Annals of the Institute of Statistical Mathematics (2020) 72:1257–1286
https://doi.org/10.1007/s10463-019-00727-1

Bias-corrected support vector machine with Gaussian kernel in high-dimension, low-sample-size settings

Yugo Nakayama[1] · Kazuyoshi Yata[2] · Makoto Aoshima[2]

Received: 28 September 2018 / Revised: 15 April 2019 / Published online: 15 July 2019

Abstract

In this paper, we study asymptotic properties of nonlinear support vector machines (SVM) in high-dimension, low-sample-size settings. 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 actual data analyses.

Keywords Geometric representation · HDLSS · Imbalanced data · Radial basis function kernel

We are very grateful to the associate editor and the reviewer for their constructive comments. The research of the second author was partially supported by Grant-in-Aid for Scientific Research (C), Japan Society for the Promotion of Science (JSPS), under Contract Number 18K03409. The research of the third author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 15H01678 and 17K19956.

✉ Makoto Aoshima
aoshima@math.tsukuba.ac.jp

Yugo Nakayama
n-yougo@math.tsukuba.ac.jp

Kazuyoshi Yata
yata@math.tsukuba.ac.jp

[1] Graduate School of Pure and Applied Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8571, Japan

[2] Institute of Mathematics, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8571, Japan

1 Introduction

A common feature of high-dimensional data is that the data dimension is high; however, the sample size is relatively low. We call such data "HDLSS" data. The current work handles the classification problem in the HDLSS framework. 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, $\boldsymbol{x}_{i1},\dots,\boldsymbol{x}_{in_i}$, from each Π_i. We assume $n_i \ge 2$. 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 $d,N\to\infty$.

In the HDLSS context, Hall et al. (2005), Marron et al. (2007) and Qiao et al. (2010) considered distance weighted classifiers. Hall et al. (2008), Chan and Hall (2009) and Aoshima and Yata (2014) considered distance-based classifiers. Aoshima and Yata (2019) considered a distance-based classifier based on a data transformation technique. Aoshima and Yata (2011, 2015) considered geometric classifiers based on a geometric representation of HDLSS data. Aoshima and Yata (2018b) considered quadratic classifiers in general and discussed an optimality of the classifiers under high-dimension, non-sparse settings. On the other hand, Hall et al. (2005), Chan and Hall (2009), Qiao and Zhang (2015) and Nakayama et al. (2017) investigated asymptotic properties of the linear support vector machine (SVM) in the HDLSS context. Huang (2017) investigated the SVM in the high-dimension, large-sample-size context as $d/N\to c>0$. Vapnik (2000), Schölkopf and Smola (2002), Hall et al. (2005) and Qiao and Zhang (2015) investigated the versatility of the SVM for both low-dimensional and high-dimensional data. Hall et al. (2005), Chan and Hall (2009) and Qiao and Zhang (2015) showed that the misclassification rates of the linear SVM tend to zero as $d\to\infty$ under certain strict conditions in the HDLSS context. Under mild conditions in the HDLSS context, Nakayama et al. (2017) pointed out the strong inconsistency of the linear SVM when n_is are imbalanced. Nakayama et al. (2017) gave a bias-corrected linear SVM and showed its superiority to the linear SVM. As long as we know, asymptotic properties of nonlinear SVMs seem not to have been sufficiently studied in the HDLSS context. In the current paper, we investigate nonlinear SVMs in the HDLSS context.

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

$$\operatorname{tr}(\boldsymbol{\Sigma}_i^2)/\operatorname{tr}(\boldsymbol{\Sigma}_i)^2 \to 0 \text{ as } d\to\infty. \tag{1}$$

We note that the ratio, $\operatorname{tr}(\boldsymbol{\Sigma}_i^2)/\operatorname{tr}(\boldsymbol{\Sigma}_i)^2$, is a measure of sphericity and (1) is equivalent to "$\lambda_{\max}(\boldsymbol{\Sigma}_i)/\operatorname{tr}(\boldsymbol{\Sigma}_i)\to 0$ 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 (2019). If we assume (1) and (A-ii) given in Sect. 3, we have that

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

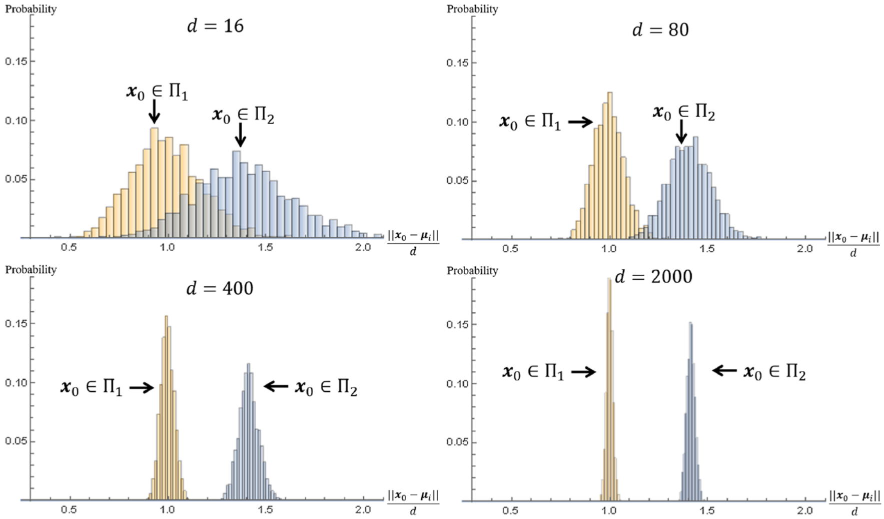


Fig. 1 The histograms of $\|\boldsymbol{x}_0 - \boldsymbol{\mu}_i\|/d^{1/2}$ for $\boldsymbol{x}_0 \in \Pi_i$, $i = 1, 2$, when $d = 16, 80, 400$ and 2000

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 (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 Fig. 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 center $\boldsymbol{\mu}_i$ and radius $\mathrm{tr}(\boldsymbol{\Sigma}_i)^{1/2}$ as in Fig. 2. In this paper, we focus on the geometric representation for high-dimensional classification.

In Sect. 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 Sect. 3, we give asymptotic properties of the BC-SVM for both 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 Sect. 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. Finally, in Sect. 5, we examine the performance of the BC-SVM with the Gaussian kernel for several choices of the scale parameter by numerical simulations and actual data analyses.

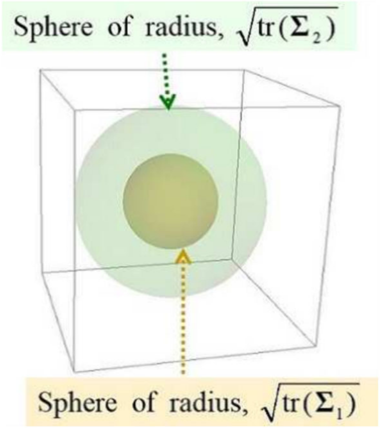

Sphere of radius, $\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_2)}$
Sphere of radius, $\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_1)}$

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

2 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) $d \to \infty$ either when $N \to \infty$ as $d \to \infty$ or N is fixed.

2.1 Setup of SVM

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

$$y(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b, \tag{2}$$

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:

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^{N} \alpha_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{j'=1}^{N} \alpha_j \alpha_{j'} t_j t_{j'} k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}), \tag{3}$$

where $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \phi(\boldsymbol{x}_j)^{\mathrm{T}}\phi(\boldsymbol{x}_{j'})$ is a kernel function, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^{\mathrm{T}}$ 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: $\mathrm{argmax}_{\boldsymbol{\alpha}}\, L(\boldsymbol{\alpha})$ subject to

$$\alpha_j \geq 0,\ j = 1, \ldots, N, \ \text{ and } \sum_{j=1}^{N} \alpha_j t_j = 0. \tag{4}$$

Let us write that

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N)^{\mathrm{T}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L(\boldsymbol{\alpha}) \ \text{ subject to (4)}.$$

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 (2) is given by

$$\hat{y}(\boldsymbol{x}) = \sum_{j \in \hat{S}} \hat{\alpha}_j t_j k(\boldsymbol{x}, \boldsymbol{x}_j) + \hat{b}. \tag{5}$$

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.

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

$$e(i) \to 0 \quad \text{as } d \to \infty \text{ for } i = 1, 2. \tag{6}$$

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

(I) The linear kernel: $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{x}_{j'}$ 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 Sect. 4. We examine the following kernels numerically.

(III) The polynomial kernel: $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = (\zeta + \boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{x}_{j'})^r$ 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.

We also investigate the soft-margin SVM in Sect. 6.

