数理解析研究所講究録
第2047巻 2017年 10-18

Asymptotic properties of support vector machines in HDLSS settings

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

Abstract

In this paper, we consider asymptotic properties of the support vector machine (SVM) in high-dimension, low-sample-size (HDLSS) settings. We first show that the linear SVM holds a consistency property in which misclassification rates tend to zero as the dimension goes to infinity under certain severe conditions. Next, we consider a non-linear SVM based on the Gaussian kernel in HDLSS settings. We also show that the non-linear SVM holds the consistency property under mild conditions. Finally, we check the performance of the SVMs by numerical simulations.

Keywords and phrases: Hard-margin SVM; Large p small n; Radial basis function kernel

1 Introduction

High-dimension, low-sample-size (HDLSS) data situations occur in many areas of modern science such as genetic microarrays, medical imaging, text recognition, finance, chemometrics, and so on. Suppose we have independent and d-variate two populations, $\pi_i,\ i=1,2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i\ (\geq \boldsymbol{O})$. We assume that $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d \in (0,\infty)$ as $d\to\infty$ for $i=1,2$. Here, for a function, $f(\cdot)$, "$f(d)\in(0,\infty)$ as $d\to\infty$" implies $\liminf_{d\to\infty} f(d)>0$ and $\limsup_{d\to\infty} f(d)<\infty$. Let $\Delta=\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2$, where $\|\cdot\|$ denotes the Euclidean norm. We assume that $\limsup_{d\to\infty}\Delta/d<\infty$. We have independent and identically distributed (i.i.d.) observations, $\boldsymbol{x}_{i1},...,\boldsymbol{x}_{in_i}$, 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. We assume $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent. Let $N = n_1 + n_2$.

In the HDLSS context, Hall et al. [6] and Marron et al. [7] considered distance weighted classifiers. Aoshima and Yata [2] and Chan and Hall [5] 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. They showed that the misclassification rates tend to 0 as d increases, i.e.,

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

under the non-sparsity such as $\Delta \to \infty$ as $d \to \infty$, where $e(i)$ denotes the error rate of misclassifying an individual from π_i into the other class. We call (1) "the consistency property".

In the field of machine learning, there are many studies about the classification in the context of supervised learning. A typical method is the support vector machine (SVM). The SVM has versatility and effectiveness both for low-dimensional and high-dimensional data. See Schölkopf and Smola [9] and Vapnik [10] for the details. Even though the SVM is quite popular, its asymptotic properties seem to have not been studied sufficiently. Recently, Nakayama et al. [8] investigated asymptotic properties of a linear SVM for HDLSS data.

In this paper, we investigate linear and non-linear SVMs in the HDLSS context where $d \to \infty$ while N is fixed. In Section 2, we show that the linear SVM holds (1) under certain severe conditions. In Section 3, we consider a non-linear SVM based on the Gaussian kernel in HDLSS settings. We also show that the non-linear SVM holds (1) under mild conditions. Finally, we check the performance of the SVMs by numerical simulations.

2 Linear SVM in HDLSS settings

In this section, we give asymptotic properties of the linear SVM in HDLSS settings. Since HDLSS data are linearly separable by a hyperplane, we consider the hard-margin linear SVM.

2.1 Hard-margin linear SVM

We consider the following linear classifier:

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

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

$$\operatorname*{argmin}_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \quad \text{subject to} \quad t_j(\boldsymbol{w}^T\boldsymbol{x}_j + b) \geq 1,\ j = 1, ..., N.$$

A Lagrangian formulation is given by

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

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

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

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

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^{N} \alpha_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} \alpha_j\alpha_k t_j t_k \boldsymbol{x}_j^T\boldsymbol{x}_k.$$

The optimization problem can be transformed into the following:

$$\operatorname*{argmax}_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha})$$

subject to

$$\alpha_j \geq 0,\ j = 1, ..., N, \quad \text{and} \quad \sum_{j=1}^{N} \alpha_j t_j = 0. \tag{3}$$

