < 1, \quad (4.15) \textit{ holds if } \liminf_{d\to\infty} \frac{\delta_{(I)}}{\Delta_{(I)}} > 1, \quad \textit{and}$$

$$(4.16) \textit{ holds if } \limsup_{d\to\infty} \frac{\delta_{(I)}}{\Delta_{(I)}} < -1$$

under (A-ii) and (C-ii), where $\delta_{(I)} = \eta_{1(I)}/n_1 - \eta_{2(I)}/n_2$.

In Chapter 2, we gave the results of Corollary 4.2 under slightly different conditions. They provided the following bias correction of the LSVM: Let $\Delta_{*(I)} = \Delta_{(I)} + \eta_{1(I)}/n_1 + \eta_{2(I)}/n_2$. Estimate $\Delta_{*(I)}$ and $\delta_{(I)}$ by

$$\widehat{\Delta}_{*(I)} = \|\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}\|^2 \text{ and } \widehat{\delta}_{(I)} = \mathrm{tr}(\boldsymbol{S}_{1n_1})/n_1 - \mathrm{tr}(\boldsymbol{S}_{2n_2})/n_2,$$

where $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})^T/(n_i - 1)$. Note that $E(\widehat{\Delta}_{*(I)}) = \Delta_{*(I)}$ and $E(\hat{\delta}_{(I)}) = \delta_{(I)}$. Let $\hat{y}_{(I)}(\boldsymbol{x}_0)$ denote $\hat{y}(\boldsymbol{x}_0)$ given by using the kernel function (I). Then, from Chapter 2, we gave the BC-LSVM as

$$\hat{y}_{BC(I)}(\boldsymbol{x}_0) = \hat{y}_{(I)}(\boldsymbol{x}_0) - \hat{\delta}_{(I)}/\widehat{\Delta}_{*(I)}. \tag{4.18}$$

One classifies $\boldsymbol{x}_0$ into π_1 if $\hat{y}_{BC(I)}(\boldsymbol{x}_0) < 0$ and into π_2 otherwise.

We note that $\widehat{\Delta}_{*(I)}$ and $\hat{\delta}_{(I)}$ are equivalent to $\widehat{\Delta}_{\kappa*}$ and $\hat{\delta}$ when $k(\cdot,\cdot)$ is the linear kernel. From Lemma 4.3 and Theorem 4.2, we have the following result.

Corollary 4.3. *Under (A-ii) and (C-ii), the BC-LSVM holds the consistency (4.2).*

The BC-LSVM has the consistency property without (C-i). Chan and Hall (2009) considered a different bias correction for the LSVM. In Chapter 2, we compared the BC-LSVM with the LSVM both in numerical simulations and real data analyses. They concluded that the BC-LSVM gives adequate performances for HDLSS data even when n_is are quite unbalanced (i.e., extremely imbalanced data).

4.3.2 Gaussian kernel

We consider the Gaussian kernel SVM (GSVM), that is, classifier (4.6) has the kernel function (II). We set $\kappa_1 = \exp\{-2\mathrm{tr}(\boldsymbol{\Sigma}_1)/\gamma\}$ $(= \kappa_{1(II)}, \text{ say})$, $\kappa_3 = \exp\{-2\mathrm{tr}(\boldsymbol{\Sigma}_2)/\gamma\}$ $(= \kappa_{3(II)}, \text{ say})$, $\kappa_2 = \kappa_4 = 1$, and $\kappa_5 = \exp[-\{\mathrm{tr}(\boldsymbol{\Sigma}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2) + \Delta_{(I)}\}/\gamma]$ $(= \kappa_{5(II)}, \text{ say})$, so that

$$\begin{aligned}\Delta_\kappa =& \kappa_{1(II)} + \kappa_{3(II)} - 2\kappa_{5(II)} \ (= \Delta_{(II)}, \text{ say}) \text{ and}\\ \eta_i =& 1 - \exp\left(-2\mathrm{tr}(\boldsymbol{\Sigma}_i)/\gamma\right) \ (= \eta_{i(II)}, \text{ say}) \text{ for } i = 1, 2.\end{aligned}$$

We note that $\Delta_{(II)} > 0$ when $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ or $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$. Let $\mathrm{tr}(\boldsymbol{\Sigma}_{\min}) = \min_{i=1,2}\mathrm{tr}(\boldsymbol{\Sigma}_i)$ and $\psi = \exp\{-2\mathrm{tr}(\boldsymbol{\Sigma}_{\min})/\gamma\}$. We assume the following condition:

(C-iii) $\dfrac{n_i\mathrm{tr}(\boldsymbol{\Sigma}_i^2) + \Delta_{(I)}\left\{n_i\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\right\}^{1/2}}{\min\{\gamma^2\Delta_{(II)}^2/\psi^2,\ \gamma^2\}} = o(1)$ for $i = 1, 2$.

We note that (C-iii) is a convergence condition of the GSVM. Under (4.1), (C-iii) holds when $\liminf_{d\to\infty} \Delta_{(II)} > 0$, $\liminf_{d\to\infty} \gamma/d > 0$ and n_is are fixed. Note that $\psi \to 1$ and $\gamma\Delta_{(II)} = 2\Delta_{(I)}\{1 + o(1)\}$ as $d \to \infty$ under $d^2/(\gamma\Delta_{(I)}) = o(1)$ as $d \to \infty$ from the fact that "$d^2/(\gamma\Delta_{(I)}) = o(1)$" implies "$d/\gamma = o(1)$". Thus, (C-iii) holds under (C-ii) and $d^2/(\gamma\Delta_{(I)}) = o(1)$. See Section 4.3.3 for the relation between the kernels (I) and (II).

We have the following result.

Lemma 4.4. *Assume (A-ii) and (C-iii). Then, the assumptions (A-i) and (A-i') are met for the kernel function (II).*

By combining Lemma 4.4 with Theorem 4.1 and Corollary 4.1, we have the following results.

Corollary 4.4. *For the GSVM, one can claim that*

$$(4.2) \textit{ holds if } \limsup_{d\to\infty} \frac{|\delta_{(II)}|}{\Delta_{(II)}} < 1, \quad (4.15) \textit{ holds if } \liminf_{d\to\infty} \frac{\delta_{(II)}}{\Delta_{(II)}} > 1, \quad \textit{and}$$

$$(4.16) \textit{ holds if } \limsup_{d\to\infty} \frac{\delta_{(II)}}{\Delta_{(II)}} < -1$$

under (A-ii) and (C-iii), where $\delta_{(II)} = \eta_{1(II)}/n_1 - \eta_{2(II)}/n_2$.

We denote $\hat{\eta}_i$ $(i = 1, 2)$ and $\widehat{\Delta}_{\kappa*}$ for the kernel function (II) by $\hat{\eta}_{i(II)}$ and $\widehat{\Delta}_{*(II)}$. Here, $\hat{\eta}_i$ and $\widehat{\Delta}_{\kappa*}$ are defined in Section 4.2.3. Let $\Delta_{*(II)} = \Delta_{(II)} + \eta_{1(II)}/n_1 + \eta_{2(II)}/n_2$ and $\hat{\delta}_{(II)} = \hat{\eta}_{1(II)}/n_1 - \hat{\eta}_{2(II)}/n_2$. Let $\hat{y}_{(II)}(\boldsymbol{x}_0)$ denote $\hat{y}(\boldsymbol{x}_0)$ given by using the kernel function (II). Then, we give the bias-corrected GSVM (BC-GSVM) as

$$\hat{y}_{BC(II)}(\boldsymbol{x}_0) = \hat{y}_{(II)}(\boldsymbol{x}_0) - \hat{\delta}_{(II)}/\widehat{\Delta}_{*(II)}. \tag{4.19}$$

One classifies $\boldsymbol{x}_0$ into π_1 if $\hat{y}_{BC(II)}(\boldsymbol{x}_0) < 0$ and into π_2 otherwise. From Theorem 4.2 and Lemma 4.4, we have the following result.

Corollary 4.5. *Under (A-ii) and (C-iii), the BC-GSVM holds the consistency (4.2).*

The BC-GSVM has the consistency property without (C-i).

Now, we consider the following condition:

$$\gamma/d \in (0, \infty) \quad \text{as } d \to \infty. \tag{4.20}$$

Let

$$\Delta_\Sigma = |\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)|, \quad \theta_1 = \exp(-\Delta_{(I)}/\gamma) \quad \text{and} \quad \theta_2 = \exp(-\Delta_\Sigma/\gamma).$$

Note that $\Delta_{(I)} = O(d)$ and

$$\Delta_{(II)}/\psi = (1 - \theta_2)^2 + 2\theta_2(1 - \theta_1). \tag{4.21}$$

If one assumes that

$$\liminf_{d\to\infty} \Delta_\Sigma/d > 0,$$

it follows that $\liminf_{d\to\infty} \Delta_{(II)} > 0$ under (4.20), so that (C-iii) holds as $d \to \infty$ while n is fixed under (4.1) and (4.20). Thus, BC-GSVM has the consistency (4.2) even when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. On the other hand, the BC-LSVM (or the LSVM) does not hold the consistency property when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. We emphasize that the BC-GSVM (or the GSVM) draws information about heteroscedasticity via the difference of $\mathrm{tr}(\boldsymbol{\Sigma}_i)$s. The accuracy becomes higher as the difference grows. See Figure 4.1.