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 under the divergence condition (D):

(A-i) $k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j'}) = \kappa_1 + o_P(\Delta)$ for all $1 \leq j < j' \leq n_1$,
$k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j}) = \kappa_2 + o_P(\Delta)$ for all $1 \leq j \leq n_1$,
$k(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j'}) = \kappa_3 + o_P(\Delta)$ for all $1 \leq j < j' \leq n_2$,
$k(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j}) = \kappa_4 + o_P(\Delta)$ for all $1 \leq j \leq n_2$,
and $k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j'}) = \kappa_5 + o_P(\Delta)$ for all $1 \leq j \leq n_1$ and $1 \leq j' \leq n_2$,

where $\Delta = \kappa_1 + \kappa_3 - 2\kappa_5$ 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 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, $\Delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, $\kappa_1 = \|\boldsymbol{\mu}_1\|^2$, $\kappa_2 = \|\boldsymbol{\mu}_1\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_1)$, $\kappa_3 = \|\boldsymbol{\mu}_2\|^2$, $\kappa_4 = \|\boldsymbol{\mu}_2\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_2)$ and $\kappa_5 = \boldsymbol{\mu}_1^T \boldsymbol{\mu}_2$ when $k(\cdot,\cdot)$ is the linear kernel. See Sect. 3.1. Also, see Sects. 3.2 and 7 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

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

where $\boldsymbol{J}_{n_1,n_2}$ denotes the $n_1 \times n_2$ matrix with all the elements 1. Let $\acute{\boldsymbol{\alpha}} = (-\alpha_1, \ldots, -\alpha_{n_1}, \alpha_{n_1+1}, \ldots, \alpha_N)^{\mathrm{T}}$. We note that $\sum_{j=1}^{n_1} \alpha_j = \sum_{j=n_1+1}^{N} \alpha_j$ $(= \alpha_\star, \ \text{say})$ under (4). Then, it holds that

$$\acute{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{K}_0 \acute{\boldsymbol{\alpha}} = \Delta \alpha_\star^2 + \eta_1 \sum_{j=1}^{n_1} \alpha_j^2 + \eta_2 \sum_{j=n_1+1}^{N} \alpha_j^2. \tag{7}$$

The second and third terms in (7) are regarded as a bias part. See Proposition 1. We have that $L(\boldsymbol{\alpha}) = 2\alpha_\star - \acute{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{K} \acute{\boldsymbol{\alpha}}/2$ under (4). Then, from (7) we claim the following lemma.

Lemma 1 *Under (4), (A-i) and (D), it holds that*

$$L(\boldsymbol{\alpha}) = 2\alpha_\star - \frac{\Delta}{2}\alpha_\star^2 - \frac{1}{2}\left(\eta_1 \sum_{j=1}^{n_1} \alpha_j^2 + \eta_2 \sum_{j=n_1+1}^{N} \alpha_j^2 \right) + o_P(\Delta \alpha_\star^2).$$

Note that

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

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). We first consider the following condition under (D):

$$\liminf_{d \to \infty} \frac{\eta_i}{n_i \Delta} > 0 \ \ \text{for } i = 1, 2. \tag{8}$$

Let $\Delta_* = \Delta + \eta_1/n_1 + \eta_2/n_2$. Note that $2\alpha_\star - \Delta_* \alpha_\star^2/2 = -\Delta_*(\alpha_\star - 2/\Delta_*)^2/2 + 2/\Delta_*$. Then, in a way similar to Sect. 2 of Nakayama et al. (2017), it follows from Lemma 1 that

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

under (4), (8), (A-i) and (D), so that $\alpha_\star \approx 2/\Delta_*$. 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 1 *Assume (A-i) and (8). It holds that*

$$\hat{\alpha}_\star = (2/\Delta_*)\{1+o_P(1)\},$$
$$\sum_{j=1}^{n_1}\hat{\alpha}_j^2 = \frac{4}{\Delta_*^2 n_1}\{1+o_P(1)\} \text{ and } \sum_{j=n_1+1}^{N}\hat{\alpha}_j^2 = \frac{4}{\Delta_*^2 n_2}\{1+o_P(1)\} \tag{9}$$

under (D). We also assume

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

It holds that under (D)

$$\hat{y}(\boldsymbol{x}_0) = \frac{\Delta}{\Delta_*}\left((-1)^i + \frac{\delta}{\Delta} + o_P(1)\right) \quad \text{when } \boldsymbol{x}_0 \in \Pi_i \text{ for } i = 1, 2, \tag{10}$$

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

We note that "δ/Δ" is a (normalized) bias term of the SVM. From Proposition 1, under (A-i) and (8), it holds that $\sum_{j=1}^{n_1}(\hat{\alpha}_j - \hat{\alpha}_\star/n_1)^2 = o_P\{(n_1\Delta_*^2)^{-1}\}$ and $\sum_{j=n_1+1}^{N}(\hat{\alpha}_j - \hat{\alpha}_\star/n_2)^2 = o_P\{(n_2\Delta_*^2)^{-1}\}$, so that

$$\hat{\alpha}_j = \frac{2}{\Delta_* n_1}\{1+o_P(1)\} \quad \text{for all } j = 1, \ldots, n_1; \quad \text{and}$$
$$\hat{\alpha}_j = \frac{2}{\Delta_* n_2}\{1+o_P(1)\} \quad \text{for all } j = n_1+1, \ldots, N \tag{11}$$

when $d \to \infty$ while N is fixed. It should be noted that all the data points are support vectors under (A-i) and (8) in the HDLSS context. Ahn and Marron (2010) called this phenomenon the "data piling."

Next, we consider the following condition instead of (8) under (D):

$$\frac{\eta_i}{n_i\Delta} = o(1) \quad \text{for } i = 1, 2. \tag{12}$$

It follows from Lemma 1 that

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = -\frac{\Delta}{2}\left(\alpha_\star - \frac{2+o_P(1)}{\Delta}\right)^2\{1+o_P(1)\} + \frac{2+o_P(1)}{\Delta} \tag{13}$$

under (4), (12), (A-i) and (D), so that $\alpha_\star \approx 2/\Delta$.

Proposition 2 *Assume (A-i) and* (12). *It holds that* $\hat{\alpha}_\star = (2/\Delta)\{1 + o_P(1)\}$ *under (D). Furthermore, we assume (A-i'). It holds that under (D)*

$$\hat{y}(\boldsymbol{x}_0) = (-1)^i + o_P(1) \quad \textit{when } \boldsymbol{x}_0 \in \Pi_i \textit{ for } i = 1, 2. \tag{14}$$

It should be noted that the data piling does not occur under (12). However, $\hat{y}(\boldsymbol{x}_0)$ has the consistency in the sense of (14). We consider the following condition under (D):

(C-i) $\limsup_{d\to\infty} \dfrac{|\delta|}{\Delta} < 1.$

Note that (C-i) is met under (12). From Proposition 1, "δ/Δ" is a normalized bias term of the SVM. From (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), (A-i') and (D). Thus, we have the following result.

Theorem 1 *Under (A-i), (A-i'), (C-i) and (D), the SVM* (5) *holds the consistency* (6).

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

Corollary 1 *Under (A-i), (A-i') and (D), the SVM* (5) *holds the following properties:*

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

$$\textit{if } \liminf_{d\to\infty} \frac{\delta}{\Delta} > 1, \textit{ and} \tag{15}$$

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

$$\textit{if } \limsup_{d\to\infty} \frac{\delta}{\Delta} < -1. \tag{16}$$

Remark 1 For the linear SVM, Hall et al. (2005), Qiao and Zhang (2015) and Nakayama et al. (2017) showed the consistency (6) and the results in Corollary 1.

From Corollary 1, if $|\delta|$ is larger than Δ, 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 imbalanced data are called the "extremely imbalanced data." In such cases, the SVM brings the strong inconsistency property as "$e(1) = 1 + o(1)$" when $\eta_1 = \eta_2$, $\Delta/\eta_i = o(1)$ and n_1 is fixed but $n_2 \to \infty$. In order to overcome such difficulties, we propose a bias-corrected SVM.

2.3 Bias-corrected nonlinear SVM

Let

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

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

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

Lemma 2 *Under (A-i) and (D) it holds that*

$$\hat{\varDelta}/\varDelta=1+o_P(1) \quad \textit{and} \quad \hat{\delta}/\hat{\varDelta}_*=\delta/\varDelta_*+o_P\big(\varDelta/\varDelta_*\big).$$

From Proposition 1 and Lemma 2, we give a bias-corrected SVM (BC-SVM) as follows:

$$\hat{y}_{BC}(\boldsymbol{x}_0)=\hat{y}(\boldsymbol{x}_0)-\frac{\hat{\delta}}{\hat{\varDelta}_*}. \tag{17}$$

One classifies $\boldsymbol{x}_0$ into $\varPi_1$ if $\hat{y}_{BC}(\boldsymbol{x}_0)<0$ and into $\varPi_2$ otherwise. We have the following result.

Theorem 2 *Under (A-i), (A-i') and (D), the BC-SVM* (17) *holds the consistency* (6).

It should be noted that the BC-SVM (17) claims the consistency without (C-i) even when $|\delta/\varDelta|\to\infty$.

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 (17) can theoretically ensure the accuracy and have the consistency property at a low computational cost even for extremely imbalanced data.

Remark 2 Nakayama et al. (2017) gave a bias-corrected linear SVM. In this paper, we generalize the concept of the BC-SVM to nonlinear kernels.

