数理解析研究所講究録 第2124巻 2019年 44-55

Soft-margin SVMs in the HDLSS context

Yugo Nakayama
Graduate School of Pure and Applied Sciences
University of Tsukuba

Kazuyoshi Yata
Institute of Mathematics
University of Tsukuba

Makoto Aoshima
Institute of Mathematics
University of Tsukuba

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 $\boldsymbol{\Sigma}_i$ for each 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 $n_i\geq 2,\ i=1,2$. Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to one of the two populations. Let $N=n_1+n_2$. We assume $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent.

In this paper, we consider classification in the High-dimension, low-sample-size (HDLSS) context such as $d\to\infty$ while N is fixed. Hall et al. [7], Chan and Hall [5] and Aoshima and Yata [2] considered distance-based classifiers. In particular, Aoshima and Yata [2] gave the misclassification rate adjusted classifier for multiclass, high-dimensional data in which misclassification rates are no more than specified thresholds. On the other hand, Aoshima and Yata [1, 3] considered geometric classifiers based on a geometric representation of HDLSS data. Aoshima and Yata [4] considered quadratic classifiers in general and discussed asymptotic properties and optimality of the classifiers under high-dimension, non-sparse settings. For linear support vector machine (SVM) in HDLSS settings, Hall et al. [6], Chan and Hall [5] and Qiao and Zhang [11] showed that the misclassification rates tend to zero as $d\to\infty$ under certain severe conditions. Nakayama et al. [8] investigated asymptotic properties of linear SVM for HDLSS data. They proposed a bias-corrected linear SVM and showed that it gives preferable performances compared to linear SVM. Nakayama [9] investigated asymptotic properties of a soft-margin linear SVM. On the other hand, Nakayama et al. [10] investigated asymptotic properties of SVM with the Gaussian kernel for HDLSS data.

In this paper, we consider the soft-margin SVM as follows:

$$y(\boldsymbol{x}) = \boldsymbol{w}^T\phi(\boldsymbol{x}) + b, \tag{1}$$

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:

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

where $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \phi(\boldsymbol{x}_j)^T\phi(\boldsymbol{x}_{j'})$ is a kernel function, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^T$ 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

$$0 \le \alpha_j \le C,\ j = 1, \ldots, N, \quad \text{and} \quad \sum_{j=1}^{N} \alpha_j t_j = 0, \tag{2}$$

where $C(> 0)$ is a regularization parameter. Let us write that

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N)^T = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L(\boldsymbol{\alpha}) \ \text{ subject to (2)}.$$

There exist some $\boldsymbol{x}_j$s satisfying that $t_jy(\boldsymbol{x}_j) = 1$ (i.e., $\hat{\alpha}_j \ne 0$). Such $\boldsymbol{x}_j$s are called the support vector. Let $\hat{S} = \{j|\hat{\alpha}_j \ne 0,\ j = 1, \ldots, N\}$ 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 (1) is defined by

$$\hat{y}(\boldsymbol{x}) = \sum_{j\in\hat{S}} \hat{\alpha}_j t_j k(\boldsymbol{x}, \boldsymbol{x}_j) + \hat{b}. \tag{3}$$

Finally, in SVM, one classifies $\boldsymbol{x}_0$ into Π_1 if $\hat{y}(\boldsymbol{x}_0) < 0$ and into Π_2 otherwise. See Vapnik [12] 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 consistency if

$$e(i) = o(1) \ \text{ as } d \to \infty \text{ for } i = 1, 2. \tag{4}$$

In this paper, we investigate the following typical kernels for the soft-margin SVM:

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

where $\gamma(>0)$ is a scale parameter and $\zeta \geq 0$ and $r \in \mathbb{N}$.

In Section 2, we investigate asymptotic properties of the soft-margin SVM with the Gaussian kernel. In Section 3, we investigate asymptotic properties of the soft-margin SVM with the polynomial kernel. We show that the SVMs are heavily biased in the HDLSS context especially for imbalanced data. In order to overcome such difficulties, we propose a bias-corrected SVM in Section 4. In Section 5, we check the performance of the BC-SVM by numerical simulations.