4.3.3 Relation between the linear kernel and Gaussian kernel

We consider the following conditions for $\gamma > 0$:

(C-iv) $\dfrac{d^2}{\gamma\Delta_{(I)}} \to 0$ as $d \to \infty$, and (C-v)

(C-v) $\dfrac{\Delta_{(I)} + \Delta_\Sigma^2/\Delta_{(I)}}{\gamma} \to 0$ as $d \to \infty$.

Note that (C-iv) implies (C-v). By noting that $\psi \to 1$ as $d \to \infty$ under (C-iv), it holds from (4.21) that under (C-iv)

$$\gamma\Delta_{(II)} = 2\Delta_{(I)}\{1 + o(1)\}. \tag{4.22}$$

Thus, the GSVM becomes close to the LSVM under (C-iv). In fact, we have the following result.

Proposition 4.3. *Under (A-ii), (C-ii) and (C-iv), it holds that*

$$\hat{y}_{(II)}(\boldsymbol{x}_0) = \hat{y}_{(I)}(\boldsymbol{x}_0)\{1 + o_P(1)\} \quad \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2.$$

Hence, the GSVM is asymptotically equivalent to the LSVM when γ satisfies (C-iv). On the other hand, it holds from (4.21) that under (C-v)

$$\gamma\Delta_{(II)} = 2\psi\Delta_{(I)}\{1 + o(1)\}. \tag{4.23}$$

Proposition 4.4. *Under (A-ii), (C-ii) and (C-v), it holds that*

$$\Big(\frac{\Delta_{(I)}}{\Delta_{*(I)}}\frac{\Delta_{*(II)}}{\Delta_{(II)}}\Big)\hat{y}_{BC(II)}(\boldsymbol{x}_0) = \hat{y}_{BC(I)}(\boldsymbol{x}_0)\{1+o_P(1)\}$$

when $\boldsymbol{x}_0 \in \pi_i$ *for* $i = 1, 2$.

Hence, the BC-GSVM is asymptotically equivalent to the BC-LSVM when γ satisfies (C-v).

4.3.4 Polynomial kernel

In this section, we consider the polynomial kernel SVM, that is, classifier (4.6) has the kernel function (III). We give some asymptotic properties of the polynomial kernel SVM. We consider the following conditions for ζ and r:

$$\zeta/d \in (0,\infty) \ \text{ and } \ r \in (0,\infty) \ \text{ as } d \to \infty. \tag{4.24}$$

We set $\kappa_1 = (\zeta + \|\boldsymbol{\mu}_1\|^2)^r$, $\kappa_2 = (\zeta + \mathrm{tr}(\boldsymbol{\Sigma}_1) + \|\boldsymbol{\mu}_1\|^2)^r$, $\kappa_3 = (\zeta + \|\boldsymbol{\mu}_2\|^2)^r$, $\kappa_4 = (\zeta + \mathrm{tr}(\boldsymbol{\Sigma}_2) + \|\boldsymbol{\mu}_2\|^2)^r$ and $\kappa_5 = (\zeta + \boldsymbol{\mu}_1^T\boldsymbol{\mu}_2)^r$. Then, we have the following result

Proposition 4.5. *Assume (4.1), (4.24) and (A-ii). Assume that n is fixed and*

$$\liminf_{d\to\infty}\Big|\frac{\|\boldsymbol{\mu}_1\|^2 - \|\boldsymbol{\mu}_2\|^2}{d}\Big| > 0. \tag{4.25}$$

Then, the assumptions (A-i) and (A-i') are met for the polynomial kernel (III). Furthermore, the BC-SVM with the polynomial kernel (III) holds the consistency (4.2).

See Figure 4.5 for the performance of the BC-SVM with the polynomial kernel (III).

Remark 14. *For the Laplace kernel (IV), it is difficult to provide asymptotic properties of the kernel SVM unless π_is are Gaussian. Detailed study of the BC-SVM with the Laplace kernel is left to a future work.*

4.4 How to choose γ in the Gaussian kernel

In this section, we discuss a choice of γ in the Gaussian kernel function (II).

4.4.1 Behaviors of $\Delta_{(II)}$ for several settings of γ

We consider the following two conditions for $\Delta_{(I)}$ and Δ_Σ:

$$\Delta_\Sigma/\Delta_{(I)} \to 0 \ \text{ as } d \to \infty, \ \text{ and} \tag{4.26}$$

$$\liminf_{d\to\infty} \Delta_\Sigma/\Delta_{(I)} > 0. \tag{4.27}$$

We first consider $\Delta_{(II)}$ under (4.26). From (4.21) it holds that $\Delta_{(II)}/\psi = 1 + \exp(-2\Delta_\Sigma/\gamma) + o(1)$ under $\liminf_{d\to\infty} \Delta_\Sigma\ /\gamma > 0$ and (4.26), so that the BC-GSVM (or GSVM) loses information about $\Delta_{(I)}$. Thus, we do not consider the case when $\liminf_{d\to\infty} \Delta_\Sigma/\gamma > 0$ under (4.26). Under (4.26) we consider the following conditions for γ, $\Delta_{(I)}$ and Δ_Σ:

$$\Delta_\Sigma/\gamma \to 0 \ \text{ as } d \to \infty, \ \text{ and} \tag{4.28}$$

$$\Delta_{(I)}/\gamma \to 0 \ \text{ as } d \to \infty. \tag{4.29}$$

From (4.21) it holds that under (4.26) and (4.28)

$$\gamma\Delta_{(II)}/\psi = 2\gamma\{1-\exp(-\Delta_{(I)}/\gamma)\}\{1+o(1)\}.$$

On the other hand, it holds from (4.23) that under (4.28) and (4.29)

$$\gamma\Delta_{(II)}/\psi = 2\Delta_{(I)}\{1+o(1)\}$$

because (C-v) holds under (4.26) and (4.29). From Proposition 4.4 we note that the BC-LSVM is asymptotically equivalent to the BC-GSVM under (4.26) and (4.29). Also, note that $\gamma\{1-\exp(-\Delta_{(I)}/\gamma)\} \leq \Delta_{(I)}$ for any $\gamma > 0$. Then, from the convergence condition (C-iii), when (4.26) is met, we recommend to use the BC-LSVM or the BC-GSVM with γ satisfying (4.29).

Next, we consider $\Delta_{(II)}$ under (4.27). From (4.21) it holds that under (4.27) and (4.28)

$$\gamma\Delta_{(II)}/\psi = 2\Delta_{(I)} + o(\Delta_\Sigma).$$

When (4.27) is met, the BC-GSVM (or GSVM) with γ satisfying (4.28) loses information about heteroscedasticity via the difference of $\mathrm{tr}(\boldsymbol{\Sigma}_i)$s. Thus, we do not consider the case when $\Delta_\Sigma/\gamma = o(1)$ as $d \to \infty$ under (4.27). Under (4.27), we consider the following conditions for γ and Δ_Σ:

$$\Delta_\Sigma/\gamma \to \infty \quad \text{as } d \to \infty, \quad \text{or} \tag{4.30}$$

$$\Delta_\Sigma/\gamma \in (0,\infty) \quad \text{as } d \to \infty. \tag{4.31}$$

It holds that under (4.27) and (4.30)

$$\gamma\Delta_{(II)}/(\psi\Delta_\Sigma) = (\gamma/\Delta_\Sigma)\{1+o(1)\} = o(1).$$

Also, it holds that under (4.27) and (4.31)

$$\liminf_{d\to\infty} \gamma\Delta_{(II)}/(\psi\Delta_\Sigma) > 0.$$

Hence, from the convergence condition (C-iii), when (4.27) is met, we recommend to use the BC-GSVM with γ satisfying (4.31).

4.4.2 Choice of γ

In this section, we give a choice of γ in the GSVM. From Section 4.4.1, we recommend to use the BC-GSVM with γ satisfying

(i) the condition (4.29) when (4.26) is met, and

(ii) the condition (4.31) when (4.27) is met.

For the dual form (4.4), from Lemma 4.1, under (4.5) and several conditions, it holds that $\acute{\boldsymbol{\alpha}}^T\boldsymbol{K}\acute{\boldsymbol{\alpha}} = \Delta\alpha_\star^2\{1+o_P(1)\} + \eta_1\sum_{j=1}^{n_1}\alpha_j^2 + \eta_2\sum_{j=n_1+1}^{n}\alpha_j^2$, so that

$$\frac{\acute{\boldsymbol{\alpha}}^T\boldsymbol{K}\acute{\boldsymbol{\alpha}}}{\alpha_\star^2\Delta} - 1 - \frac{\eta_1\sum_{j=1}^{n_1}\alpha_j^2 + \eta_2\sum_{j=n_1+1}^{n}\alpha_j^2}{\alpha_\star^2\Delta} \quad (= \mathrm{Loss}(\gamma), \text{ say}). \tag{4.32}$$

We emphasize that the accuracy of the BC-SVM (or SVM) heavily depends on the convergence rate of $\mathrm{Loss}(\gamma)$ because the bias in $\hat{y}(\boldsymbol{x}_0)$ converges to δ in Proposition 4.1. See Lemma 4.1 in Section 4.2.2.