2.4 Performance of the BC-SVM

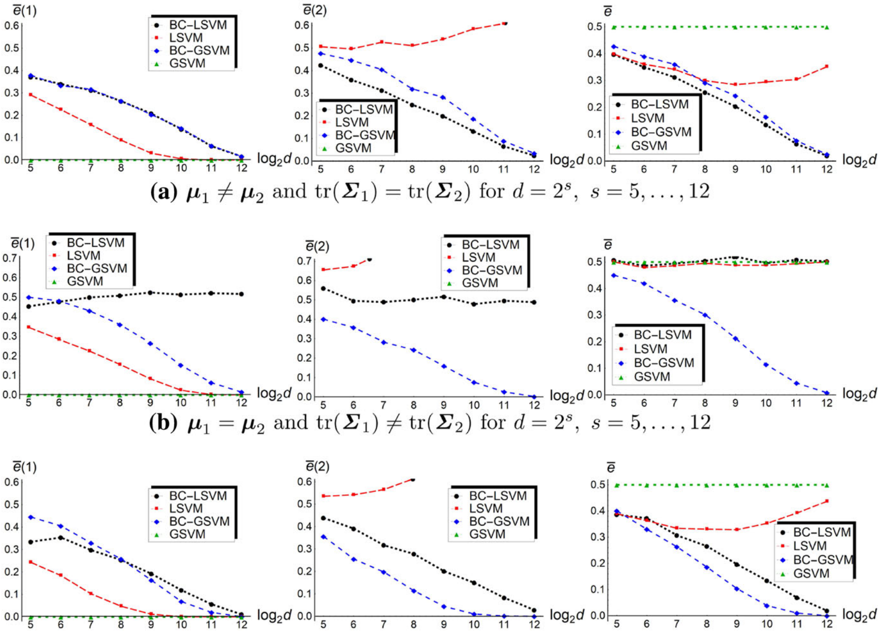
We set $\varPi_i: N_d(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i), i=1,2$, having $\boldsymbol{\mu}_2=\boldsymbol{0}$, $\boldsymbol{\Sigma}_1=c_1\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ and $\boldsymbol{\Sigma}_2=c_2\boldsymbol{B}(0.4^{|i-j|^{1/3}})\boldsymbol{B}$, where $\boldsymbol{B}=\mathrm{diag}[\{0.5+1/(d+1)\}^{1/2},\ldots,\{0.5+d/(d+1)\}^{1/2}]$.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

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

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

First, we considered the linear SVM (LSVM) and the Gaussian kernel SVM (GSVM). We compared the performance of the bias-corrected LSVM (BC-LSVM) and bias-corrected GSVM (BC-GSVM) with the above ones. See (18) and (19) 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) $\boldsymbol{\mu}_1=\boldsymbol{0}$ and $(c_1,c_2)=(0.9,1.1)$, and
(c) $\boldsymbol{\mu}_1=\boldsymbol{\mu}_\alpha$ and $(c_1,c_2)=(0.9,1.1)$.



(a) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1) = \mathrm{tr}(\boldsymbol{\Sigma}_2)$ for $d = 2^s,\ s = 5, \ldots, 12$

(b) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ for $d = 2^s,\ s = 5, \ldots, 12$

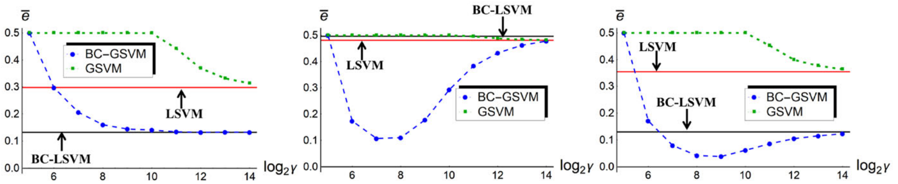
(c) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ for $d = 2^s,\ s = 5, \ldots, 12$

Fig. 3 The error rates of the BC-LSVM, LSVM, BC-GSVM and GSVM for (**a**–**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

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 Fig. 3, we plotted $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ for $d = 2^s,\ s = 5, \ldots, 12$.

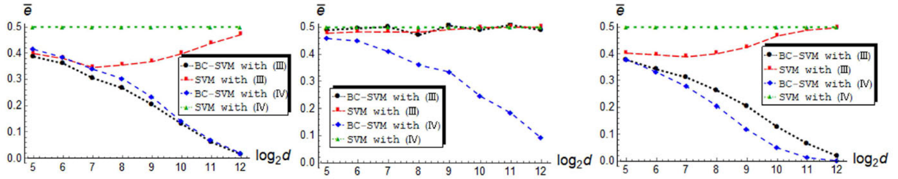
We observed that the BC-SVMs give good performances as d increases for (a) and (c). However, for (b), the error rate of the BC-LSVM is close to 0.5 because $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = 0$. On the other hand, the BC-GSVM gave good performances as d increases by drawing information about heteroscedasticity thorough the geometric representation as in Figs. 1 and 2 . For the LSVM and GSVM, $\overline{e}(1)$ and $\overline{e}(2)$ became quite unbalanced as d increases. In particular, the strong inconsistency (16) occurred for the GSVM. This is because of the bias in the GSVM. We give their theoretical backgrounds in Sect. 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 Fig. 3, we calculated the average error rate $\overline{e}$ by 2000 replications and plotted the results in Fig. 4. We observed that the BC-GSVM and GSVM are close to the BC-LSVM and LSVM, respectively, as γ increases for



(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)** $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$

Fig. 4 The average error rate, $\overline{e}$, of the BC-GSVM and GSVM for (**a**–**c**) when $d = 1024$ and $\gamma = 2^s$, $s = 5, \dots, 14$. The average error rates of the BC-LSVM and LSVM are described by the lines



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

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

(a) and (c). We give their theoretical backgrounds in Sect. 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 Sect. 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 $d = 2^s$, $s = 5, \dots, 12$. Similar to Fig. 3, we calculated the average error rate $\overline{e}$ by 2000 replications and plotted the results in Fig. 5. 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 $\boldsymbol{\Sigma}_i$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 Sect. 7.

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 z_{ij}, $j = 1, \ldots, n_i$, be i.i.d. random p_i-vectors having $E(z_{ij}) = \mathbf{0}$ and $\mathrm{Var}(z_{ij}) = \boldsymbol{I}_{p_i}$ for each i $(= 1, 2)$ and some p_i. Let $z_{ij} = (z_{i1j}, \ldots, z_{ip_i j})^{\mathrm{T}}$ 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 $\boldsymbol{x}_{ij} = \boldsymbol{\Gamma}_i z_{ij} + \boldsymbol{\mu}_i$, $j = 1, \ldots, n_i$, where $\boldsymbol{\Gamma}_i$ is a $d \times p_i$ matrix such that $\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_i^{\mathrm{T}} = \boldsymbol{\Sigma}_i$.

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 (2018a) for the details.

3.1 Linear kernel

We consider the linear SVM (LSVM); that is, the classifier (5) 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 $\kappa_5 = \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_2$, so that

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

We note that the LSVM is invariant to linear transformations on the data set. Thus, in Sect. 3.1, we assume $\boldsymbol{\mu}_2 = \mathbf{0}$ 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 under (D):

(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 (1), (C-ii) holds when $\liminf_{d\to\infty} \Delta_{(I)}/d > 0$ and n_is are fixed. We have the following result.

Lemma 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 3 with Theorem 1 and Corollary 1, we have the following results.

Corollary 2 *For the LSVM, one can claim that*

$$(6) \textit{ holds if } \limsup_{d\to\infty} \frac{|\delta_{(I)}|}{\Delta_{(I)}} < 1, \quad (15) \textit{ holds if } \liminf_{d\to\infty} \frac{\delta_{(I)}}{\Delta_{(I)}} > 1, \quad \textit{and}$$

$$(16) \textit{ holds if } \limsup_{d\to\infty} \frac{\delta_{(I)}}{\Delta_{(I)}} < -1$$

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

Nakayama et al. (2017) gave the results of Corollary 2 under slightly different conditions. They provided the following bias correction of the linear SVM: Let $\varDelta_{*(I)} = \varDelta_{(I)} + \eta_{1(I)}/n_1 + \eta_{2(I)}/n_2$. Estimate $\varDelta_{*(I)}$ and $\delta_{(I)}$ by

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

where $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})^{\mathrm{T}}/(n_i - 1)$. Note that $E(\hat{\varDelta}_{*(I)}) = \varDelta_{*(I)}$ 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, Nakayama et al. (2017) gave the bias-corrected linear SVM (BC-LSVM) as

$$\hat{y}_{BC(I)}(\boldsymbol{x}_0) = \hat{y}_{(I)}(\boldsymbol{x}_0) - \hat{\delta}_{(I)}/\hat{\varDelta}_{*(I)}. \tag{18}$$

One classifies $\boldsymbol{x}_0$ into $\varPi_1$ if $\hat{y}_{BC(I)}(\boldsymbol{x}_0) < 0$ and into $\varPi_2$ otherwise.

We note that $\hat{\varDelta}_{*(I)}$ and $\hat{\delta}_{(I)}$ are equivalent to $\hat{\varDelta}_*$ and $\hat{\delta}$ when $k(\cdot,\cdot)$ is the linear kernel. From Lemma 3 and Theorem 2, we have the following result.

Corollary 3 *Under (A-ii), (C-ii) and (D), the BC-LSVM holds the consistency* (6).

The BC-LSVM has the consistency property without (C-i). Chan and Hall (2009) considered a different bias correction for the LSVM. Nakayama et al. (2017) compared the BC-LSVM with the LSVM in both numerical simulations and actual 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).