2 Asymptotic properties of the soft-margin SVM with the Gaussian kernel

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 Aoshima and Yata [2], 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 $\boldsymbol{z}_{ij} = (z_{i1j}, \ldots, z_{ip_ij})^\top$ 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 $\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^\top = \boldsymbol{\Sigma}_i$.

Note that (A-i) naturally holds when the Π_is are Gaussian.

We consider the soft-margin Gaussian kernel SVM (sm-GSVM), that is, the classifier (3) with the Gaussian kernel. Let $\Delta_\mu = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$. Let $\kappa_{1(I)} = \exp\{-2\mathrm{tr}(\boldsymbol{\Sigma}_1)/\gamma\}$, $\kappa_{2(I)} = \exp\{-2\mathrm{tr}(\boldsymbol{\Sigma}_2)/\gamma\}$, $\kappa_{3(I)} = \exp[-\{\mathrm{tr}(\boldsymbol{\Sigma}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2) + \Delta_\mu\}/\gamma]$, and

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

We note that $\Delta_{(I)} > 0$ when $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ or $\mathrm{tr}(\boldsymbol{\Sigma}_1) \neq \mathrm{tr}(\boldsymbol{\Sigma}_2)$. We consider the following condition:

$$\liminf_{d\to\infty} \frac{\eta_{i(I)}}{\Delta_{(I)}} > 0 \quad \text{for } i = 1, 2. \tag{5}$$

Let $\Delta_{*(I)} = \Delta_{(I)} + \eta_{1(I)}/n_1 + \eta_{2(I)}/n_2$ and $n_{\min} = \min\{n_1, n_2\}$. We consider the following condition for C:

$$\liminf_{d\to\infty} \frac{C\Delta_{*(I)} n_{\min}}{2} > 1. \tag{6}$$

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 as $d \to \infty$:

(A-ii) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)+\Delta_\mu\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}^{1/2}}{\min\{\gamma^2\Delta_{(I)}^2/\psi^2,\ \gamma^2\}}=o(1)$ for $i=1,2$.

Let $\delta_{(I)}=\eta_{1(I)}/n_1-\eta_{2(I)}/n_2$. Let $\hat{y}_{(I)}(\boldsymbol{x}_0)$ denote $\hat{y}(\boldsymbol{x}_0)$ given by using the kernel function (I). Then, from Sections 2 and 6 in Nakayama et al. [10], we have the following results.

Theorem 1. *Assume (A-i) and (A-ii). Assume also (5) and (6). Then, it holds that as* $d\to\infty$

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

Assume also

(A-iii) $\limsup_{d\to\infty}\dfrac{|\delta_{(I)}|}{\Delta_{(I)}}<1.$

Then, the sm-GSVM holds consistency (4).

Corollary 1. *For the sm-GSVM, one can claim that*

$$\begin{aligned}&e(1)=1+o(1)\quad \textit{and}\quad e(2)=o(1)\quad \textit{as } d\to\infty\\ &\textit{if}\quad \liminf_{d\to\infty}\frac{\delta_{(I)}}{\Delta_{(I)}}>1;\qquad \textit{and}\\ &e(1)=o(1)\quad \textit{and}\quad e(2)=1+o(1)\quad \textit{as } d\to\infty\\ &\textit{if}\quad \limsup_{d\to\infty}\frac{\delta_{(I)}}{\Delta_{(I)}}<-1.\end{aligned}$$

under (A-i), (A-ii) and (5) and (6).

From Corollary 1, if $|\delta_{(I)}|$ is larger than $\Delta_{(I)}$, the sm-GSVM would give a bad performance. In order to overcome such difficulties, we propose a bias-corrected SVM in Section 4.