Thus, for the Gaussian kernel (II), we consider such γ as to have a higher convergence rate of Loss(γ). From Proposition 4.1 and (4.43) to (4.48) in Appendix, we can evaluate that under several conditions

$$\begin{aligned}\mathrm{Loss}(\gamma) &= \frac{1}{\gamma\Delta_{(II)}}\Big(\frac{n_1(n_1-1)\kappa_{1(II)}}{n_1^2}+\frac{n_2(n_2-1)\kappa_{3(II)}}{n_2^2}+2\kappa_{5(II)}\Big)\times O_P(\varepsilon)\\ &= \frac{\kappa_{1(II)}+\kappa_{3(II)}+2\kappa_{5(II)}}{\gamma\Delta_{(II)}}\times O_P(\varepsilon),\end{aligned}$$

where $\varepsilon = \max_{i=1,2}[\mathrm{tr}(\boldsymbol{\Sigma}_i^2)+\Delta_{(I)}\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}^{1/2}]^{1/2}$. Thus from (4.32), one may consider γ as

$$\gamma_0 = \operatorname*{argmin}_{\gamma>0}\frac{\kappa_{1(II)}+\kappa_{3(II)}+2\kappa_{5(II)}}{\gamma\Delta_{(II)}}. \tag{4.33}$$

When (4.26) is met, we have the following result.

Proposition 4.6. *Under (4.26) it holds that $\Delta_{(I)}/\gamma_0 \to 0$ as $d\to\infty$.*

Hence, when (4.26) is met, the BC-GSVM with γ_0 is asymptotically equivalent to the BC-LSVM because (C-v) is met under (4.26) and (4.29). See Proposition 4.6.

Next, we consider the case when

$$\limsup_{d\to\infty}\Delta_{(I)}/\Delta_\Sigma \le 1. \tag{4.34}$$

Proposition 4.7. *Under (4.34) it holds that $\Delta_\Sigma/\gamma_0 \in (0,\infty)$ as $d\to\infty$.*

Finally, we consider the case when

$$\liminf_{d\to\infty}\Delta_{(I)}/\Delta_\Sigma \ge 1 \quad\text{and}\quad \limsup_{d\to\infty}\Delta_{(I)}/\Delta_\Sigma < \infty. \tag{4.35}$$

Since it is very difficult to evaluate γ_0 under (4.35), we investigate the behavior of γ_0 numerically. Let $\gamma_\star = \gamma/\Delta_\Sigma$ and $\omega = \Delta_{(I)}/\Delta_\Sigma$. By noting that $\Delta_{(II)}/\psi = 1+\theta_2^2-2\theta_1\theta_2$ and $(\kappa_{1(II)}+\kappa_{3(II)}+2\kappa_{5(II)})/\psi = 1+\theta_2^2+2\theta_1\theta_2$, it holds that

$$\begin{aligned}\Delta_\Sigma\frac{\kappa_{1(II)}+\kappa_{3(II)}+2\kappa_{5(II)}}{\gamma\Delta_{(II)}} &= \frac{\Delta_\Sigma}{\gamma}\Big(1+\frac{4\theta_1\theta_2}{1+\theta_2^2-2\theta_1\theta_2}\Big)\\ &= \frac{1}{\gamma_\star}\Big(1+\frac{4\exp\big\{-(\omega+1)/\gamma_\star\big\}}{1+\exp(-2/\gamma_\star)-2\exp\big\{-(\omega+1)/\gamma_\star\big\}}\Big)\quad\big(=F(\gamma_\star),\ \text{say}\big).\end{aligned} \tag{4.36}$$

Thus, we consider the following minimization:

$$\gamma_{0\star} = \operatorname*{argmin}_{\gamma_\star>0} F(\gamma_\star).$$

Note that $\gamma_0 = \Delta_\Sigma\gamma_{0\star}$. Hence, (4.33) depends only on ω. We plotted $\gamma_{0\star}$ and $\gamma_{0\star}/(\omega^3/3)$ for $\omega = 1,\ldots,100$ in Figure 4.6.

We observed that $\gamma_{0\star}$ behaves around $\omega^3/3$. One may conclude that $\gamma_{0\star} = O(\omega^3)$, so that from Proposition 4.7 it holds that $\Delta_\Sigma/\gamma_0 = 1/\gamma_{0\star} \in (0,\infty)$ as $d\to\infty$ when (4.27) is met.

In conclusion, we recommend to use the BC-GSVM with γ_0. From (4.36) we estimate γ_0 as

$$\hat\gamma_0 = \operatorname*{argmin}_{\gamma>0}\gamma^{-1}\{1+4\hat\theta_1\hat\theta_2/(1+\hat\theta_2^2-2\hat\theta_1\hat\theta_2)\}, \tag{4.37}$$

where $\hat\theta_1 = \exp(-\widehat\Delta_{*(I)}/\gamma)$ and $\hat\theta_2 = \exp(-\widehat\Delta_\Sigma/\gamma)$ with $\widehat\Delta_\Sigma = |\mathrm{tr}(\boldsymbol{S}_{1n_1})-\mathrm{tr}(\boldsymbol{S}_{2n_2})|$. See Section 4.5.1 for the performance of the BC-SVM with $\hat\gamma_0$.

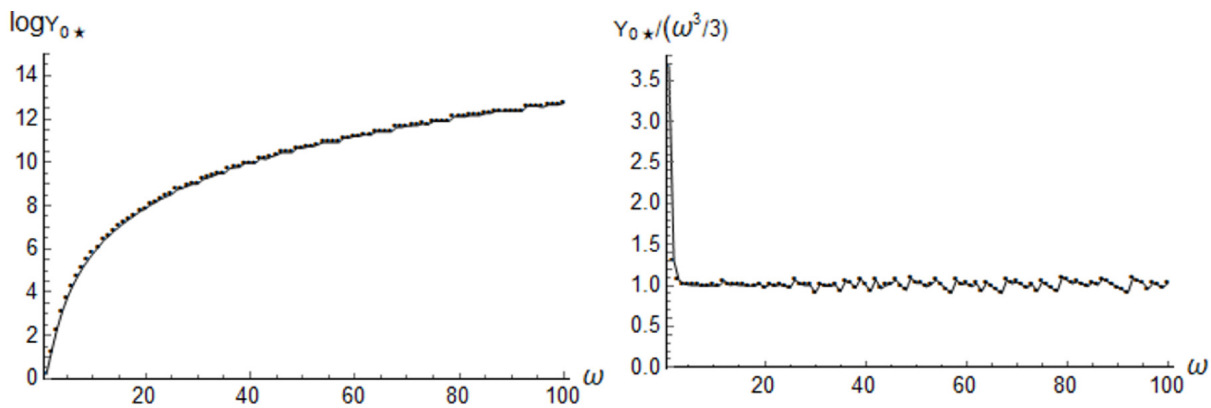


Figure 4.6: The left panel displays $\log \gamma_{0\star}$ and the right panel displays $\gamma_{0\star}/(\omega^3/3)$ for $\omega = 1, \ldots, 100$.

Remark 15. *We note that* $E(\widehat{\Delta}_{(I)}) = \Delta_{(I)}$, *where* $\widehat{\Delta}_{(I)} = \widehat{\Delta}_{*(I)} - \mathrm{tr}(\boldsymbol{S}_{1n_1})/n_1 - \mathrm{tr}(\boldsymbol{S}_{2n_2})/n_2$. *However, it does not hold* $P(\widehat{\Delta}_{(I)} \geq 0) = 1$. *Thus we use* $\widehat{\Delta}_{*(I)}$ *in (4.37) since* $P(\widehat{\Delta}_{*(I)} \geq 0) = 1$.

Remark 16. *Note that* $E(\widehat{\Delta}_{*(I)}) = \Delta_{*(I)}$, *and* $\mathrm{Var}(\widehat{\Delta}_{(I)}) = O[\sum_{i=1}^{2}\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i^2 + \Delta_{(I)}\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}/n_i\}]$ *and* $\mathrm{Var}\{\mathrm{tr}(\boldsymbol{S}_{in_i})\} = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i\}$ *under (A-ii). Thus, if* $\mathrm{tr}(\boldsymbol{\Sigma}_i)/(n_i\Delta_{(I)}) = o(1)$ *and* $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/(n_i\Delta_{\Sigma}^2) = o(1)$ *as* $d, n \to \infty$ *for* $i = 1, 2$, *it holds that* $\widehat{\Delta}_{*(I)} = \Delta_{(I)}\{1 + o_P(1)\}$ *and* $\widehat{\Delta}_{\Sigma} = \Delta_{\Sigma}\{1 + o_P(1)\}$ *as* $d, n \to \infty$ *since* $\mathrm{tr}(\boldsymbol{\Sigma}_i^2) \leq \mathrm{tr}(\boldsymbol{\Sigma}_i)^2$, *so that* $\hat{\gamma}_0$ *becomes close to* γ_0 *in (4.33).*

4.5 Performances

In this section, we check the performance of the BC-SVM both in numerical simulations and real data analyses.