Let us write that

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, ..., \hat{\alpha}_N)^T = \operatorname*{argmax}_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) \quad \text{subject to (3).}$$

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

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

Then, the linear classifier in (2) is defined by

$$\hat{y}(\boldsymbol{x}) = \sum_{k\in\hat{S}} \hat{\alpha}_k t_k \boldsymbol{x}_k^T\boldsymbol{x} + \hat{b}. \tag{4}$$

Finally, in the SVM, one classifies $\boldsymbol{x}_0$ into π_1 if $\hat{y}(\boldsymbol{x}_0) < 0$ and into π_2 otherwise. See Vapnik [10] for the details.

2.2 Asymptotic properties of the linear SVM in the HDLSS context

We assume the following assumptions:

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

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

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

$$\delta = \Delta + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1)}{n_1} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_2)}{n_2} \quad \text{and} \quad \kappa = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1)}{n_1} - \frac{\mathrm{tr}(\boldsymbol{\Sigma}_2)}{n_2}.$$

Then, Nakayama et al. [8] gave the following results.

Lemma 2.1 ([8]). *Under (A-i) and (A-ii), it holds that as $d \to \infty$*

$$\hat{\alpha}_j = \frac{2}{\delta n_1}\{1 + o_p(1)\} \quad \textit{for } j = 1, ..., n_1; \quad \textit{and}$$
$$\hat{\alpha}_j = \frac{2}{\delta n_2}\{1 + o_p(1)\} \quad \textit{for } j = n_1 + 1, ..., N.$$

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

$$\hat{y}(\boldsymbol{x}_0) = \frac{(-1)^i\Delta}{\delta} + \frac{\kappa}{\delta} + o_p\Big(\frac{\Delta}{\delta}\Big) \quad \textit{when } \boldsymbol{x}_0 \in \pi_i,\ i = 1, 2.$$

From Lemma 2.1, it holds that as $d \to \infty$

$$\frac{\delta}{\Delta}\hat{y}(\boldsymbol{x}_0) = (-1)^i + \frac{\kappa}{\Delta} + o_p(1) \tag{5}$$

when $\boldsymbol{x}_0 \in \pi_i$, $i = 1, 2$. Hence, "κ/Δ" is the bias term of the (normalized) SVM. We consider the following assumption:

(A-iii) $\limsup_{d\to\infty} \dfrac{|\kappa|}{\Delta} < 1.$

Then, Nakayama et al. [8] gave the following results.

Theorem 2.1 ([8]). *Under (A-i) to (A-iii), the linear SVM holds (1).*

Corollary 2.1 ([8]). *Under (A-i) and (A-ii), the linear SVM holds the following properties:*

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

We expect from (5) that, for sufficiently large d, $e(1)$ and $e(2)$ for the SVM become small and $e(1)$ (or $e(2)$) is larger than $e(2)$ (or $e(1)$) if $\kappa/\Delta > 0$ (or $\kappa/\Delta < 0$). In addition, from Corollary 2.1, if $\liminf_{d\to\infty} |\kappa|/\Delta > 1$, one should not use the SVM. In order to overcome the difficulties, Nakayama et al. [8] proposed a bias-corrected SVM (BC-SVM). They showed that the BC-SVM gives preferable performances even when (A-iii) is not met.

3 Non-linear SVM in HDLSS settings

In this section, we consider a non-linear SVM based on the Gaussian kernel. We give asymptotic properties of the non-linear SVM in HDLSS settings.