3 Asymptotic properties of the soft-margin SVM with the polynomial kernel

In this section, we consider the soft-margin polynomial kernel SVM (sm-PSVM), that is, the classifier (3) with the polynomial kernel.

Let $\kappa_{1(II)}=(\zeta+\|\boldsymbol{\mu}_1\|^2)^r$, $\kappa_{2(II)}=(\zeta+\|\boldsymbol{\mu}_2\|^2)^r$, $\kappa_{3(II)}=(\zeta+\boldsymbol{\mu}_1^T\boldsymbol{\mu}_2)^r$, and

$$\begin{aligned}\Delta_{(II)}=&\kappa_{1(II)}+\kappa_{2(II)}-2\kappa_{3(II)}\quad\text{and}\\ \eta_{i(II)}=&(\zeta+\mathrm{tr}(\boldsymbol{\Sigma}_i)+\|\boldsymbol{\mu}_i\|^2)^r-\kappa_{i(II)}\quad\text{for } i=1,2.\end{aligned}$$

We consider the following condition:

$$\liminf_{d\to\infty} \frac{\eta_{i(II)}}{\Delta_{(II)}} > 0 \quad \text{for } i = 1, 2. \tag{7}$$

Let $\Delta_{*(II)} = \Delta_{(II)} + \eta_{1(II)}/n_1 + \eta_{2(II)}/n_2$. We consider the following condition for C:

$$\liminf_{d\to\infty} \frac{C\Delta_{*(II)} n_{\min}}{2} > 1. \tag{8}$$

We assume the following conditions for ζ and r:

$$\zeta/d \in (0, \infty) \quad \text{and} \quad r \in (0, \infty) \quad \text{as } d \to \infty. \tag{9}$$

We also assume the following condition:

(A-iv) $\liminf_{d\to\infty} \left| \frac{\|\boldsymbol{\mu}_1\|^2 - \|\boldsymbol{\mu}_2\|^2}{d} \right| > 0.$

Let $\delta_{(II)} = \eta_{1(II)}/n_1 - \eta_{2(II)}/n_2$. Let $\hat{y}_{(II)}(\boldsymbol{x}_0)$ denote $\hat{y}(\boldsymbol{x}_0)$ given by using the kernel function (II). Then, from Sections 2 and 7 in Nakayama et al. [10], we have the following results.

Theorem 2. *Assume (A-i) and (A-iv). Assume also (7) to (9). Then, it holds that as* $d \to \infty$

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

Assume also

(A-v) $\limsup_{d\to\infty} \frac{|\delta_{(II)}|}{\Delta_{(II)}} < 1.$

Then, the sm-PSVM holds consistency (4).

Corollary 2. *For the sm-PSVM, one can claim that*

$$\begin{aligned}
&e(1) = 1 + o(1) \quad \textit{and} \quad e(2) = o(1) \quad \textit{as } d \to \infty \\
&\textit{if} \quad \liminf_{d\to\infty} \frac{\delta_{(II)}}{\Delta_{(II)}} > 1; \quad \textit{and} \\
&e(1) = o(1) \quad \textit{and} \quad e(2) = 1 + o(1) \quad \textit{as } d \to \infty \\
&\textit{if} \quad \limsup_{d\to\infty} \frac{\delta_{(II)}}{\Delta_{(II)}} < -1.
\end{aligned}$$

under (A-i), (A-iv) and (7) to (9).

Similar to the sm-GSVM, if $|\delta_{(II)}|$ is larger than $\Delta_{(II)}$, the sm-PSVM would give a bad performance.