4.5.1 Simulations

For the settings (a) to (c) in Section 4.2.4, we first checked the performance of the BC-GSVM with $\hat{\gamma}_0$. Similar to Section 4.2.4, we calculated the error rates, $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$, of the BC-GSVM and the GSVM with $\gamma = \hat{\gamma}_0$ by 2000 replications and plotted the results in Figure 4.7. We laid $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ for the BC-LSVM and the LSVM by borrowing them from Figure 4.3. In the rth replication, we evaluated $\hat{\gamma}_{0r}$ by (4.37) and calculated $\overline{\gamma}_0 = \sum_{r=1}^{2000}\hat{\gamma}_{0r}/2000$. In Figure 4.8, we plotted $\Delta_{(I)}/\overline{\gamma}_0$, $\Delta_{(I)}/\gamma_0$, $\Delta_{\Sigma}/\overline{\gamma}_0$ and Δ_{Σ}/γ_0 for (a) to (c). As expected theoretically, we observed that the BC-GSVM with $\hat{\gamma}_0$ is asymptotically equivalent to the BC-LSVM for (a). See Section 4.4.2. On the other hand, $\hat{\gamma}_0$ did not become close to γ_0 for (b) and (c). However, one may conclude that $\Delta_{\Sigma}/\overline{\gamma}_0 < \infty$ as $d \to \infty$. The BC-GSVM draws information about heteroscedasticity via the difference of $\mathrm{tr}(\boldsymbol{\Sigma}_i)$s. See Section 4.4.1. This is the reason why the BC-GSVM with $\hat{\gamma}_0$ gave adequate performances for (b) and (c).

Next, we compared the performance of the BC-SVMs with the SVMs in non-Gaussian and imbalanced settings. We set $\boldsymbol{\mu}_2 = \boldsymbol{0}$, $\boldsymbol{\Sigma}_1 = 1.3\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_2 = 0.7\boldsymbol{B}(0.4^{|i-j|^{1/3}})\boldsymbol{B}$. Let $d_* = 2\lceil d^{1/2}/2\rceil$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. We set $\boldsymbol{\mu}_2 = (1, \ldots, 1, 0, \ldots, 0, -1, \ldots, -1)^T$ whose first $d_*/2$ elements are 1 and last $d_*/2$ elements are -1. Note that $\Delta_{(I)} = d_* \approx d^{1/2}$, so that (C-ii) does not hold. We generated $\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i$ $(= \boldsymbol{\Sigma}_i^{1/2}(z_{i1j}, \ldots, z_{idj})^T)$, $j = 1, 2, \ldots (i = 1, 2)$ independently from $z_{irj} = (y_{irj} - 1)/2^{1/2}$ $(r = 1, \ldots, d)$ in which y_{irj}s are i.i.d. as the chi-squared distribution with 1 degree of freedom. Note that (A-ii) holds. We considered two cases for $d = 2^s$, $s = 5, \ldots, 12$:

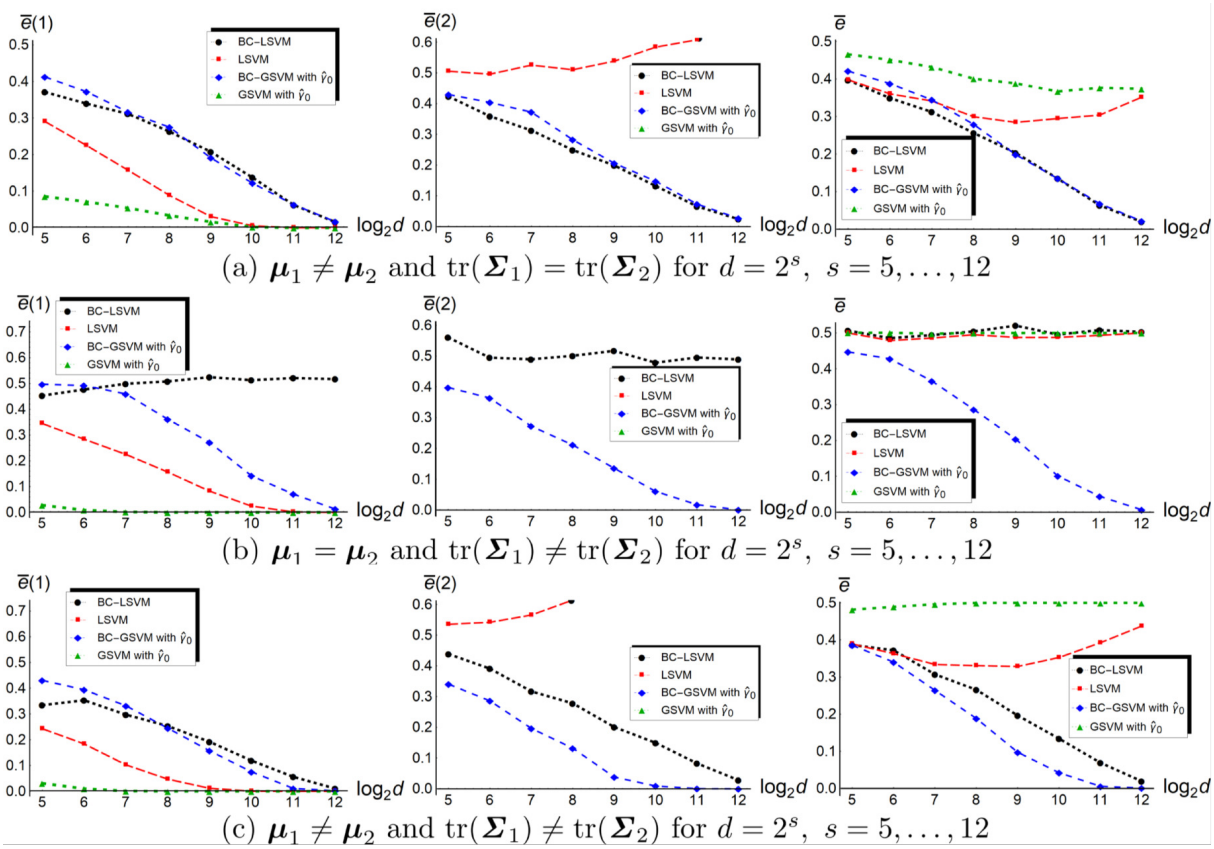
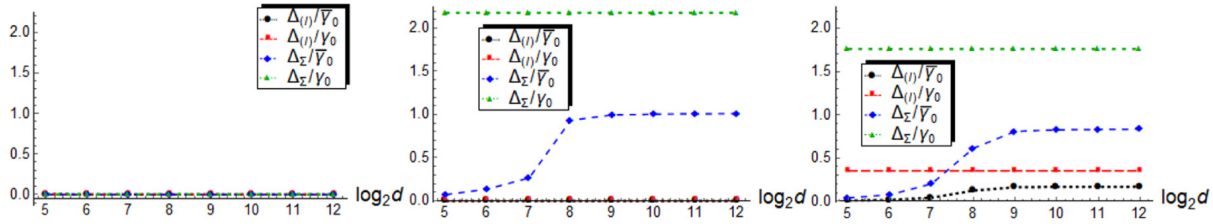


Figure 4.7: The error rates of the BC-LSVM, LSVM, BC-GSVM with $\gamma = \hat{\gamma}_0$ and GSVM with $\gamma = \hat{\gamma}_0$ for (d) and (e). The left panels display $\overline{e}(1)$, the middle panels display $\overline{e}(2)$ and the right panels display $\overline{e}$ for $d = 2^s,\ s = 5, \ldots, 12$. Their standard deviations are less than 0.0112.



(a) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\boldsymbol{\Sigma}_1) = \mathrm{tr}(\boldsymbol{\Sigma}_2)$ (b) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ (c) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$

Figure 4.8: Behaviors of $\Delta_{(I)}/\overline{\gamma}_0$, $\Delta_{(I)}/\gamma_0$, $\Delta_{\Sigma}/\overline{\gamma}_0$ and Δ_{Σ}/γ_0 for (a) to (c).

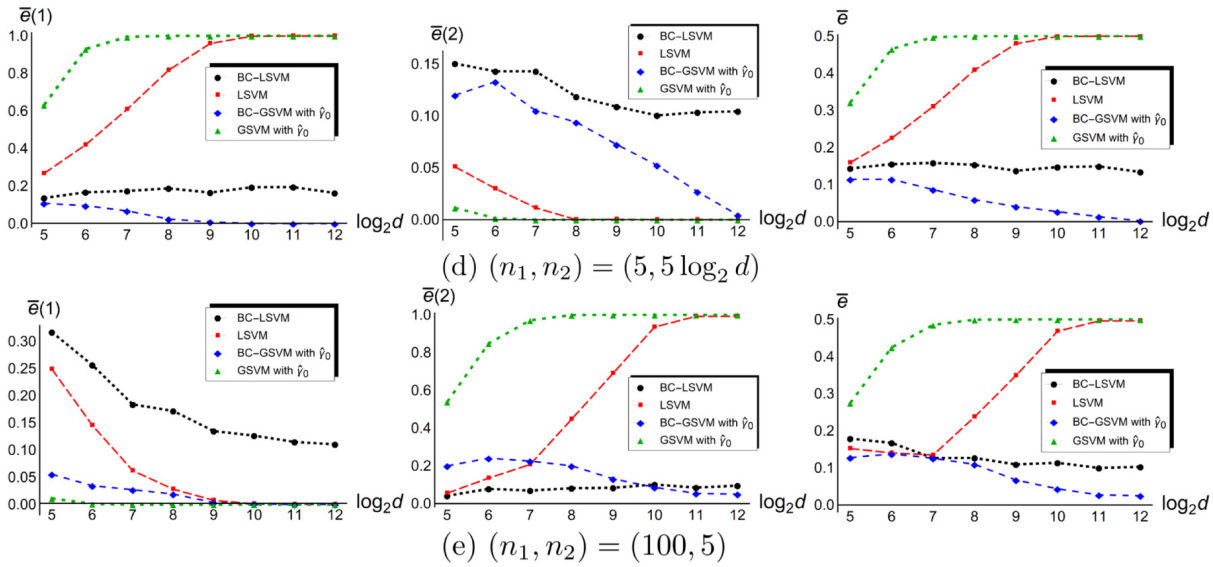


Figure 4.9: The error rates of the BC-LSVM, LSVM, BC-GSVM with $\gamma = \hat{\gamma}_0$ and GSVM with $\gamma = \hat{\gamma}_0$ for (d) and (e). The left panels display $\overline{e}(1)$, the middle panels display $\overline{e}(2)$ and the right panels display $\overline{e}$ for $d = 2^s,\ s = 5, \ldots, 12$. Their standard deviations are less than 0.0112.