3.2 Gaussian kernel

We consider the Gaussian kernel SVM (GSVM); that is, the classifier (5) 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 $\kappa_5 = \exp[-\{\mathrm{tr}(\boldsymbol{\Sigma}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2) + \varDelta_{(I)}\}/\gamma]$ $(= \kappa_{5(II)},\ \text{say})$, so that

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

We note that $\varDelta_{(II)} > 0$ 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 under (D):

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

We note that (C-iii) is a convergence condition of the GSVM. Under (1), (C-iii) holds when $\liminf_{d\to\infty} \varDelta_{(II)} > 0$, $\liminf_{d\to\infty} \gamma/d > 0$ and n_is are fixed. Note that $\psi \to 1$ and $\gamma\varDelta_{(II)} = 2\varDelta_{(I)}\{1 + o(1)\}$ as $d \to \infty$ under $d^2/(\gamma\varDelta_{(I)}) = o(1)$ as $d \to \infty$ from the fact that "$d^2/(\gamma\varDelta_{(I)}) = o(1)$" implies "$d/\gamma = o(1)$." Thus, (C-iii) holds under (C-ii) and $d^2/(\gamma\varDelta_{(I)}) = o(1)$. See Sect. 3.3 for the relation between the kernels (I) and (II). We have the following result.

Lemma 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 with Theorem 1 and Corollary 1, we have the following results.

Corollary 4 *For the GSVM, one can claim that*

$$(6)\ \textit{holds if}\ \limsup_{d\to\infty}\frac{|\delta_{(II)}|}{\Delta_{(II)}}<1,\quad (15)\ \textit{holds if}\ \liminf_{d\to\infty}\frac{\delta_{(II)}}{\Delta_{(II)}}>1,\quad \textit{and}$$

$$(16)\ \textit{holds if}\ \limsup_{d\to\infty}\frac{\delta_{(II)}}{\Delta_{(II)}}<-1$$

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

We denote $\hat{\eta}_i$ $(i=1,2)$ and $\hat{\Delta}_*$ for the kernel function (II) by $\hat{\eta}_{i(II)}$ and $\hat{\Delta}_{*(II)}$. Here, $\hat{\eta}_i$ and $\hat{\Delta}_*$ are defined in Sect. 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

$$\hat{y}_{BC(II)}(\boldsymbol{x}_0)=\hat{y}_{(II)}(\boldsymbol{x}_0)-\hat{\delta}_{(II)}/\hat{\Delta}_{*(II)}. \tag{19}$$

One classifies $\boldsymbol{x}_0$ into Π_1 if $\hat{y}_{BC(II)}(\boldsymbol{x}_0)<0$ and into Π_2 otherwise. From Theorem 2 and Lemma 4, we have the following result.

Corollary 5 *Under (A-ii), (C-iii) and (D), the BC-GSVM holds the consistency* (6).

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

Now, we consider the following condition:

$$\gamma/d\in(0,\infty)\ \text{ as } d\to\infty. \tag{20}$$

Let

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

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

$$\Delta_{(II)}/\psi=(1-\theta_2)^2+2\theta_2(1-\theta_1). \tag{21}$$

If one assumes that

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

it follows that $\liminf_{d\to\infty}\Delta_{(II)}>0$ under (20), so that (C-iii) holds as $d\to\infty$ while N is fixed under (1) and (20). Thus, the BC-GSVM has the consistency (6) 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 Fig. 3.

3.3 Relation between the linear kernel and Gaussian kernel

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

$$\text{(C-iv)}\ \ \frac{d^2}{\gamma \Delta_{(I)}} \to 0 \text{ as } d \to \infty, \text{ and }\ \ \text{(C-v)}\ \ \frac{\Delta_{(I)} + \Delta_{\Sigma}^2/\Delta_{(I)}}{\gamma} \to 0 \text{ as } d \to \infty.$$

Note that (C-iv) implies (C-v). By noting that $\psi \to 1$ as $d \to \infty$ under (C-iv), it holds from (21) that under (C-iv)

$$\gamma \Delta_{(II)} = 2\Delta_{(I)}\{1 + o(1)\}. \tag{22}$$

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

Proposition 3 *Under (A-ii), (C-ii), (C-iv) and (D), 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 (21) that under (C-v)

$$\gamma \Delta_{(II)} = 2\psi \Delta_{(I)}\{1 + o(1)\}. \tag{23}$$

Proposition 4 *Under (A-ii), (C-ii), (C-v) and (D), it holds that*

$$\left(\frac{\Delta_{(I)}}{\Delta_{*(I)}} \frac{\Delta_{*(II)}}{\Delta_{(II)}}\right) \hat{y}_{BC(II)}(\boldsymbol{x}_0) = \hat{y}_{BC(I)}(\boldsymbol{x}_0)\{1 + o_P(1)\}$$
$$\textit{when } \boldsymbol{x}_0 \in \Pi_i \textit{ for } i = 1, 2.$$

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

4 How to choose γ in the Gaussian kernel

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

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

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

$$\varDelta_{\Sigma}/\varDelta_{(I)} \to 0 \text{ as } d \to \infty, \quad \text{and} \tag{24}$$

$$\liminf_{d\to\infty} \varDelta_{\Sigma}/\varDelta_{(I)} > 0. \tag{25}$$

We first consider $\varDelta_{(II)}$ under (24). From (21), it holds that $\varDelta_{(II)}/\psi = 1 + \exp(-2\varDelta_{\Sigma}/\gamma) + o(1)$ under $\liminf_{d\to\infty} \varDelta_{\Sigma}/\gamma > 0$ and (24), so that the BC-GSVM (or GSVM) loses information about $\varDelta_{(I)}$. Thus, we do not consider the case when $\liminf_{d\to\infty} \varDelta_{\Sigma}/\gamma > 0$ under (24). Under (24) we consider the following conditions for γ, $\varDelta_{(I)}$ and $\varDelta_{\Sigma}$:

$$\varDelta_{\Sigma}/\gamma \to 0 \text{ as } d \to \infty, \quad \text{and} \tag{26}$$

$$\varDelta_{(I)}/\gamma \to 0 \text{ as } d \to \infty. \tag{27}$$

From (21), it holds that under (24) and (26)

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

On the other hand, it holds from (23) that under (24) and (27)

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

because (C-v) holds under (24) and (27). From Proposition 4, we note that the BC-LSVM is asymptotically equivalent to the BC-GSVM under (24) and (27). Also, note that $\gamma\{1 - \exp(-\varDelta_{(I)}/\gamma)\} \le \varDelta_{(I)}$ for any $\gamma > 0$. Then, from the convergence condition (C-iii), when (24) is met, we recommend to use the BC-LSVM or the BC-GSVM with γ satisfying (27).

Next, we consider $\varDelta_{(II)}$ under (25). From (21), it holds that under (25) and (26)

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

When (25) is met, the BC-GSVM (or GSVM) with γ satisfying (26) loses information about heteroscedasticity via the difference of $\mathrm{tr}(\boldsymbol{\Sigma}_i)$s. Thus, we do not consider the case when $\varDelta_{\Sigma}/\gamma = o(1)$ as $d \to \infty$ under (25). Under (25), we consider the following conditions for γ and $\varDelta_{\Sigma}$:

$$\varDelta_{\Sigma}/\gamma \to \infty \text{ as } d \to \infty, \quad \text{or} \tag{28}$$

$$\varDelta_{\Sigma}/\gamma \in (0, \infty) \text{ as } d \to \infty. \tag{29}$$

It holds that under (25) and (28)

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

Also, it holds that under (25) and (29)

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

Hence, from the convergence condition (C-iii), when (25) is met, we recommend to use the BC-GSVM with γ satisfying (29).

4.2 Choice of γ in the GSVM

In this section, we give a choice of γ in the GSVM. From Sect. 4.1, we recommend to use the BC-GSVM with γ satisfying

(i) the condition (27) when (24) is met, and
(ii) the condition (29) when (25) is met.

For the dual form (3), from Lemma 1, under (4) and several conditions, it holds that $\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{K}\hat{\boldsymbol{\alpha}} = \Delta\alpha_\star^2\{1+o_P(1)\} + \eta_1\sum_{j=1}^{n_1}\alpha_j^2 + \eta_2\sum_{j=n_1+1}^{N}\alpha_j^2$, so that

$$\frac{\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{K}\hat{\boldsymbol{\alpha}}}{\alpha_\star^2\Delta} - 1 - \frac{\eta_1\sum_{j=1}^{n_1}\alpha_j^2 + \eta_2\sum_{j=n_1+1}^{N}\alpha_j^2}{\alpha_\star^2\Delta} \quad (= \mathrm{Loss}(\gamma),\ \text{say}). \tag{30}$$

We emphasize that the accuracy of the BC-SVM (or SVM) heavily depends on the convergence rate of Loss(γ) because the bias in $\hat{y}(\boldsymbol{x}_0)$ converges to δ in Proposition 1. See Lemma 1 in Sect. 2. Thus, for the Gaussian kernel (II), we consider such γ as to have a higher convergence rate of Loss(γ). From Proposition 1 and (47) to (52) in Sect. 8, we can evaluate that under several conditions

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

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 (30), one may consider γ as

$$\gamma_0 = \operatorname*{argmin}_{\gamma>0} \frac{\kappa_{1(II)}+\kappa_{3(II)}+2\kappa_{5(II)}}{\gamma\Delta_{(II)}}. \tag{31}$$

When (24) is met, we have the following result.