The optimization problem of the non-linear SVM can be written as follows: Let

$$L_*(\boldsymbol{\alpha}) = \sum_{j=1}^{N} \alpha_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} \alpha_j\alpha_k t_j t_k \exp\Big(- \frac{\|\boldsymbol{x}_j - \boldsymbol{x}_k\|^2}{\gamma}\Big),$$

where $\gamma > 0$ is a tuning parameter. The optimization problem can be transformed into the following:

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

subject to (3). Let us write that

$$\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, ..., \tilde{\alpha}_N)^T = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L_*(\boldsymbol{\alpha}) \ \text{ subject to (3)}.$$

Let $\tilde{S} = \{j | \tilde{\alpha}_j \neq 0,\ j = 1, ..., N\}$ and $N_{\tilde{S}} = \#\tilde{S}$. The intercept term is given by

$$\tilde{b} = \frac{1}{N_{\tilde{S}}} \sum_{j\in\tilde{S}} \Big(t_j - \sum_{k\in\tilde{S}} \tilde{\alpha}_k t_k \exp\Big(- \frac{\|\boldsymbol{x}_j - \boldsymbol{x}_k\|^2}{\gamma}\Big)\Big).$$

Then, the classifier is given by

$$\tilde{y}(\boldsymbol{x}) = \sum_{k\in\tilde{S}} \tilde{\alpha}_k t_k \exp\Big(- \frac{\|\boldsymbol{x}_k - \boldsymbol{x}\|^2}{\gamma}\Big) + \tilde{b}. \qquad (6)$$

Finally, in the non-linear SVM, one classifies $\boldsymbol{x}_0$ into π_1 if $\tilde{y}(\boldsymbol{x}_0) < 0$ and into π_2 otherwise.

We assume the following condition for γ:

(A-iv) $\gamma/d \in (0, \infty)$ as $d \to \infty$.

Let

$$c_i = \exp\Big(- \frac{2\mathrm{tr}(\boldsymbol{\Sigma}_i)}{\gamma}\Big),\ \ i = 1, 2; \quad \text{and} \quad c_3 = \exp\Big(- \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2) + \Delta}{\gamma}\Big).$$

Let $\Delta_* = c_1 + c_2 - 2c_3$, $\delta_* = \Delta_* + \sum_{i=1}^{2}(1 - c_i)/n_i$ and $\kappa_* = (1 - c_1)/n_1 - (1 - c_2)/n_2$. Here, we assume the following assumptions:

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

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

We have the following result.

Lemma 3.1. *Assume (A-iv) to (A-vi). It holds that as $d \to \infty$*

$$\begin{aligned} \tilde{\alpha}_j &= \frac{2}{\delta_* n_1}\{1 + o_p(1)\} \quad \textit{for } j = 1, ..., n_1; \quad \textit{and} \\ \tilde{\alpha}_j &= \frac{2}{\delta_* n_2}\{1 + o_p(1)\} \quad \textit{for } j = n_1 + 1, ..., N. \end{aligned}$$

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

$$\tilde{y}(\boldsymbol{x}_0) = \frac{(-1)^i\Delta_*}{\delta_*} + \frac{\kappa_*}{\delta_*} + o_p\Big(\frac{\Delta_*}{\delta_*}\Big) \quad \textit{when } \boldsymbol{x}_0 \in \pi_i \textit{ for } i = 1, 2. \tag{7}$$

We consider the following assumption:

(A-vii) $\limsup_{d\to\infty} \dfrac{|\kappa_*|}{\Delta_*} < 1$.

Then, from Lemma 3.1, we have the following result.

Theorem 3.1. *Under (A-iv) to (A-vii), the non-linear SVM holds (1).*

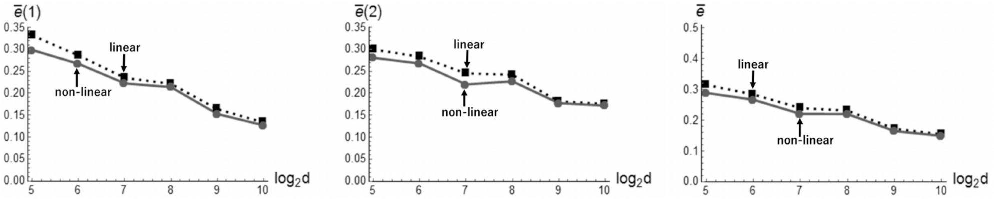
Now, we consider the following conditions:

$$\mathrm{Var}(\|\boldsymbol{x}_{ij}-\boldsymbol{\mu}_i\|^2) = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\} \text{ and } \mathrm{tr}(\boldsymbol{\Sigma}_i^2)/d^2 \to 0 \text{ as } d \to \infty \text{ for } i = 1, 2. \tag{8}$$