(d) $(n_1, n_2) = (5, 5\log_2 d)$ and (e) $(n_1, n_2) = (100, 5)$.

For the BC-LSVM, LSVM, BC-GSVM with $\gamma = \hat{\gamma}_0$ and GSVM with $\gamma = \hat{\gamma}_0$, similar to Section 4.2.4, we calculated the error rates by 2000 replications and plotted the results in Figure 4.9. We observed that the BC-SVMs give adequate performances even when $n_i/n_{i'} \to 0$ for some $i\ (\neq i')$.

Throughout the simulations, $\hat{\gamma}_0$ by (4.37) was a preferable choice. We recommend to use a cross-validation procedure for γ around $\hat{\gamma}_0$. See Section 4.5.2.

4.5.2 Real data analyses

In this section, we analyze gene expression data sets by using the BC-SVMs and SVMs. We summarized the information on the data sets together with $\widehat{\Delta}_\Sigma/\widehat{\Delta}_{(I)}$ in Table 4.1, where $\widehat{\Delta}_{(I)}$ and $\widehat{\Delta}_\Sigma$ are given in Section 4.4.2.

We randomly split the data sets from (π_1, π_2) into training data sets of sizes (n_1, n_2) and test data sets of sizes $(m_1 - n_1, m_2 - n_2)$. We constructed the BC-SVM and SVM by using the training data sets. We checked accuracy by using the test data set for each π_i and denoted the misclassification rates by $\widehat{e}(1)_r$ and $\widehat{e}(2)_r$. We repeated this procedure 100 times and obtained $\widehat{e}(1)_r$ and $\widehat{e}(2)_r$, $r = 1, \ldots, 100$, for the BC-LSVM, LSVM, BC-GSVM and GSVM. For the BC-GSVM and GSVM, we used the average of the parameters selected by 5-fold cross-validation among $\gamma = (2s-1)\hat{\gamma}_0$ $(s = 1, \ldots, 5)$ with $\hat{\gamma}_0$ given by (4.37). We used the BC-GSVM and GSVM with $\hat{\gamma}_0$ (without applying the cross-validation) for Breast cancer because m_is are quite small for the dataset. We calculated the average misclassification rates, $\overline{e}(1)$ $(= \sum_{r=1}^{100} \widehat{e}(1)_r/100)$, $\overline{e}(2)$ $(= \sum_{r=1}^{100} \widehat{e}(2)_r/100)$ and $\overline{e}$ $(= \{\overline{e}(1)+\overline{e}(2)\}/2)$ for the SVMs and BC-SVMs in various combinations of (n_1, n_2) in Table 4.2.

We observed that the BC-SVMs give adequate performances compared to the SVMs especially when n_1 and n_2 are unbalanced. See Sections 4.3.1 and 4.3.2 for theoretical reasons. On the other hand, the BC-GSVM gave adequate performances compared to the BC-SVM for HGG and Breast cancer data sets. This is because $\widehat{\Delta}_\Sigma/\widehat{\Delta}_{(I)}$ is large for those data sets, so that the BC-GSVM can draw information

about heteroscedasticity via the difference of $\mathrm{tr}(\boldsymbol{\Sigma}_i)$s.

Table 4.1: Microarray data sets and $\widehat{\Delta}_\Sigma/\widehat{\Delta}_{(I)}$.

Data set	Number of genes	Sample size		$\frac{\widehat{\Delta}_\Sigma}{\widehat{\Delta}_{(I)}}$
	d	m_1	m_2	
Colon cancer by Alon et al. (1999)	2000	40	22	0.03
Leukemia by Golub et al. (1999)	7129	25	47	0.093
DLBCL by Shipp et al. (2002)	7129	58	19	0.668
HGG by Nutt et al. (2003)	12625	28	22	2.66
Breast cancer by Chang et al. (2003)	12625	14	10	0.78

4.6 Soft-margin nonlinear SVM

In Sections 4.2 to 4.5, we discussed asymptotic properties and the performance of the hmSVMs. In this section, we consider soft-margin SVMs (smSVM). The smSVM is given by $\hat{y}(\boldsymbol{x})$ after replacing (4.5) with

$$0 \le \alpha_j \le C,\ j = 1, \ldots, n, \quad \text{and} \quad \sum_{j=1}^{n} \alpha_j t_j = 0, \tag{4.38}$$

where $C(> 0)$ is a regularization parameter. Let $n_{\min} = \min\{n_1, n_2\}$. From (4.11) in Section 4.2, we can asymptotically claim that $\hat{\alpha}_j \le 2/(\Delta_{\kappa*} n_{\min})$ for all j. Thus we consider the following condition for C:

$$\liminf_{d\to\infty} \frac{C\Delta_{\kappa*} n_{\min}}{2} > 1. \tag{4.39}$$

Let $\hat{y}_{(S)}(\boldsymbol{x}_0)$ and $\hat{y}_{BC(S)}(\boldsymbol{x}_0)$ denote $\hat{y}(\boldsymbol{x}_0)$ and $\hat{y}_{BC}(\boldsymbol{x}_0)$ after replacing (4.5) with (4.38), respectively. Then, we have the following result.

Proposition 4.8. *Assume (A-i), (A-i') and (4.8). Under (4.39) it holds that when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$*

$$\hat{y}_{(S)}(\boldsymbol{x}_0) = \frac{\Delta_\kappa}{\Delta_{\kappa*}}\Big((-1)^i + \frac{\delta}{\Delta_\kappa} + o_P(1)\Big) \quad \textit{and} \quad \hat{y}_{BC(S)}(\boldsymbol{x}_0) = \frac{\Delta_\kappa}{\Delta_{\kappa*}}\{(-1)^i + o_P(1)\}.$$

From Proposition 4.8, the bias-corrected smSVM (BC-smSVM) holds the consistency (4.2) even when $|\delta/\Delta_\kappa| \to \infty$. Hence, for smSVMs, we recommend to use the BC-smSVM.

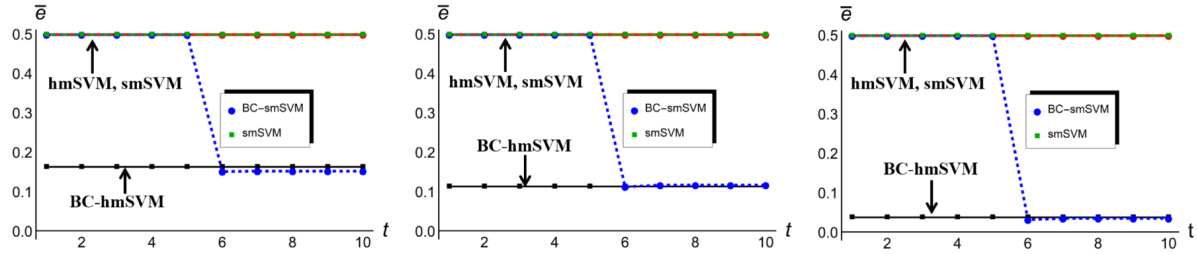
For the settings (a) to (c) in Section 4.2.4, we checked the performance of the BC-smSVM and smSVM together with the hmSVM and bias-corrected hmSVM (BC-hmSVM) for the kernel function (II). We set $(n_1, n_2) = (20, 10)$, $d = 1024$ $(= 2^{10})$ and $\gamma = d/4$. We set $C = 2^{-5+t}/(n_{\min}\Delta_{\kappa*})$, $t = 1, \ldots, 10$, for the smSVMs. Similar to Figure 4.3, we calculated $\overline{e}$ by 2000 replications and plotted the results in Figure 4.10. We observed that smSVMs give bad performances when $C < 2/(n_{\min}\Delta_{\kappa*})$. As expected, the smSVMs are close to the hmSVMs when $C > 2/(n_{\min}\Delta_{\kappa*})$.

Appendix 4

Proof of Lemma 4.1. Note that $L(\boldsymbol{\alpha}) = \sum_{j=1}^{n} \alpha_j - \boldsymbol{\acute{\alpha}}^T \boldsymbol{K} \boldsymbol{\acute{\alpha}}/2$. The result is obtained from (4.7) straightforwardly. □

Table 4.2: The average error rate $\overline{e}$ for five microarray data sets in Table 4.1.