Proposition 5 *Under* (24) *it holds that* $\Delta_{(I)}/\gamma_0 \to 0$ *as* $d \to \infty$.

Hence, when (24) is met, the BC-GSVM with γ_0 is asymptotically equivalent to the BC-LSVM because (C-v) is met under (24) and (27). See Proposition 4.

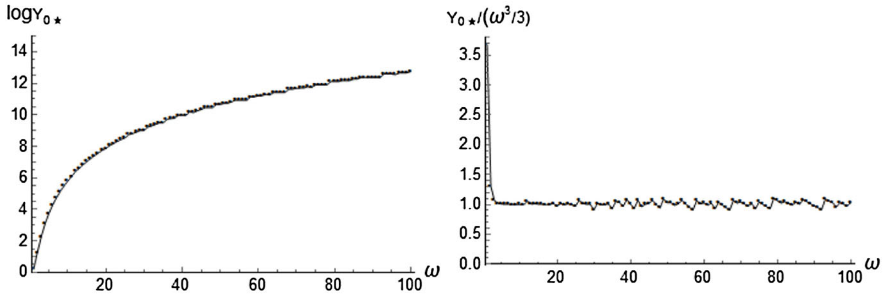


Fig. 6 The left panel displays $\log \gamma_{0\star}$, and the right panel displays $\gamma_{0\star}/(\omega^3/3)$ for $\omega = 1, \ldots, 100$

Next, we consider the case when

$$\limsup_{d\to\infty} \Delta_{(I)}/\Delta_\Sigma \le 1. \tag{32}$$

Proposition 6 *Under (32) it holds that $\Delta_\Sigma/\gamma_0 \in (0, \infty)$ as $d \to \infty$.*

Finally, we consider the case when

$$\liminf_{d\to\infty} \Delta_{(I)}/\Delta_\Sigma \ge 1 \ \text{ and } \ \limsup_{d\to\infty} \Delta_{(I)}/\Delta_\Sigma < \infty. \tag{33}$$

Since it is very difficult to evaluate γ_0 under (33), 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

$$\begin{aligned} \Delta_\Sigma \frac{\kappa_{1(II)} + \kappa_{3(II)} + 2\kappa_{5(II)}}{\gamma \Delta_{(II)}} &= \frac{\Delta_\Sigma}{\gamma}\left(1 + \frac{4\theta_1\theta_2}{1 + \theta_2^2 - 2\theta_1\theta_2}\right) \\ &= \frac{1}{\gamma_\star}\left(1 + \frac{4\exp\big\{-(\omega+1)/\gamma_\star\big\}}{1 + \exp(-2/\gamma_\star) - 2\exp\big\{-(\omega+1)/\gamma_\star\big\}}\right) \quad \big(= F(\gamma_\star),\ \text{say}\big). \end{aligned} \tag{34}$$

Thus, we consider the following minimization:

$$\gamma_{0\star} = \underset{\gamma_\star > 0}{\operatorname{argmin}}\, F(\gamma_\star).$$

Note that $\gamma_0 = \Delta_\Sigma \gamma_{0\star}$. Hence, (31) depends only on ω. We plotted $\gamma_{0\star}$ and $\gamma_{0\star}/(\omega^3/3)$ for $\omega = 1, \ldots, 100$ in Fig. 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 6 it holds that $\Delta_\Sigma/\gamma_0 = 1/\gamma_{0\star} \in (0, \infty)$ as $d \to \infty$ when (25) is met.

In conclusion, we recommend to use the BC-GSVM with γ_0. From (34) we estimate γ_0 as

$$\hat{\gamma}_0 = \underset{\gamma>0}{\operatorname{argmin}}\, \gamma^{-1}\left\{1+4\hat{\theta}_1\hat{\theta}_2/\left(1+\hat{\theta}_2^2-2\hat{\theta}_1\hat{\theta}_2\right)\right\}, \tag{35}$$

where $\hat{\theta}_1 = \exp(-\hat{\Delta}_{*(I)}/\gamma)$ and $\hat{\theta}_2 = \exp(-\hat{\Delta}_{\Sigma}/\gamma)$ with $\hat{\Delta}_{\Sigma} = |\mathrm{tr}(\boldsymbol{S}_{1n_1})-\mathrm{tr}(\boldsymbol{S}_{2n_2})|$. See Sect. 5 for the performance of the BC-SVM with $\hat{\gamma}_0$.

Remark 3 We note that $E(\hat{\Delta}_{(I)}) = \Delta_{(I)}$, where $\hat{\Delta}_{(I)} = \hat{\Delta}_{*(I)} - \mathrm{tr}(\boldsymbol{S}_{1n_1})/n_1 - \mathrm{tr}(\boldsymbol{S}_{2n_2})/n_2$. However, it does not hold $P(\hat{\Delta}_{(I)} \geq 0) = 1$. Thus, we use $\hat{\Delta}_{*(I)}$ in (35) since $P(\hat{\Delta}_{*(I)} \geq 0) = 1$.

Remark 4 Note that $E(\hat{\Delta}_{*(I)}) = \Delta_{*(I)}$, and $\mathrm{Var}(\hat{\Delta}_{(I)}) = O[\sum_{i=1}^{2}\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i^2 + \Delta_{(I)}\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}/n_i\}]$ 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 $\hat{\Delta}_{*(I)} = \Delta_{(I)}\{1+o_P(1)\}$ and $\hat{\Delta}_{\Sigma} = \Delta_{\Sigma}\{1+o_P(1)\}$ 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 (31).

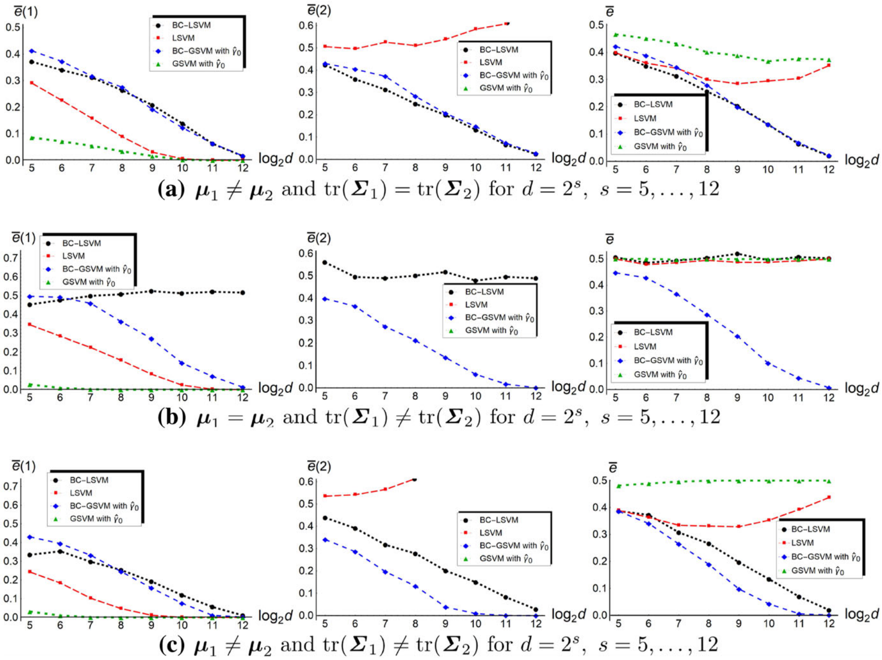
5 Performance of the BC-SVM

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

5.1 Simulations

For the settings (a) to (c) in Sect. 2.4, we first checked the performance of the BC-GSVM with $\hat{\gamma}_0$. Similar to Sect. 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 Fig. 7. We laid $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ for the BC-LSVM and the LSVM by borrowing them from Fig. 3. In the rth replication, we evaluated $\hat{\gamma}_{0r}$ by (35) and calculated $\overline{\gamma}_0 = \sum_{r=1}^{2000}\hat{\gamma}_{0r}/2000$. In Fig. 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 Sect. 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 Sect. 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 $\boldsymbol{\mu}_2 = (1,\ldots,1,0,\ldots,0,-1,\ldots,-1)^{\mathrm{T}}$ 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 $\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i$ $(= \boldsymbol{\Sigma}_i^{1/2}(z_{i1j},\ldots,z_{idj})^T)$, $j = 1, 2, \ldots$ $(i = 1, 2)$



(a) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1) = \mathrm{tr}(\boldsymbol{\Sigma}_2)$ for $d = 2^s,\ s = 5, \ldots, 12$

(b) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ for $d = 2^s,\ s = 5, \ldots, 12$

(c) $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ for $d = 2^s,\ s = 5, \ldots, 12$

Fig. 7 The error rates of the BC-LSVM, LSVM, BC-GSVM with $\gamma = \hat{\gamma}_0$ and GSVM with $\gamma = \hat{\gamma}_0$ for (**a**–**c**). The left panels display $\bar{e}(1)$, the middle panels display $\bar{e}(2)$, and the right panels display $\bar{e}$ for $d = 2^s,\ s = 5, \ldots, 12$. For the LSVM and GSVM, $\bar{e}(2)$ was too high to describe. Their standard deviations are less than 0.0112

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 $d = 2^s,\ s = 5, \ldots, 12$:

(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 Sect. 2.4, we calculated the error rates by 2000 replications and plotted the results in Fig. 9.

