A Bahadur Representation of the Linear Support Vector Machine

Ja-Yong Koo JYKOO@KOREA.AC.KR
Department of Statistics
Korea University
Seoul, 136-701, Korea

Yoonkyung Lee YKLEE@STAT.OSU.EDU
Department of Statistics
The Ohio State University
Columbus, OH 43210, USA

Yuwon Kim YUWONKIM@NAVER.COM
Data Mining Team
NHN Inc.
Gyeonggi-do 463-847, Korea

Changyi Park PARK463@UOS.AC.KR
Department of Statistics
University of Seoul
Jeonnong-dong 90, Dongdaemun-gu
Seoul, 130-743, Korea

Abstract

The support vector machine has been successful in a variety of applications. Also on the theoretical front, statistical properties of the support vector machine have been studied quite extensively with a particular attention to its Bayes risk consistency under some conditions. In this paper, we study somewhat basic statistical properties of the support vector machine yet to be investigated, namely the asymptotic behavior of the coefficients of the linear support vector machine. A Bahadur type representation of the coefficients is established under appropriate conditions, and their asymptotic normality and statistical variability are derived on the basis of the representation. These asymptotic results do not only help further our understanding of the support vector machine, but also they can be useful for related statistical inferences.

Keywords: Asymptotic Normality, Bahadur Representation, Classification, Convexity Lemma, Radon Transform

1. Introduction

The support vector machine (SVM) introduced by Cortes and Vapnik (1995) has been successful in many applications due to its high classification accuracy and flexibility. For reference, see Vapnik (1996), Schölkopf and Smola (2002), and Cristianini and Shawe-Taylor (2000). In parallel with a wide range of applications, statistical properties of the SVM have been studied by many researchers recently in addition to the statistical learning theory by

Vapnik (1996) that originally motivated the SVM. These include studies on the Bayes risk consistency of the SVM (Lin, 2002; Zhang, 2004; Steinwart, 2005) and its rate of convergence to the Bayes risk (Lin, 2000; Blanchard, Bousquet, and Massart, 2004; Scovel and Steinwart, 2006; Bartlett, Jordan, and McAuliffe, 2006). While the existing theoretical analysis of the SVM largely concerns its asymptotic risk, there are some basic statistical properties of the SVM that seem to have eluded our attention. For example, to the best of our knowledge, large sample properties of the coefficients in the linear SVM have not been studied so far although the magnitude of each coefficient is often the determining factor of feature selection for the SVM in practice.

In this paper, we address this basic question of the statistical behavior of the linear SVM as a first step to the study of more general properties of the SVM. We mainly investigate asymptotic properties of the coefficients of variables in the SVM solution for linear classification. The investigation is done in the standard way that parametric methods are studied in a finite dimensional setting, that is, the number of variables is assumed to be fixed and the sample size grows to infinity. Additionally, in the asymptotics, the effect of regularization through maximization of the class margin is assumed to vanish at a certain rate so that the solution is ultimately governed by the empirical risk. Due to these assumptions, the asymptotic results become more pertinent to the classical parametric setting where the number of features is moderate compared to the sample size and the virtue of regularization is minute than to the situation with high dimensional inputs. Despite the difference between the practical situation where the SVM methods are effectively used and the setting theoretically posited in this paper, the asymptotic results shed a new light on the SVM from a classical parametric point of view. In particular, we establish a Bahadur type representation of the coefficients as in the studies of sample quantiles and estimates of regression quantiles. See Bahadur (1966) and Chaudhuri (1991) for reference. It turns out that the Bahadur type representation of the SVM coefficients depends on Radon transform of the second moments of the variables. This representation illuminates how the so called margins of the optimal separating hyperplane and the underlying probability distribution within and around the margins determine the statistical behavior of the estimated coefficients. Asymptotic normality of the coefficients then follows immediately from the representation. The proximity of the hinge loss function that defines the SVM solution to the absolute error loss and its convexity allow such asymptotic results akin to those for least absolute deviation regression estimators in Pollard (1991).

In addition to providing an insight into the asymptotic behavior of the SVM, we expect that our results can be useful for related statistical inferences on the SVM, for instance, feature selection. For introduction to feature selection, see Guyon and Elisseeff (2003), and for an extensive empirical study of feature selection using SVM-based criteria, see Ishak and Ghattas (2005). In particular, Guyon, Weston, Barnhill, and Vapnik (2002) proposed a recursive feature elimination procedure for the SVM with an application to gene selection in microarray data analysis. Its selection or elimination criterion is based on the absolute value of a coefficient not its standardized value. The asymptotic variability of estimated coefficients that we provide can be used in deriving a new feature selection criterion which takes inherent statistical variability into account.

This paper is organized as follows. Section 2 contains the main results of a Bahadur type representation of the linear SVM coefficients and their asymptotic normality under mild

conditions. An illustrative example is then provided in Section 3 followed by simulation studies in Section 4. Proofs of technical lemmas and theorems are collected in Section 5, ensued by a discussion in Section 6.