Data set	(n_1, n_2)	BC-GSVM	GSVM	BC-LSVM	LSVM
Colon cancer	$(10, 10)$	0.157	0.158	0.163	0.159
	$(20, 10)$	0.148	0.166	0.159	0.173
	$(30, 10)$	0.135	0.172	0.178	0.213
	$(10, 15)$	0.149	0.15	0.17	0.17
	$(20, 15)$	0.131	0.142	0.154	0.157
	$(30, 15)$	0.133	0.133	0.159	0.181
Leukemia	$(5, 10)$	0.055	0.071	0.06	0.08
	$(10, 10)$	0.041	0.04	0.04	0.041
	$(20, 10)$	0.035	0.041	0.039	0.05
	$(5, 20)$	0.049	0.099	0.049	0.102
	$(10, 20)$	0.037	0.033	0.035	0.041
	$(20, 20)$	0.03	0.029	0.037	0.037
DLBCL	$(10, 5)$	0.082	0.096	0.079	0.079
	$(30, 5)$	0.072	0.096	0.055	0.115
	$(50, 5)$	0.099	0.137	0.069	0.147
	$(10, 15)$	0.042	0.052	0.045	0.054
	$(30, 15)$	0.028	0.027	0.021	0.021
	$(50, 15)$	0.019	0.025	0.017	0.019
HGG	$(5, 10)$	0.282	0.333	0.304	0.316
	$(10, 10)$	0.269	0.277	0.28	0.286
	$(20, 10)$	0.231	0.29	0.288	0.292
	$(5, 15)$	0.279	0.476	0.313	0.344
	$(10, 15)$	0.246	0.387	0.281	0.281
	$(20, 15)$	0.246	0.262	0.268	0.267
Breast cancer	$(3, 3)$	0.226	0.236	0.245	0.239
	$(6, 3)$	0.202	0.264	0.228	0.243
	$(9, 3)$	0.182	0.369	0.234	0.253
	$(3, 5)$	0.218	0.277	0.257	0.276
	$(6, 5)$	0.168	0.176	0.226	0.225
	$(9, 5)$	0.149	0.217	0.211	0.206



(a) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\boldsymbol{\Sigma}_1) = \mathrm{tr}(\boldsymbol{\Sigma}_2)$ (b) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ (c) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$

Figure 4.10: The average error rate, $\bar{e}$, of the BC-smSVM, smSVM, BC-hmSVM and hmSVM with (II) for (a) to (c) when $d = 1024$ and $C = 2^{-5+t}/(n_{\min}\Delta_{\kappa*})$, $t = 1, \ldots, 10$. The average error rates of the BC-smSVM and smSVM are described by the dashed lines and the average error rates of the BC-hmSVM and hmSVM are described by the solid lines.

Proofs of Proposition 4.1 and Proposition 4.3. We assume (A-i) and (A-i'). From Lemma 4.1 it holds that under (4.8)

$$\eta_1 \sum_{j=1}^{n_1} \hat{\alpha}_j^2/\hat{\alpha}_\star^2 = \eta_1/n_1 + o_P(\Delta_\kappa) \quad \text{and} \quad \eta_2 \sum_{j=n_1+1}^{n} \hat{\alpha}_j^2/\hat{\alpha}_\star^2 = \eta_2/n_2 + o_P(\Delta_\kappa), \tag{4.40}$$

so that $L(\hat{\boldsymbol{\alpha}}) = 2\hat{\alpha}_\star - \Delta_{\kappa*}\hat{\alpha}_\star^2\{1 + o_P(\Delta_\kappa/\Delta_{\kappa*})\}/2$. Then, it holds that

$$\hat{\alpha}_\star = (2/\Delta_{\kappa*})\{1 + o_P(\Delta_\kappa/\Delta_{\kappa*})\}. \tag{4.41}$$

Also, from (4.40) we have (4.9) under (4.8).

Next, we consider the second result of Proposition 4.2. Let $\hat{S}_1 = \{j|\hat{\alpha}_j \neq 0,\ j = 1, \ldots, n_1\}$, $\hat{S}_2 = \{j|\hat{\alpha}_j \neq 0,\ j = n_1 + 1, \ldots, n\}$, $\hat{n}_1 = \#\hat{S}_1$ and $\hat{n}_2 = \#\hat{S}_2$. Then, we have that when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$,

$$\begin{aligned}
&\sum_{j=1}^{n} \hat{\alpha}_j t_j k(\boldsymbol{x}_0, \boldsymbol{x}_j) + \frac{1}{n_{\hat{S}}} \sum_{j \in \hat{S}} \Big(t_j - \sum_{j' \in \hat{S}} \hat{\alpha}_{j'} t_{j'} k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) \Big) \\
&= (-1)^i \hat{\alpha}_\star (\kappa_{2i-1} - \kappa_5) + \frac{\hat{n}_2 - \hat{n}_1}{n_{\hat{S}}} \\
&\quad - \hat{\alpha}_\star \Big(\frac{-\kappa_1 \hat{n}_1 - \eta_1 + \kappa_3 \hat{n}_2 + \eta_2 + (\hat{n}_1 - \hat{n}_2)\kappa_5}{n_{\hat{S}}} \Big) + o_P(\Delta_\kappa \hat{\alpha}_\star) \\
&= (-1)^i \hat{\alpha}_\star (\kappa_{2i-1} - \kappa_5) + \frac{(\hat{n}_2 - \hat{n}_1)(1 - \hat{\alpha}_\star \Delta_{\kappa*}/2)}{n_{\hat{S}}} + \frac{\hat{\alpha}_\star(\kappa_1 - \kappa_3)}{2} \\
&\quad + \hat{\alpha}_\star \frac{\eta_1/n_1 - \eta_2/n_2}{2} + \hat{\alpha}_\star \frac{\eta_1(1 - \hat{n}_1/n_1) - \eta_2(1 - \hat{n}_2/n_2)}{n a_{\hat{S}}} + o_P(\Delta_\kappa \hat{\alpha}_\star).
\end{aligned} \tag{4.42}$$

Here, we note that $\eta_1 \sum_{j=1}^{n_1} \hat{\alpha}_j^2/\hat{\alpha}_\star^2 \geq \eta_1/\hat{n}_1$. Thus from (4.40) it holds that

$$\hat{n}_1(\eta_1/\hat{n}_1 - \eta_1/n_1) = \eta_1(1 - \hat{n}_1/n_1) = o_P(\hat{n}_1 \Delta_\kappa) \tag{4.43}$$

under (4.8). Similarly, we have $\eta_2(1 - \hat{n}_2/n_2) = o_P(\hat{n}_2 \Delta_\kappa)$ under (4.8). Then, from (4.41) and (4.42), we have that when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$,

$$\begin{aligned}
\hat{y}(\boldsymbol{x}_0) &= 2(-1)^i \frac{\kappa_{2i-1} - \kappa_5}{\Delta_{\kappa*}} + \frac{\kappa_1 - \kappa_3}{\Delta_{\kappa*}} + \frac{\eta_1/n_1 - \eta_2/n_2}{\Delta_{\kappa*}} + o_P\Big(\frac{\Delta_\kappa}{\Delta_{\kappa*}}\Big) \\
&= (-1)^i \Delta_\kappa/\Delta_{\kappa*} + \delta/\Delta_{\kappa*} + o_P(\Delta_\kappa/\Delta_{\kappa*})
\end{aligned} \tag{4.44}$$

under (4.8). Hence, we conclude the second result of Proposition 4.1.

Finally, we consider the proof of Proposition 4.2. In view of (4.13), we claim the first result. By noting that $\Delta_{\kappa*}/\Delta_\kappa \to 1$ and $\delta/\Delta_\kappa = o(1)$ under (4.12), it holds from (4.42) that $\hat{y}(\boldsymbol{x}_0) = (-1)^i + o_P(1)$ under (4.12). We conclude the second result. □

Proofs of Theorem 4.1 and Corollary 4.1. We assume (A-i) and (A-i'). We consider the following conditions:

$$\liminf_{d\to\infty} \eta_2/(n_2\Delta_\kappa) > 0 \ \text{ and } \ \eta_1/(n_1\Delta_\kappa) = o(1). \tag{4.45}$$

Let $\Delta_{*2} = \Delta_\kappa + \eta_2/n_2$. Note that $\eta_1 \sum_{j=1}^{n_1} \hat{\alpha}_j^2/\hat{\alpha}_\star^2 = o_P(\Delta_\kappa)$ under (4.45). Similar to (4.41), it holds from (4.42) and (4.43) that $\hat{\alpha}_\star = (2/\Delta_{*2})\{1 + o_P(\Delta_\kappa/\Delta_{*2})\}$ and

$$\hat{y}(\boldsymbol{x}_0) = (-1)^i \Delta_\kappa/\Delta_{*2} + \delta/\Delta_{*2} + o_P(\Delta_\kappa/\Delta_{*2}) \tag{4.46}$$

under (4.45) when $\boldsymbol{x}_0 \in \pi_i$ for $i = 1, 2$. Note that $\Delta_{\kappa*}/\Delta_\kappa \to 1$ and $\delta/\Delta_{\kappa*} \to 0$ under (4.12) and $\Delta_{\kappa*}/\Delta_{*2} \to 1$ under (4.45). From Propositions 4.1, 4.2 and (4.46), we obtain (4.44) without (4.8). Thus, from (4.44), we conclude the results of Theorem 4.1 and Corollary 4.1. □

Proofs of Lemma 4.2 and Theorem 4.2. Under (A-i), it holds that $\widehat{\Delta}_{\kappa*} = \Delta_{\kappa*} + o_P(\Delta_\kappa)$ and $\hat{\eta}_i = \eta_i + o_P(\Delta_\kappa)$ for $i = 1, 2$. Thus we can conclude the result of Lemma 4.2. From the proofs of Theorem 4.1 and Corollary 4.1, we obtain (4.44) under (A-i). By combining (4.44) with Lemma 4.2, we conclude the result of Theorem 2.2. □