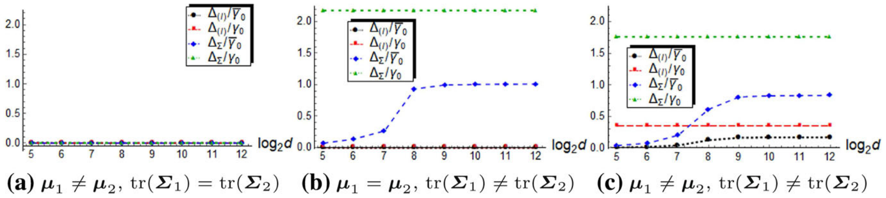
We observed that the BC-SVMs give adequate performances even when $n_i/n_{i'} \to 0$ for some i $(\neq i')$.

Throughout the simulations, $\hat{\gamma}_0$ by (35) was a preferable choice. We recommend to use a cross-validation procedure for γ around $\hat{\gamma}_0$. See Sect. 5.2.

5.2 Examples: microarray data sets

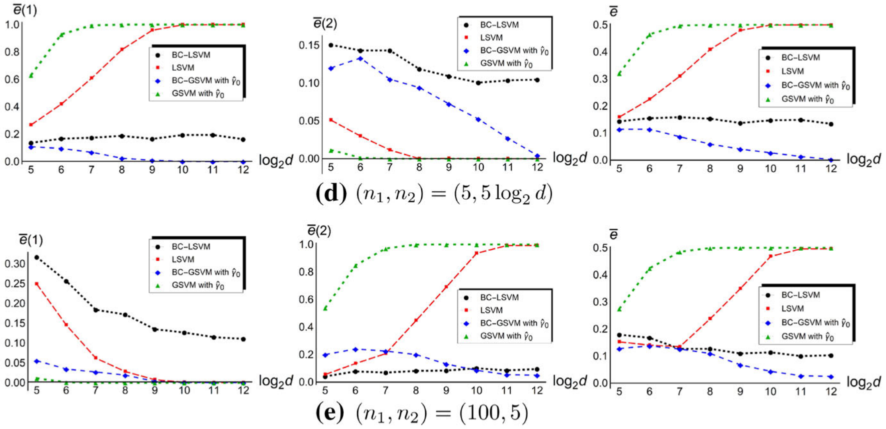
In this section, we analyze gene expression data sets by using the BC-SVMs and SVMs. We summarized the information on the data sets together with $\hat{\Delta}_{\Sigma}/\hat{\Delta}_{(I)}$ in Table 1, where $\hat{\Delta}_{(I)}$ and $\hat{\Delta}_{\Sigma}$ are given in Sect. 4.2.

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



(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)** $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$

Fig. 8 Behaviors of $\Delta_{(I)}/\overline{\gamma}_0$, $\Delta_{(I)}/\gamma_0$, $\Delta_{\Sigma}/\overline{\gamma}_0$ and Δ_{Σ}/γ_0 for (**a**–**c**)



(d) $(n_1, n_2) = (5, 5\log_2 d)$

(e) $(n_1, n_2) = (100, 5)$

Fig. 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

Table 1 Microarray data sets and $\hat{\Delta}_{\Sigma}/\hat{\Delta}_{(I)}$

Data set	Number of genes	Sample size		$\dfrac{\hat{\Delta}_{\Sigma}}{\hat{\Delta}_{(I)}}$
	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)	12,625	28	22	2.66
Breast cancer by Chang et al. (2003)	12,625	14	10	0.78

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 (35). We used the BC-GSVM and GSVM with $\hat{\gamma}_0$ (without applying the cross-validation) for Breast cancer because

Table 2 The average error rate $\overline{e}$ for five microarray data sets in Table 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

m_is are quite small for the data set. 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 2.

We observed that the BC-SVMs give adequate performances compared to the SVMs especially when n_1 and n_2 are unbalanced. See Sects. 3.1 and 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 $\hat{\Delta}_{\Sigma}/\hat{\Delta}_{(I)}$ 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.

6 Appendix A: soft-margin SVM

In Sects. 2–5, we discussed asymptotic properties and the performance of the hard-margin SVMs (hmSVM). In this section, we consider soft-margin SVMs (smSVM). The smSVM is given by $\hat{y}(\boldsymbol{x})$ after replacing (4) with

$$0 \le \alpha_j \le C, \ \ j = 1, \dots, N, \ \ \text{and} \ \ \sum_{j=1}^{N} \alpha_j t_j = 0, \tag{36}$$

where $C(> 0)$ is a regularization parameter. Let $n_{\min} = \min\{n_1, n_2\}$. From (11) in Sect. 2, we can asymptotically claim that $\hat{\alpha}_j \le 2/(\Delta_* n_{\min})$ for all j. Thus, we consider the following condition for C:

$$\liminf_{d \to \infty} \frac{C \Delta_* n_{\min}}{2} > 1. \tag{37}$$

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) with (36), respectively. Then, we have the following result.

Proposition 7 *Assume (A-i), (A-i') and* (8). *Under* (37), *it holds that when* $\boldsymbol{x}_0 \in \Pi_i$ *for* $i = 1, 2$

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

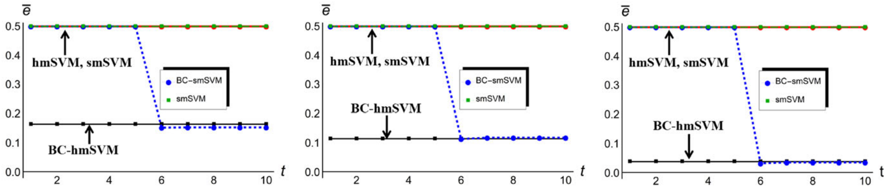
From Proposition 7, the bias-corrected smSVM (BC-smSVM) holds the consistency (6) even when $|\delta/\Delta| \to \infty$. Hence, for smSVMs, we recommend to use the BC-smSVM.

For the settings (a) to (c) in Sect. 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 $C = 2^{-5+t}/(n_{\min}\Delta_*),\ t = 1, \dots, 10$, for the smSVMs. Similar to Fig. 3, we calculated $\overline{e}$ by 2000 replications and plotted the results in Fig. 10. We observed that smSVMs give bad performances when $C < 2/(n_{\min}\Delta_*)$. As expected, the smSVMs are close to the hmSVMs when $C > 2/(n_{\min}\Delta_*)$.

7 Appendix B: Polynomial kernel SVM

In this section, we consider the polynomial kernel SVM; that is, the classifier (5) has the kernel function (III). We give some asymptotic properties of the polynomial kernel SVM. We consider the following conditions for ζ and r:

$$\zeta/d \in (0, \infty) \ \ \text{and} \ \ r \in (0, \infty) \ \ \text{as } d \to \infty. \tag{38}$$



(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)** $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$

Fig. 10 The average error rate, $\bar{e}$, of the BC-smSVM, smSVM, BC-hmSVM and hmSVM with (II) for (**a**–**c**) when $d = 1024$ and $C = 2^{-5+t}/(n_{\min}\Delta_*)$, $t = 1, \dots, 10$. 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

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 8 *Assume* (1), (38) *and (A-ii). Assume that N is fixed and*

$$\liminf_{d\to\infty} \left| \frac{\|\boldsymbol{\mu}_1\|^2 - \|\boldsymbol{\mu}_2\|^2}{d} \right| > 0. \tag{39}$$

Then, the assumptions (A-i) and (A-i') are met for the polynomial kernel (III). Furthermore, the BC-SVM (17) *with the polynomial kernel (III) holds the consistency* (6).

See Fig. 5 for the performance of the BC-SVM with the polynomial kernel (III).

Remark 5 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.

8 Appendix C: proofs

8.1 Proof of Lemma 1

Note that $L(\boldsymbol{\alpha}) = \sum_{j=1}^{N} \alpha_j - \acute{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{K} \acute{\boldsymbol{\alpha}}/2$. The result is obtained from (7) straightforwardly. □

8.2 Proofs of Propositions 1 and 2

We assume (A-i) and (A-i'). From Lemma 1, it holds that under (8) and (D)

$$\eta_1 \sum_{j=1}^{n_1} \hat{\alpha}_j^2/\hat{\alpha}_\star^2 = \eta_1/n_1 + o_P(\Delta) \quad \text{and} \quad \eta_2 \sum_{j=n_1+1}^{N} \hat{\alpha}_j^2/\hat{\alpha}_\star^2 = \eta_2/n_2 + o_P(\Delta), \tag{40}$$

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

$$\hat{\alpha}_\star = (2/\Delta_*)\{1 + o_P(\Delta/\Delta_*)\}. \tag{41}$$

Also, from (40) we have (9) under (8).