4 Bias-corrected SVM

Let

$$\hat{\eta}_i = \sum_{j=1}^{n_i} \frac{k(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij})}{n_i - 1} - \sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \frac{k(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'})}{n_i(n_i - 1)} \quad \text{for } i = 1, 2; \text{ and} \tag{10}$$

$$\hat{\Delta}_* = \sum_{i=1}^{2} \left(\sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \frac{k(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'})}{n_i^2} \right) - 2 \sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} \frac{k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j'})}{n_1 n_2}. \tag{11}$$

We consider estimating δ as $\hat{\delta} = \hat{\eta}_1/n_1 - \hat{\eta}_2/n_2$. We give a bias-corrected SVM (BC-SVM) as follows:

$$\hat{y}_{BC}(\boldsymbol{x}_0) = \hat{y}(\boldsymbol{x}_0) - \frac{\hat{\delta}}{\hat{\Delta}_*}. \tag{12}$$

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 3. *Assume (A-i) and (A-ii). Assume also (5) and (6). For the classifier (12) with the Gaussian kernel, it holds the consistency (4).*

For the Gaussian kernel, the BC-SVM claims the consistency without (A-iii).

Theorem 4. *Assume (A-i) and (A-iv). Assume also (7) to (9). For the classifier (12) with the polynomial kernel, it holds the consistency (4).*

For the polynomial kernel, the BC-SVM claims the consistency without (A-v).

Remark 1. *Nakayama et al. [8] gave a bias-corrected linear SVM. Nakayama [9] also proposed a robust SVM in HDLSS settings for the linear kernel.*

5 Simulation

In this section, we compared the performance of the sm-GSVM, sm-PSVM and BC-SVM with the kernel functions (I) and (II). 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 $\boldsymbol{B} = \text{diag}[\{0.5+1/(d+1)\}^{1/2}, \ldots, \{0.5+d/(d+1)\}^{1/2}]$. Note that $\text{tr}(\boldsymbol{\Sigma}_i) = c_i d$ 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 $r = 1, \ldots, d$. We set $(n_1, n_2) = (20, 10)$, $\gamma = d/4$ in the Gaussian kernel and $\zeta = d$, $r = 2$ in the polynomial kernel. 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 $(c_1, c_2) = (0.9, 1.1)$.

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 set $C = 4/(n_{\min}\hat{\Delta}_*)$ for both kernel (I) and (II). From Lemma 2 in Nakayama et al. [10], it holds that $\hat{\Delta}_* = \Delta_*\{1 + o_P(1)\}$, so that (6) and (8) hold. 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 $\mathrm{Var}\{\overline{e}(i)\} = e(i)\{1 - e(i)\}/2000 \le 1/8000$. In Figures 1 to 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-SVM with the polynomial kernel is 0.5 because (A-iv) does not hold. On the other hand, the BC-SVM with the Gaussian kernel gave good performances drawing information about heteroscedasticity. For the sm-GSVM and the sm-PSVM, $\overline{e}(1)$ and $\overline{e}(2)$ became quite unbalanced. This is because of the bias in the SVM. See Corollaries 1 and 2 for the details.

Next, we considered (a) to (c) for $(n_1, n_2) = (20, 10)$, $d = 1024$ $(= 2^{10})$ and $C = 2^{-7+t}/(n_{\min}\Delta_*)$, $t = 1, \ldots, 10$ for the kernel function (I) and (II). Similar to Figures 1 to 3, we calculated the average error rate $\overline{e}$ by 2000 replications and plotted the results in Figure 4. We observed that the sm-GSVM and the sm-PSVM give bad performances for all C. However the BC-SVMs gave good performances when $C > 2/(n_{\min}\Delta_*)$.