Proofs of Lemma 4.3, Corollaries 4.2 and 4.3. We assume (A-ii) and (C-ii). Assume also $\boldsymbol{\mu}_2 = \boldsymbol{0}$ without loss of generality. Note that $\kappa_1 = \|\boldsymbol{\mu}_1\|^2$, $\kappa_2 = \|\boldsymbol{\mu}_1\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1)$, $\kappa_3 = \kappa_5 = 0$, $\kappa_4 = \eta_{2(I)}$ and $\Delta_{(I)} = \|\boldsymbol{\mu}_1\|^2$. Also, note that

$$\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_i \boldsymbol{\mu}_1 \le \Delta_{(I)} \lambda_{\max}(\boldsymbol{\Sigma}_i) \le \Delta_{(I)} \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}. \tag{4.47}$$

Then, by using Chebyshev's inequality, for any $\tau > 0$ we have that

$$\begin{aligned} &\sum_{j=1}^{n_1} P(|\boldsymbol{\mu}_1^T(\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1)| \ge \tau\Delta_{(I)}) \le n_1(\tau\Delta_{(I)})^{-4} E[\{\boldsymbol{\mu}_1^T(\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1)\}^4] \\ &= O\Big\{n_1\Big((\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_i\boldsymbol{\mu}_1)^2 + \sum_{r=1}^{p_1}(\boldsymbol{\gamma}_r^T\boldsymbol{\mu}_1)^4\Big)/\Delta_{(I)}^4\Big\} = O(n_1 \mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_{(I)}^2) \to 0 \end{aligned} \tag{4.48}$$

from the fact that $\sum_{r=1}^{p_1}(\boldsymbol{\gamma}_r^T\boldsymbol{\mu}_1)^4 \le (\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_i\boldsymbol{\mu}_1)^2$, where $\boldsymbol{\Gamma}_1 = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{p_1}]$. On the other hand, we have that

$$\begin{aligned} &\sum_{j<j'}^{n_i} P(|(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)^T(\boldsymbol{x}_{ij'} - \boldsymbol{\mu}_i)| \ge \tau\Delta_{(I)}) \\ &\le \sum_{j<j'}^{n_i} (\tau\Delta_{(I)})^{-4} E[\{(\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1)^T(\boldsymbol{x}_{1j'} - \boldsymbol{\mu}_1)\}^4] = O\Big(n_i^2 \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^2/\Delta_{(I)}^4\Big) \to 0. \end{aligned} \tag{4.49}$$

Note that $\boldsymbol{x}_{1j}^T\boldsymbol{x}_{1j'} - \kappa_1 = (\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1)^T(\boldsymbol{x}_{1j'} - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_1^T(\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1 + \boldsymbol{x}_{1j'} - \boldsymbol{\mu}_1)$. Thus, from (4.48) and (4.49), it holds that

$$\boldsymbol{x}_{1j}^T\boldsymbol{x}_{1j'} = \kappa_1 + o_P(\Delta_{(I)}) \ \text{ for all } j < j' \le n_1. \tag{4.50}$$

Note that

$$\sum_{j=1}^{n_1}\sum_{j'=1}^{n_2} P(|(\boldsymbol{x}_{1j}-\boldsymbol{\mu}_1)^T(\boldsymbol{x}_{2j'}-\boldsymbol{\mu}_2)| \geq \tau\Delta_{(I)})$$
$$= O\Big(n_1 n_2\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}/\Delta_{(I)}^4\Big) \to 0 \tag{4.51}$$

from the fact that $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) \leq \{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}^2$. Then, similar to (4.50), we have that

$$\begin{aligned}
&\boldsymbol{x}_{2j}^T\boldsymbol{x}_{2j'} = \kappa_3 + o_P(\Delta_{(I)}) \text{ for all } j < j' \leq n_2,\\
&\boldsymbol{x}_{1j}^T\boldsymbol{x}_{2j'} = \kappa_5 + o_P(\Delta_{(I)}) \text{ for all } j = 1,\ldots,n_1;\ j' = 1,\ldots,n_2,\\
&\boldsymbol{x}_0^T\boldsymbol{x}_{ij} = \kappa_{2i-1} + o_P(\Delta_{(I)}) \text{ for all } 1 \leq j \leq n_i, i = 1,2, \text{ when } \boldsymbol{x}_0 \in \pi_i\\
&\text{and } \boldsymbol{x}_0^T\boldsymbol{x}_{i'j} = \kappa_5 + o_P(\Delta_{(I)}) \text{ for all } 1 \leq j \leq n_i, i = 1,2\ (i' \neq i) \text{ when } \boldsymbol{x}_0 \in \pi_i.
\end{aligned}$$

In addition, for any $\tau > 0$ we have that

$$\sum_{j=1}^{n_i} P\big(\big|\|\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i\|^2 - \mathrm{tr}(\boldsymbol{\Sigma}_i)\big| \geq \tau\Delta_{(I)}\big) = O\Big(n_i\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_{(I)}^2\Big) \to 0 \tag{4.52}$$

for $i = 1,2$. Thus, from (4.48) and (4.52), it holds that for all $j = 1,\ldots,n_i;\ i = 1,2$

$$\boldsymbol{x}_{ij}^T\boldsymbol{x}_{ij} = \kappa_{2i} + o_P(\Delta_{(I)}).$$

It concludes Lemma 4.3.

For the proofs of Corollaries 4.2 and 4.3, from Theorems 4.1, 4.2 and Corollary 4.1, we conclude the results. □

Proofs of Lemma 4.4, Corollaries 4.4 and 4.5. We assume (A-ii). Let $\Omega = \min\{\gamma\Delta_{(II)}/\psi,\ \gamma\}$. Similar to (4.48), for any $\tau > 0$, we have that under (C-iii)

$$\sum_{j=1}^{n_i} P(|(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i)| \geq \tau\Omega) \to 0$$

for $i = 1,2$, so that $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i) = o_P(\Omega)$ for all $j = 1,\ldots,n_i;\ i = 1,2$. Similarly, $\|\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i\|^2 = \mathrm{tr}(\boldsymbol{\Sigma}_i) + o_P(\Omega)$ for all $j = 1,\ldots,n_i;\ i = 1,2$, and $(\boldsymbol{x}_{1j}-\boldsymbol{\mu}_1)^T(\boldsymbol{x}_{2j'}-\boldsymbol{\mu}_2) = o_P(\Omega)$ for all $j = 1,\ldots,n_1;\ j' = 1,\ldots,n_2$. Then, under (C-iii), we have that for all $j = 1,\ldots,n_1;\ j' = 1,\ldots,n_2$

$$\begin{aligned}
\exp(-\|\boldsymbol{x}_{1j}-\boldsymbol{x}_{2j'}\|^2/\gamma) &= \exp(-\|(\boldsymbol{x}_{1j}-\boldsymbol{\mu}_1)-(\boldsymbol{x}_{2j'}-\boldsymbol{\mu}_2)+\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2/\gamma)\\
&= \kappa_{5(II)} + o_P(\kappa_{5(II)}\Omega/\gamma) = \kappa_{5(II)} + o_P(\Delta_{(II)})
\end{aligned} \tag{4.53}$$

from the fact that $\kappa_{5(II)} \leq \psi$. Similar to (4.53), we can conclude that the assumptions (A-i) and (A-i') are met. It concludes Lemma 4.4.

For the proofs of Corollaries 4.4 and 4.5, from Theorems 4.1, 4.2 and Corollary 4.1, we conclude the results. □

Proofs of Propositions 4.3 and 4.4. From (4.23), (C-iii) holds under (C-ii) and (C-v). Thus, from (4.44) and Lemmas 4.2 to 4.4, we conclude Proposition 4.4. For the proof of Proposition 4.3, we note that $\mathrm{tr}(\boldsymbol{\Sigma}_i)/\gamma \to 0$ for $i = 1,2$, under (C-iv) from the fact that $\Delta_{(I)} = O(d)$. Thus it holds that $\psi \to 1$ and $\gamma\eta_{i(II)} = 2\mathrm{tr}(\boldsymbol{\Sigma}_i) + O(d^2/\gamma)$ for $i = 1,2$, under (C-iv). In addition, from (4.22) it holds that $\delta_{(II)}/\Delta_{(II)} = \delta_{(I)}\{1+o(1)\}/\Delta_{(I)} + o(1)$ under (C-iv). Thus from (4.44), Lemmas 4.3 and 4.4, we conclude Proposition 4.3. □