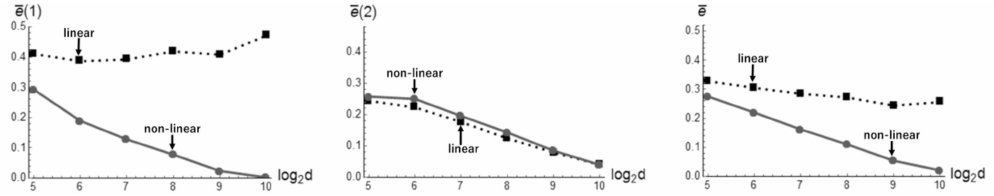
We note that

$$\Delta_* \geq \big[\exp\{-\mathrm{tr}(\boldsymbol{\Sigma}_1)/\gamma\} - \exp\{-\mathrm{tr}(\boldsymbol{\Sigma}_2)/\gamma\}\big]^2.$$

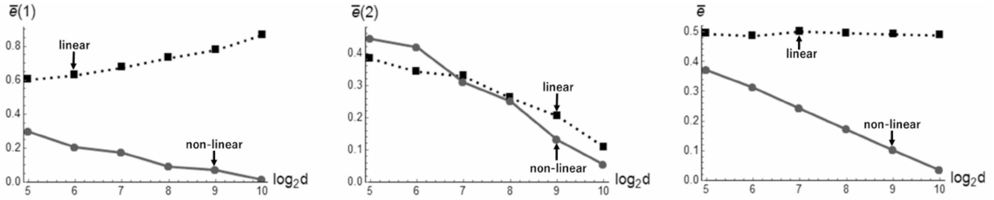
If one can assume that $\liminf_{d\to\infty} |\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2) - 1| > 0$, it follows $\liminf_{d\to\infty} \Delta_* > 0$ under (A-iv), so that (A-v) and (A-vi) hold under (8). Thus the non-linear SVM has the consistency even when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. We emphasize that the non-linear SVM based on the Gaussian kernel draws information about heteroscedasticity via the difference of $\mathrm{tr}(\boldsymbol{\Sigma}_i)$s.



(a) $\boldsymbol{\mu}_1 = \mathbf{1}_d$, $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_*$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Phi}_1$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Phi}_2$.



(b) $\boldsymbol{\mu}_1 = \mathbf{1}_d$, $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_*$, $\boldsymbol{\Sigma}_1 = 1.4\boldsymbol{\Phi}_1$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Phi}_2$.



(c) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{1}_d$, $\boldsymbol{\Sigma}_1 = 1.4\boldsymbol{\Phi}_1$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Phi}_2$.

Figure 1: The performance of the linear SVM and the non-linear SVM for (a) to (c). The error rates of the linear SVM are denoted by the dotted lines, and those of the non-linear SVM are denoted by the solid lines.

4 Simulation

In this section, we compare the performance of the linear SVM given by (4) and the non-linear SVM given by (6) in numerical simulations.

We set $d = 2^s$, $s = 5, ..., 10$, and $(n_1, n_2) = (5, 5)$. We generated $\boldsymbol{x}_{ij}$, $j = 1, 2, ...$, $(i = 1, 2)$ independently from $\pi_i : N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. We set $\boldsymbol{\mu}_* = (1, ..., 1, 0, ..., 0)^T$ whose last $\lceil d^{2/3} \rceil$ elements are 0, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. Let $\boldsymbol{\Phi}_1 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$, $\boldsymbol{\Phi}_2 = \boldsymbol{B}(0.4^{|i-j|^{1/3}})\boldsymbol{B}$ and

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

We considered three cases :

(a) $\boldsymbol{\mu}_1 = \boldsymbol{1}_d = (1, ..., 1)^T$, $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_*$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Phi}_1$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Phi}_2$;
(b) $\boldsymbol{\mu}_1 = \boldsymbol{1}_d$, $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_*$, $\boldsymbol{\Sigma}_1 = 1.4\boldsymbol{\Phi}_1$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Phi}_2$; and
(c) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{1}_d$, $\boldsymbol{\Sigma}_1 = 1.4\boldsymbol{\Phi}_1$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Phi}_2$.