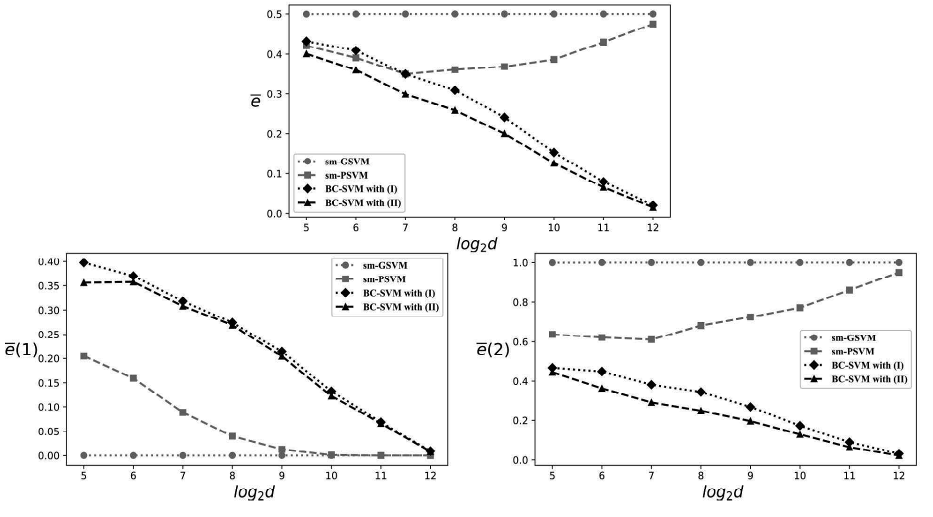

Figure 1: The error rates of the BC-SVM with (I), BC-SVM with (II), sm-GSVM and sm-PSVM for (a). The left panel displays $\overline{e}(1)$, the right panel displays $\overline{e}(2)$ and the top panel displays $\overline{e}$ for $d = 2^s,\ s = 5, \ldots, 12$.

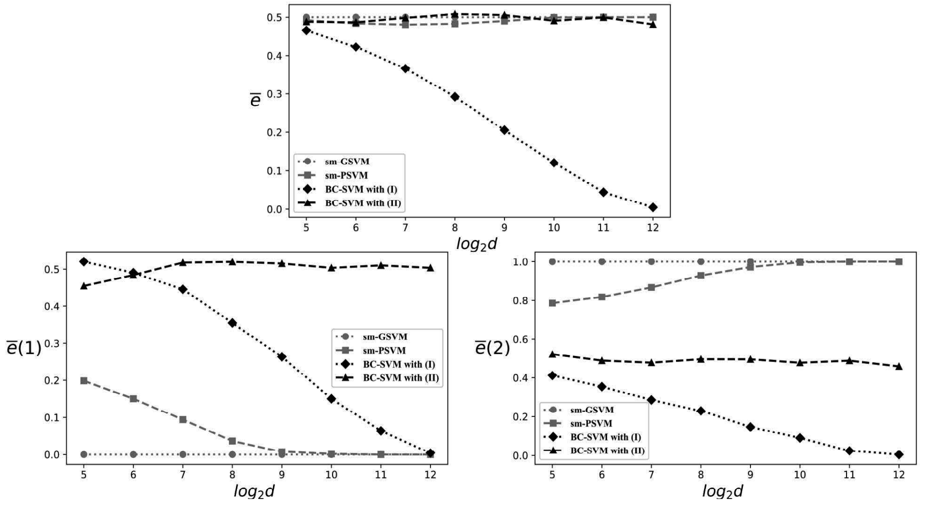

Figure 2: The error rates of the BC-SVM with (I), BC-SVM with (II), sm-GSVM and sm-PSVM for (b). The left panel displays $\overline{e}(1)$, the right panel displays $\overline{e}(2)$ and the top panel displays $\overline{e}$ for $d = 2^s,\ s = 5, \ldots, 12$.