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

$$\begin{aligned}
&\sum_{j=1}^N \hat{\alpha}_j t_j k(\boldsymbol{x}_0, \boldsymbol{x}_j) + \frac{1}{N_{\hat{S}}} \sum_{j\in\hat{S}} \Big(t_j - \sum_{j'\in\hat{S}} \hat{\alpha}_{j'} t_{j'} k(\boldsymbol{x}_j, \boldsymbol{x}_{j'})\Big)\\
&= (-1)^i \hat{\alpha}_\star(\kappa_{2i-1} - \kappa_5) + \frac{\hat{n}_2 - \hat{n}_1}{N_{\hat{S}}}\\
&\quad - \hat{\alpha}_\star\Big(\frac{-\kappa_1\hat{n}_1 - \eta_1 + \kappa_3\hat{n}_2 + \eta_2 + (\hat{n}_1 - \hat{n}_2)\kappa_5}{N_{\hat{S}}}\Big) + o_P(\Delta\hat{\alpha}_\star)\\
&= (-1)^i \hat{\alpha}_\star(\kappa_{2i-1} - \kappa_5) + \frac{(\hat{n}_2 - \hat{n}_1)(1 - \hat{\alpha}_\star\Delta_*/2)}{N_{\hat{S}}} + \frac{\hat{\alpha}_\star(\kappa_1 - \kappa_3)}{2}\\
&\quad + \hat{\alpha}_\star\frac{\eta_1/n_1 - \eta_2/n_2}{2} + \hat{\alpha}_\star\frac{\eta_1(1 - \hat{n}_1/n_1) - \eta_2(1 - \hat{n}_2/n_2)}{N_{\hat{S}}} + o_P(\Delta\hat{\alpha}_\star).
\end{aligned} \tag{42}$$

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 (40) it holds that

$$\hat{n}_1(\eta_1/\hat{n}_1 - \eta_1/n_1) = \eta_1(1 - \hat{n}_1/n_1) = o_P(\hat{n}_1\Delta) \tag{43}$$

under (8). Similarly, we have $\eta_2(1 - \hat{n}_2/n_2) = o_P(\hat{n}_2\Delta)$ under (8). Then, from (41) and (42), we have that when $\boldsymbol{x}_0 \in \Pi_i$ for $i = 1, 2$,

$$\begin{aligned}
\hat{y}(\boldsymbol{x}_0) &= 2(-1)^i \frac{\kappa_{2i-1} - \kappa_5}{\Delta_*} + \frac{\kappa_1 - \kappa_3}{\Delta_*} + \frac{\eta_1/n_1 - \eta_2/n_2}{\Delta_*} + o_P\Big(\frac{\Delta}{\Delta_*}\Big)\\
&= (-1)^i \Delta/\Delta_* + \delta/\Delta_* + o_P(\Delta/\Delta_*)
\end{aligned} \tag{44}$$

under (8). Hence, we conclude the second result of Proposition 1.

Finally, we consider the proof of Proposition 2. In view of (13), we claim the first result. By noting that $\Delta_*/\Delta \to 1$ and $\delta/\Delta = o(1)$ under (12) and (D), it holds from (42) that $\hat{y}(\boldsymbol{x}_0) = (-1)^i + o_P(1)$ under (12) and (D). We conclude the second result. □

8.3 Proofs of Theorem 1 and Corollary 1

We assume (A-i) and (A-i'). We consider the following conditions under (D):

$$\liminf_{d\to\infty} \eta_2/(n_2\Delta) > 0 \ \text{ and } \ \eta_1/(n_1\Delta) = o(1). \tag{45}$$

Let $\Delta_{*2} = \Delta + \eta_2/n_2$. Note that $\eta_1 \sum_{j=1}^{n_1} \hat{\alpha}_j^2/\hat{\alpha}_\star^2 = o_P(\Delta)$ under (45). Similar to (41), it holds from (42) and (43) that $\hat{\alpha}_\star = (2/\Delta_{*2})\{1 + o_P(\Delta/\Delta_{*2})\}$ and

$$\hat{y}(\boldsymbol{x}_0) = (-1)^i \Delta/\Delta_{*2} + \delta/\Delta_{*2} + o_P(\Delta/\Delta_{*2}) \tag{46}$$

under (45) when $\boldsymbol{x}_0 \in \Pi_i$ for $i = 1, 2$. Note that $\Delta_*/\Delta \to 1$ and $\delta/\Delta_* \to 0$ under (12) and $\Delta_*/\Delta_{*2} \to 1$ under (45). From Propositions 1, 2 and (46), we obtain (44) under (D) without (8). Thus, from (44), we conclude the results of Theorem 1 and Corollary 1. □

8.4 Proofs of Lemma 2 and Theorem 2

Under (A-i) and (D), it holds that $\hat{\Delta}_* = \Delta_* + o_P(\Delta)$ and $\hat{\eta}_i = \eta_i + o_P(\Delta)$ for $i = 1, 2$. Thus, we can conclude the result of Lemma 2. From the proofs of Theorem 1 and Corollary 1, we obtain (44) under (A-i) and (D). By combining (44) with Lemma 2, we conclude the result of Theorem 2. □

8.5 Proofs of Lemma 3, Corollaries 2 and 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

$$\boldsymbol{\mu}_1^{\mathrm{T}} \boldsymbol{\Sigma}_i \boldsymbol{\mu}_1 \le \Delta_{(I)} \lambda_{\max}(\boldsymbol{\Sigma}_i) \le \Delta_{(I)} \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2}. \tag{47}$$

Then, by using Chebyshev's inequality, for any $\tau > 0$ we have that

$$\begin{aligned} &\sum_{j=1}^{n_1} P\left(|\boldsymbol{\mu}_1^{\mathrm{T}}(\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1)| \ge \tau\Delta_{(I)}\right) \le n_1 \left(\tau\Delta_{(I)}\right)^{-4} E\left[\left\{\boldsymbol{\mu}_1^{\mathrm{T}}(\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1)\right\}^4\right] \\ &\quad = O\left\{n_1 \left(\left(\boldsymbol{\mu}_1^{\mathrm{T}} \boldsymbol{\Sigma}_i \boldsymbol{\mu}_1\right)^2 + \sum_{r=1}^{p_1} \left(\boldsymbol{\gamma}_r^{\mathrm{T}} \boldsymbol{\mu}_1\right)^4\right) / \Delta_{(I)}^4\right\} = O\left(n_1 \mathrm{tr}\left(\boldsymbol{\Sigma}_i^2]\right) / \Delta_{(I)}^2\right) \to 0 \end{aligned} \tag{48}$$

from the fact that $\sum_{r=1}^{p_1} (\boldsymbol{\gamma}_r^{\mathrm{T}} \boldsymbol{\mu}_1)^4 \le (\boldsymbol{\mu}_1^{\mathrm{T}} \boldsymbol{\Sigma}_i \boldsymbol{\mu}_1)^2$, where $\boldsymbol{\Gamma}_1 = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{p_1}]$. On the other hand, we have that

$$\begin{aligned} &\sum_{j<j'}^{n_i} P(|(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)^{\mathrm{T}}(\boldsymbol{x}_{ij'} - \boldsymbol{\mu}_i)| \ge \tau\Delta_{(I)}) \\ &\quad \le \sum_{j<j'}^{n_i} (\tau\Delta_{(I)})^{-4} E\left[\{(\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1)^{\mathrm{T}}(\boldsymbol{x}_{1j'} - \boldsymbol{\mu}_1)\}^4\right] = O\left(n_i^2 \mathrm{tr}\left(\boldsymbol{\Sigma}_i^2\right)^2 / \Delta_{(I)}^4\right) \to 0. \end{aligned} \tag{49}$$

Note that $\boldsymbol{x}_{1j}^{\mathrm{T}}\boldsymbol{x}_{1j'}-\kappa_1=(\boldsymbol{x}_{1j}-\boldsymbol{\mu}_1)^{\mathrm{T}}(\boldsymbol{x}_{1j'}-\boldsymbol{\mu}_1)+\boldsymbol{\mu}_1^{\mathrm{T}}(\boldsymbol{x}_{1j}-\boldsymbol{\mu}_1+\boldsymbol{x}_{1j'}-\boldsymbol{\mu}_1)$. Thus, from (48) and (49), it holds that

$$\boldsymbol{x}_{1j}^{\mathrm{T}}\boldsymbol{x}_{1j'}=\kappa_1+o_P(\Delta_{(I)}) \text{ for all } j<j'\le n_1. \tag{50}$$

Note that

$$\sum_{j=1}^{n_1}\sum_{j'=1}^{n_2}P(|(\boldsymbol{x}_{1j}-\boldsymbol{\mu}_1)^{\mathrm{T}}(\boldsymbol{x}_{2j'}-\boldsymbol{\mu}_2)|\ge\tau\Delta_{(I)})$$
$$=O\Big(n_1n_2\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}^2+\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}/\Delta_{(I)}^4\Big)\to 0 \tag{51}$$

from the fact that $\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\le\{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\}^2$. Then, similar to (50), we have that

$$\begin{aligned}
&\boldsymbol{x}_{2j}^{\mathrm{T}}\boldsymbol{x}_{2j'}=\kappa_3+o_P(\Delta_{(I)}) \text{ for all } j<j'\le n_2,\\
&\boldsymbol{x}_{1j}^{\mathrm{T}}\boldsymbol{x}_{2j'}=\kappa_5+o_P(\Delta_{(I)}) \text{ for all } j=1,\ldots,n_1;\ \ j'=1,\ldots,n_2,\\
&\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{x}_{ij}=\kappa_{2i-1}+o_P(\Delta_{(I)}) \text{ for all } 1\le j\le n_i, i=1,2, \text{ when } \boldsymbol{x}_0\in\Pi_i\\
&\text{and } \boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{x}_{i'j}=\kappa_5+o_P(\Delta_{(I)}) \text{ for all } 1\le j\le n_i, i=1,2\ (i'\ne i) \text{ when } \boldsymbol{x}_0\in\Pi_i.
\end{aligned}$$

In addition, for any $\tau>0$ we have that

$$\sum_{j=1}^{n_i}P\left(\left|\|\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i\|^2-\mathrm{tr}(\boldsymbol{\Sigma}_i)\right|\ge\tau\Delta_{(I)}\right)=O\left(n_i\mathrm{tr}\left(\boldsymbol{\Sigma}_i^2\right)/\Delta_{(I)}^2\right)\to 0 \tag{52}$$

for $i=1,2$. Thus, from (48) and (52), it holds that for all $j=1,\ldots,n_i;\ i=1,2$

$$\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{x}_{ij}=\kappa_{2i}+o_P(\Delta_{(I)}).$$

It concludes Lemma 3.