For $\boldsymbol{x}_0 \in \pi_i$ $(i = 1, 2)$ we calculated each classifier 2000 times to confirm if each rule does (or does not) classify $\boldsymbol{x}_0$ correctly and defined $P_{ir} = 0$ (or 1) accordingly for each π_i. 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, $\overline{e} = \{\overline{e}(1) + \overline{e}(2)\}/2$. For the Gaussian kernel, we chose γ from the candidates, $d^{(t+5)/10}$, $t = 1, ..., 10$, by a cross-validation procedure. Their standard deviations are less than 0.011. In Figure 1, we plotted $\overline{e}(1)$, $\overline{e}(2)$ and $\overline{e}$ for (a) to (c).

We observed that the SVMs give preferable performances for (a) in Figure 1. However, the linear SVM gave a quite bad performance for (c). This is because of $\Delta = 0$ for (c). On the other hand, the non-linear SVM gave a better performance compared to the linear SVM for (b) and (c). This is because the non-linear SVM draws information about heteroscedasticity from the difference of $\text{tr}(\boldsymbol{\Sigma}_i)$s. See Section 3 for the details.

5 Appendix

Proof of Lemma 3.1. Similarly to the proof of Lemma 1 in Nakayama et al. [8], we have that as $d \to \infty$

$$L_*(\boldsymbol{\alpha}) = 2\alpha_\star - \frac{\Delta_*}{2}\alpha_\star^2\{1 + o_p(1)\} - \frac{1}{2}\Big((1 - c_1)\sum_{j=1}^{n_1}\alpha_j^2 + (1 - c_2)\sum_{j=n_1+1}^{N}\alpha_j^2\Big)$$

subject to (3) under (A-iv) to (A-vi), where $\alpha_\star = \sum_{j=1}^{n_1}\alpha_j$. Then, by noting

$$\liminf_{d\to\infty}(1 - c_i)/\Delta_* > 0, \ i = 1, 2,$$

under (A-iv), in a way similar to the proof of Lemma 2 in Nakayama et al. [8], we can obtain the result. □

Proof of Theorem 3.1. By using (7), the result is obtained straightforwardly. □

Acknowledgements

Research of the second author was partially supported by Grant-in-Aid for Young Scientists (B), Japan Society for the Promotion of Science (JSPS), under Contract Number 26800078. Research of the third author was partially supported by Grants-in-Aid for Scientific Research (A), JSPS, under Contract Number 15H01678.

References

[1] Aoshima, M., Yata, K., 2011. Two-stage procedures for high-dimensional data. Sequential Anal. (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. Ann. Inst. Statist. Math. 66, 983-1010.

[3] Aoshima, M., Yata, K., 2015a. Geometric classifier for multiclass, high-dimensional data. Sequential Anal. 34, 279-294.

[4] Aoshima, M., Yata, K., 2015b. High-dimensional quadratic classifiers in non-sparse settings. arXiv:1503.04549.

[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. J. R. Statist. Soc. B 67, 427-444.

[7] Marron, J.S., Todd, M.J., Ahn, J., 2007. Distance-weighted discrimination. J. Amer. Statist. Assoc. 102, 1267-1271.

[8] Nakayama, Y., Yata, K., Aoshima, M., 2017. Support vector machine and its bias correction in high-dimension, low-sample-size settings. Revised in J. Stat. Plan. Infer. (arXiv:1702.08019).

[9] Schölkopf, B., Smola, A.J., 2002. Learning with Kernels. MIT Press, Cambridge.

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