Proof of Proposition 4.5. We set that $\kappa_1 = (\zeta + \|\boldsymbol{\mu}_1\|^2)^r$, $\kappa_2 = (\zeta + \mathrm{tr}(\boldsymbol{\Sigma}_1) + \|\boldsymbol{\mu}_1\|^2)^r$, $\kappa_3 = (\zeta + \|\boldsymbol{\mu}_2\|^2)^r$, $\kappa_4 = (\zeta + \mathrm{tr}(\boldsymbol{\Sigma}_2) + \|\boldsymbol{\mu}_2\|^2)^r$ and $\kappa_5 = (\zeta + \boldsymbol{\mu}_1^T \boldsymbol{\mu}_2)^r$. From (4.1) we note that $\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_{i'} \boldsymbol{\mu}_i \leq \|\boldsymbol{\mu}_i\|^2 \lambda_{\max}(\boldsymbol{\Sigma}_i) = o(d^2)$ as $d \to \infty$ for $i, i' = 1, 2$. Then, similar to (4.46) to (4.48), for the polynomial kernel, we have that $\boldsymbol{x}_{ij}^T \boldsymbol{x}_{ij'} = \|\boldsymbol{\mu}_i\|^2 + o_P(d)$ for all $j < j'$, $i = 1, 2$, $\boldsymbol{x}_{ij}^T \boldsymbol{x}_{ij} = \mathrm{tr}(\boldsymbol{\Sigma}_i) + \|\boldsymbol{\mu}_i\|^2 + o_P(d)$ for all i, j, and $\boldsymbol{x}_{1j}^T \boldsymbol{x}_{2j'} = \boldsymbol{\mu}_1^T \boldsymbol{\mu}_2 + o_P(d)$ for all j, j', so that $k(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}) = \kappa_{2i-1} + o_P(d^r)$ for all $j < j'$, $i = 1, 2$, $k(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij}) = \kappa_{2i} + o_P(d^r)$ for all i, j, and $k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j'}) = \kappa_5 + o_P(d^r)$ for all j, j'. Here, note that

$$(\zeta + \|\boldsymbol{\mu}_1\|^2)^r + (\zeta + \|\boldsymbol{\mu}_2\|^2)^r - 2(\zeta + \boldsymbol{\mu}_1^T \boldsymbol{\mu}_2)^r \geq \{(\zeta + \|\boldsymbol{\mu}_1\|^2)^{r/2} - (\zeta + \|\boldsymbol{\mu}_2\|^2)^{r/2}\}^2$$

from the fact that $(\zeta + \boldsymbol{\mu}_1^T \boldsymbol{\mu}_2)^r \leq (\zeta + \|\boldsymbol{\mu}_1\|^2)^{r/2} (\zeta + \|\boldsymbol{\mu}_2\|^2)^{r/2}$. Then, it holds that $\liminf_{d \to \infty} \Delta_\kappa / d^r > 0$ from (4.25). Thus, we have (A-i). Similarly, we can conclude (A-i'). From Theorem 4.2, the BC-SVM (4.17) holds (4.2) for the polynomial kernel. It concludes Proposition 4.5. □

Proof of Proposition 4.6. We assume (4.26). Note that $1/\omega \to 0$ under (4.28). First, we consider the case when $\limsup_{d \to \infty} \gamma_\star < \infty$. Then, it holds that $F(\gamma_\star) = \{1 + o(1)\}/\gamma_\star$, so that $\liminf_{d \to \infty} F(\gamma_\star) > 0$. Next, we consider the case when $\gamma_\star \to \infty$. Let $\nu = \omega / \gamma_\star$ (> 0). Note that $\nu = \Delta_{(I)}/\gamma$. Then, it holds that

$$\omega F(\gamma_\star) = \nu + \frac{2\nu \exp(-\nu)\{1 + o(\nu)\}}{\{1 - \exp(-\nu)\} + o(\nu)}.$$

Let $g(\nu) = \nu + 2\nu \exp(-\nu)/\{1 - \exp(-\nu)\}$. Note that $g(\nu)$ is a monotonically increasing function and $g(\nu) \to 2$ as $\nu \to 0$, so that $F(\gamma_\star) = 2\{1 + o(1)\}/\omega = o(1)$ when $\nu \to 0$. We can conclude the result. □

Proof of Proposition 4.6. When $\omega \leq 1$, it holds that $F(\gamma_\star) = 2\{1 + o(1)\}/\omega$ under $\gamma_\star \to \infty$. When $\omega \leq 1$ and $\gamma_\star = 1$, it holds that

$$F(\gamma_\star) = 1 + \frac{4}{\exp(\omega + 1) + \exp(\omega - 1) - 2} < 1 + 1/\omega \leq 2/\omega$$

from the facts that $\exp(\omega + 1) > 1 + (\omega + 1) + (\omega + 1)^2/2 \geq 2 + 3\omega$ and $\exp(\omega - 1) \geq \omega$. Hence, when $\omega \leq 1$, we have that $\Delta_\Sigma / \gamma_0 \in (0, \infty)$ as $d \to \infty$. It concludes the result. □

Proof of Proposition 4.7. From Proposition 4.1, Lemma 4.2 and (4.11), we can conclude the results. □

Bibliography

Ahn, J., Marron, J., Muller, K. M., and Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):760–766.

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

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750.

Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Analysis (Editor's special invited paper)*, 30(4):356–399.

Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, 66(5):983–1010.

Aoshima, M. and Yata, K. (2015a). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions. *Methodology and Computing in Applied Probability*, 17(2):419–439.

Aoshima, M. and Yata, K. (2015b). Geometric classifier for multiclass, high-dimensional data. *Sequential Analysis*, 34(3):279–294.

Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica*, 28(1):43–62.

Aoshima, M. and Yata, K. (2019a). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Annals of the Institute of Statistical Mathematics*, 71:473–503.

Aoshima, M. and Yata, K. (2019b). High-dimensional quadratic classifiers in non-sparse settings. *Methodology and Computing in Applied Probability*, 21:663–682.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., and Korsmeyer, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–47.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6(2):311–329.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97:1382–1408.

Benjamin, X. W. and Nathalie, J. (2010). Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, 25(1):1–20.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Carmichael, I. and Marron, J. (2017). Geometric insights into support vector machine behavior using the kkt conditions. *arXiv preprint arXiv:1704.00767*.

Chan, Y.-B. and Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, 96(2):469–478.

Chang, J. C., Wooten, E. C., Tsimelzon, A., Hilsenbeck, S. G., Gutierrez, M. C., Elledge, R., Mohsin, S., Osborne, C. K., Chamness, G. C., Allred, D. C., and O'Connell, P. (2003). Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *The Lancet*, 362(9381):362–369.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics*, 38(2):808–835.

Chen, S. X., Zhang, L.-X., and Zhong, P.-S. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, 105(490):810–819.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Friedman, J. (1996). Another Approach to Polychotomous Classification. Technical report.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Huang, H. (2017). Asymptotic behavior of support vector machine for spiked population model. *Journal of Machine Learning Research*, 18(45):1–21.

Ishii, A., Yata, K., and Aoshima, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. *Journal of Statistical Planning and Inference*, 170:186–199.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130.

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, 30(4):1081–1102.

Naderi, A., Teschendorff, A. E., Barbosa-Morais, N. L., Pinder, S. E., Green, A. R., Powe, D. G., Robertson, J. F. R., Aparicio, S., Ellis, I. O., Brenton, J. D., and Caldas, C. (2007). A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, 26(10):1507–1516.

Nagao, H. (1973). On some test criteria for covariance matrix. *Annals of Statistics*, 1(4):700–709.

Nakayama, Y. (2019). Robust support vector machine for high-dimensional imbalanced data. *Communications in Statistics - Simulation and Computation, in press (doi: 10.1080/03610918.2019.1586922).*

Nakayama, Y., Yata, K., and Aoshima, M. (2017). Support vector machine and its bias correction in high-dimension, low-sample-size settings. *Journal of Statistical Planning and Inference*, 191:88–100.

Nakayama, Y., Yata, K., and Aoshima, M. (2019). Bias-corrected support vector machine with gaussian kernel in high-dimension, low-sample-size settings. *Annals of Institute of Mathematical Statistics, in press (doi: 10.1007/s10463-019-00727-1).*

Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., Black, P. M., von Deimling, A., Pomeroy, S. L., Golub, T. R., and Louis, D. N. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, 63(7):1602–1607.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642.

Qiao, X. and Zhang, L. (2015). Flexible high-dimensional classification machines and their asymptotic properties. *Journal of Machine Learning Research*, 16(45):1547–1572.

Shen, D., Shen, H., Zhu, H., and Marron, J. S. (2016). The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica*, 26(4):1747–1770.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., and Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74.

Srivastava, M. S., Kolloand, T., and von Rosen, D. (2011). Some tests for the covariance matrix with fewer observations than the dimension under non-normal. *Journal of Multivariate Analysis*, 102(6):1090–1103.

Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., and Shaughnessy, J. D. J. (2003). The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. *New England Journal of Medicine*, 349(26):2483–2494.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory (second ed.)*. Springer, New York.

Yata, K. and Aoshima, M. (2009). PCA consistency for non-gaussian data in high dimension, low sample size context. *Communications in Statistics - Theory and Methods, Special Issue Honoring Zacks, S. (ed. Mukhopadhyay, N.)*, 38:2634–2652.

Yata, K. and Aoshima, M. (2012a). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis*, 105(1):193–215.

Yata, K. and Aoshima, M. (2012b). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of Multivariate Analysis*, 101(9):2060–2077.

Yata, K. and Aoshima, M. (2013). Correlation tests for high-dimensional data using extended cross-data-matrix methodology. *Journal of Multivariate Analysis*, 117:313–331.

Yata, K. and Aoshima, M. (2016). High-dimensional inference on covariance structures via the extended cross-data-matrix methodology. *Journal of Multivariate Analysis*, 151:151–166.

Yata, K., Aoshima, M., and Nakayama, Y. (2018). A test of sphericity for high-dimensional data and its application for detection of divergently spiked noise. *Sequential Analysis*, 37:397–411.