For the proofs of Corollaries 2 and 3 , from Theorems 1, 2 and Corollary 1, we conclude the results. □

8.6 Proofs of Lemma 4, Corollaries 4 and 5

We assume (A-ii). Let $\Omega=\min\{\gamma\Delta_{(II)}/\psi,\ \gamma\}$. Similar to (48), for any $\tau>0$, we have that under (C-iii) and (D)

$$\sum_{j=1}^{n_i}P(|(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i)|\ge\tau\Omega)\to 0$$

for $i=1,2$, so that $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i)=o_P(\Omega)$ 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

$(\boldsymbol{x}_{1j}-\boldsymbol{\mu}_1)^{\mathrm{T}}(\boldsymbol{x}_{2j'}-\boldsymbol{\mu}_2)=o_P(\Omega)$ for all $j=1,\dots,n_1;\ \ j'=1,\dots,n_2$. Then, under (C-iii), we have that for all $j=1,\dots,n_1;\ \ j'=1,\dots,n_2$

$$\begin{aligned}\exp(-\|\boldsymbol{x}_{1j}-\boldsymbol{x}_{2j'}\|^2/\gamma)&=\exp(-\|(\boldsymbol{x}_{1j}-\boldsymbol{\mu}_1)-(\boldsymbol{x}_{2j'}-\boldsymbol{\mu}_2)+\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2/\gamma)\\&=\kappa_{5(II)}+o_P(\kappa_{5(II)}\Omega/\gamma)=\kappa_{5(II)}+o_P(\Delta_{(II)})\qquad(53)\end{aligned}$$

from the fact that $\kappa_{5(II)}\le\psi$. Similar to (53), we can conclude that the assumptions (A-i) and (A-i') are met. It concludes Lemma 4.

For the proofs of Corollaries 4 and 5 , from Theorems 1, 2 and Corollary 1, we conclude the results. □

8.7 Proofs of Propositions 3 and 4

From (23), (C-iii) holds under (C-ii) and (C-v). Thus, from (44) and Lemmas 2 to 4, we conclude Proposition 4. For the proof of Proposition 3, we note that $\mathrm{tr}(\boldsymbol{\Sigma}_i)/\gamma\to0$ for $i=1,2$, under (C-iv) from the fact that $\Delta_{(I)}=O(d)$. Thus, it holds that $\psi\to1$ 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 (22) it holds that $\delta_{(II)}/\Delta_{(II)}=\delta_{(I)}\{1+o(1)\}/\Delta_{(I)}+o(1)$ under (C-iv). Thus, from (44), Lemmas 3 and 4 , we conclude Proposition 3. □

8.8 Proof of Proposition 5

We assume (24). Note that $1/\omega\to0$ under (24). 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 $g(\nu)\to2$ as $\nu\to0$, so that $F(\gamma_\star)=2\{1+o(1)\}/\omega=o(1)$ when $\nu\to0$. We can conclude the result. □

8.9 Proof of Proposition 6

When $\omega\le1$, it holds that $F(\gamma_\star)=2\{1+o(1)\}/\omega$ under $\gamma_\star\to\infty$. When $\omega\le1$ and $\gamma_\star=1$, it holds that

$$F(\gamma_\star)=1+\frac{4}{\exp(\omega+1)+\exp(\omega-1)-2}<1+1/\omega\le2/\omega$$

from the facts that $\exp(\omega+1)>1+(\omega+1)+(\omega+1)^2/2\ge2+3\omega$ and $\exp(\omega-1)\ge\omega$. Hence, when $\omega\le1$, we have that $\Delta_{\Sigma}/\gamma_0\in(0,\infty)$ as $d\to\infty$. It concludes the result. □

8.10 Proof of Proposition 7

From Proposition 1, Lemma 2 and (11), we can conclude the results. □

8.11 Proof of Proposition 8

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 (1), we note that $\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_{i'}\boldsymbol{\mu}_i \le \|\boldsymbol{\mu}_i\|^2\lambda_{\max}(\boldsymbol{\Sigma}_i) = o(d^2)$ as $d \to \infty$ for $i, i' = 1, 2$. Then, similar to (50)–(52), for the polynomial kernel, we have that $\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{x}_{ij'} = \|\boldsymbol{\mu}_i\|^2 + o_P(d)$ for all $j < j'$, $i = 1, 2$, $\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{x}_{ij} = \mathrm{tr}(\boldsymbol{\Sigma}_i) + \|\boldsymbol{\mu}_i\|^2 + o_P(d)$ for all i, j, and $\boldsymbol{x}_{1j}^{\mathrm{T}}\boldsymbol{x}_{2j'} = \boldsymbol{\mu}_1^T\boldsymbol{\mu}_2 + o_P(d)$ 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

$$\begin{aligned}&(\zeta + \|\boldsymbol{\mu}_1\|^2)^r + (\zeta + \|\boldsymbol{\mu}_2\|^2)^r - 2(\zeta + \boldsymbol{\mu}_1^T\boldsymbol{\mu}_2)^r\\ &\quad \ge \{(\zeta + \|\boldsymbol{\mu}_1\|^2)^{r/2} - (\zeta + \|\boldsymbol{\mu}_2\|^2)^{r/2}\}^2\end{aligned}$$

from the fact that $(\zeta + \boldsymbol{\mu}_1^T\boldsymbol{\mu}_2)^r \le (\zeta + \|\boldsymbol{\mu}_1\|^2)^{r/2}(\zeta + \|\boldsymbol{\mu}_2\|^2)^{r/2}$. Then, it holds that $\liminf_{d\to\infty} \Delta/d^r > 0$ from (39). Thus, we have (A-i). Similarly, we can conclude (A-i'). From Theorem 2, the BC-SVM (17) holds (6) for the polynomial kernel. It concludes Proposition 8. □

References

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

Ahn, J., Marron, J. S., Muller, K. M., Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, *94*, 760–766.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., 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 of the United States of America*, *96*, 6745–6750.

Aoshima, M., Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Analysis*, *30*, 356–399. (Editor's special invited paper).

Aoshima, M., Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, *66*, 983–1010.

Aoshima, M., Yata, K. (2015). Geometric classifier for multiclass, high-dimensional data. *Sequential Analysis*, *34*, 279–294.

Aoshima, M., Yata, K. (2018a). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica*, *28*, 43–62.

Aoshima, M., Yata, K. (2018b). High-dimensional quadratic classifiers in non-sparse settings. Methodology and Computing in Applied Probability. https://doi.org/10.1007/s11009-018-9646-z.

Aoshima, M., Yata, K. (2019). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Annals of the Institute of Statistical Mathematics*, *71*, 473–503.

Bai, Z., Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, *6*, 311–329.

Benjamin, X. W., Nathalie, J. (2010). Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, *25*, 1–20.

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., O'Connell, P. (2003). Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *The Lancet*, *362*, 362–369.

Chan, Y.-B., Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, *96*, 469–478.

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., Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*, 531–537.

Hall, P., Marron, J. S., Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society, Series B*, *67*, 427–444.

Hall, P., Pittelkow, Y., Ghosh, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *Journal of the Royal Statistical Society, Series B*, *70*, 159–173.

He, H., Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284.

Huang, H. (2017). Asymptotic behavior of support vector machine for spiked population model. *Journal of Machine Learning Research*, *18*, 1–21.

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

Nakayama, Y., Yata, K., 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.

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., Louis, D. N. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, *63*, 1602–1607.

Qiao, X., Zhang, L. (2015). Flexible high-dimensional classification machines and their asymptotic properties. *Journal of Machine Learning Research*, *16*, 1547–1572.

Qiao, X., Zhang, H. H., Liu, Y., Todd, M. J., Marron, J. S. (2010). Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statistical Association*, *105*, 401–414.

Schölkopf, B., Smola, A. J. (2002). *Learning with Kernels*. Cambridge: MIT Press.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, *8*, 68–74.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*2nd ed. New York: Springer.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.