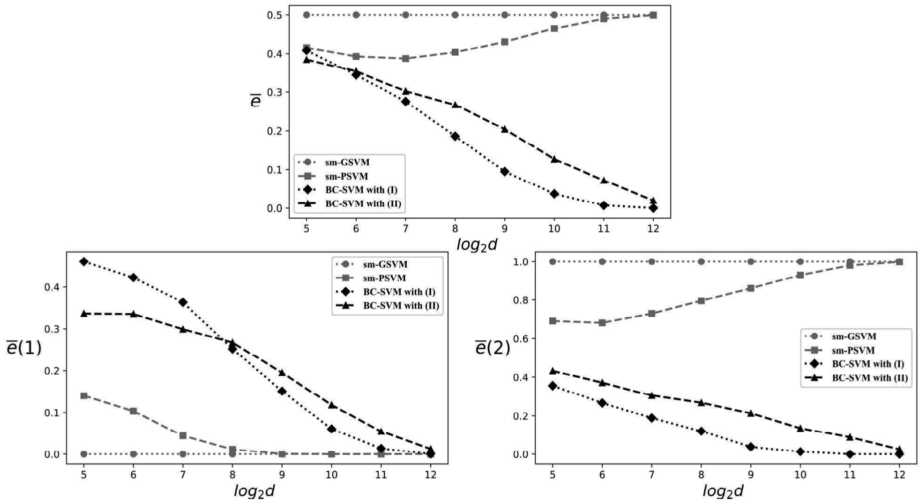

Figure 3: The error rates of the BC-SVM with (I), BC-SVM with (II), sm-GSVM and sm-PSVM for (c). The left panel displays $\overline{e}(1)$, the right panel displays $\overline{e}(2)$ and the top panel displays $\overline{e}$ for $d = 2^s,\ s = 5, \ldots, 12$.

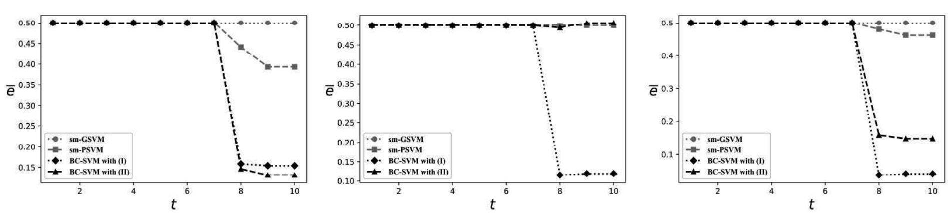

Figure 4: The error rates of the BC-SVM with (I), BC-SVM with (II), sm-GSVM and sm-PSVM for (a) to (c) when $d = 1024$ and $C = 2^{-7+t}/(n_{\min}\Delta_*),\ t = 1, \ldots, 10$. The left panel displays (a), the middle panel displays (b) and the right panel displays (c).

6 Proofs

6.1 Proofs of Theorem 1 and Corollary 1

Assume (A-i), (A-ii) and (5) and (6). From Proposition 1 and Lemma 4 in Nakayama et al. [10], we have that as $d \to \infty$

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

for the Gaussian kernel. Then, similar to the proof of Proposition 1 in Nakayama et al. [10], we can conclude the result of Theorem 1. From Theorem 1, we conclude the results of Corollary 1.

6.2 Proofs of Theorem 2 and Corollary 2

Assume (A-i), (A-ii) and (7) to (9). From Propositions 1 and 8 in Nakayama et al. [10], we have that as $d \to \infty$

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

for the polynomial kernel. Then, similar to the proof of Proposition 1 in Nakayama et al. [10], we can conclude the result of Theorem 2. From Theorem 2, we conclude the results of Corollary 2.

6.3 Proofs of Theorems 3 and 4

By combining Theorem 2 in Nakayama et al. [10] with Theorems 1 and 2, we can conclude the results.

Acknowledgements

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), JSPS, under Contract Numbers 15H01678.

References

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

[2] 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.

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

[4] Aoshima, M., Yata, K. (2018). High-dimensional quadratic classifiers in non-sparse settings. *Methodology and Computing in Applied Probability*, in press (doi:10.1007/s11009-018-9646-z).

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

[6] 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.

[7] 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.

[8] 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.

[9] 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).

[10] Nakayama, Y., Yata, K., Aoshima, M. (2019). Bias-corrected support vector machine with Gaussian kernel in high-dimension, low-sample-size settings. Revised in *Annals of the Institute of Statistical Mathematics*.

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

[12] Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory (second ed.)*. New York: Springer.