2. Main Results

2.1 Preliminaries

Let (X, Y) be a pair of random variables with $X \in \mathcal{X} \subset \mathbb{R}^d$ and $Y \in \{1, -1\}$. The marginal distribution of Y is given by $\mathbb{P}(Y = 1) = \pi_+$ and $\mathbb{P}(Y = -1) = \pi_-$ with $\pi_+, \pi_- > 0$ and $\pi_+ + \pi_- = 1$. Let f and g be the densities of X given $Y = 1$ and -1, respectively. Let $\{(X^i, Y^i)\}_{i=1}^n$ be a set of training data, independently drawn from the distribution of (X, Y). Denote the input variables as $x = (x_1, \ldots, x_d)^\top$ and their coefficients as $\beta_+ = (\beta_1, \ldots, \beta_d)^\top$. Let $\widetilde{x} = (\widetilde{x}_0, \ldots, \widetilde{x}_d)^\top = (1, x_1, \ldots, x_d)^\top$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_d)^\top$. We consider linear classifications with hyperplanes defined by $h(x; \beta) = \beta_0 + x^\top \beta_+ = \widetilde{x}^\top \beta$. Let $\|\cdot\|$ denote the Euclidean norm of a vector. For separable cases, the SVM finds the hyperplane that maximizes the geometric margin, $2/\|\beta_+\|^2$ subject to the constraints $y^i h(x^i; \beta) \geq 1$ for $i = 1, \ldots, n$. For non-separable cases, a soft-margin SVM is introduced to minimize

$$C \sum_{i=1}^n \xi_i + \frac{1}{2}\|\beta_+\|^2$$

subject to the constraints $\xi_i \geq 1 - y^i h(x^i; \beta)$ and $\xi_i \geq 0$ for $i = 1, \ldots, n$, where $C > 0$ is a tuning parameter and $\{\xi_i\}_{i=1}^n$ are called the slack variables. Equivalently, the SVM minimizes the unconstrained objective function

$$l_{\lambda,n}(\beta) = \frac{1}{n} \sum_{i=1}^n \Big[1 - y^i h(x^i; \beta)\Big]_+ + \frac{\lambda}{2}\|\beta_+\|^2 \tag{1}$$

over $\beta \in \mathbb{R}^{d+1}$, where $[z]_+ = \max(z, 0)$ for $z \in \mathbb{R}$ and $\lambda > 0$ is a penalization parameter; see Vapnik (1996) for details. Let the minimizer of (1) be denoted by $\widehat{\beta}_{\lambda,n} = \arg\min_\beta l_{\lambda,n}(\beta)$. Note that $C = (n\lambda)^{-1}$. Choice of λ depends on the data, and usually it is estimated via cross validation in practice. In this paper, we consider only nonseparable cases and assume that $\lambda \to 0$ as $n \to \infty$. We note that separable cases require a different treatment for asymptotics because λ has to be nonzero in the limit for the uniqueness of the solution.

Before we proceed with a discussion of the asymptotics of the $\widehat{\beta}_{\lambda,n}$, we introduce some notation and definitions first. The population version of (1) without the penalty term is defined as

$$L(\beta) = \mathbb{E}\Big[1 - Y h(X; \beta)\Big]_+ \tag{2}$$

and its minimizer is denoted by $\beta^* = \arg\min_\beta L(\beta)$. Then the population version of the optimal hyperplane defined by the SVM is

$$\widetilde{x}^\top \beta^* = 0. \tag{3}$$

Sets are identified with their indicator functions. For example, $\displaystyle\int_{\mathcal{X}} x_j \{x_j > 0\} f(x) dx = \int_{\{x \in \mathcal{X} \,:\, x_j > 0\}} x_j f(x) dx$. Letting $\psi(z) = \{z \geq 0\}$ for $z \in \mathbb{R}$, we define $S(\beta) = (S(\beta)_j)$ to be

the $(d+1)$-dimensional vector given by

$$S(\beta) = -\mathbb{E}\Big(\psi(1 - Yh(X;\beta))Y\widetilde{X}\Big)$$

and $H(\beta) = (H(\beta)_{jk})$ to be the $(d+1) \times (d+1)$-dimensional matrix given by

$$H(\beta) = \mathbb{E}\Big(\delta(1 - Yh(X;\beta))\widetilde{X}\widetilde{X}^\top\Big),$$

where δ denotes the Dirac delta function with $\delta(t) = \psi'(t)$ in distributional sense. Provided that $S(\beta)$ and $H(\beta)$ are well-defined, $S(\beta)$ and $H(\beta)$ are considered as the gradient and Hessian matrix of $L(\beta)$, respectively. Formal proofs of these relationships are given in Section 5.1.

For explanation of $H(\beta)$, we introduce a Radon transformation. For a function s on $\mathcal{X}$, define the Radon transform $\mathcal{R}s$ of s for $p \in \mathbb{R}$ and $\xi \in \mathbb{R}^d$ as

$$(\mathcal{R}s)(p,\xi) = \int_{\mathcal{X}} \delta(p - \xi^\top x)s(x)dx.$$

Denote

$$s_j(x) = \widetilde{x}_j s(x), \quad s_{jk}(x) = \widetilde{x}_j\widetilde{x}_k s(x) \quad \text{for} \quad 0 \le j,k \le d.$$

Then, it can be seen that

$$H(\beta)_{jk} = \pi_+(\mathcal{R}f_{jk})(1 - \beta_0, \beta_+) + \pi_-(\mathcal{R}g_{jk})(1 + \beta_0, -\beta_+). \tag{4}$$

Equation (4) shows that the Hessian matrix $H(\beta)$ depends on the Radon transforms of f, g, f_j, g_j, f_{jk} and g_{jk}. For Radon transform and its properties in general, see Natterer (1986), Deans (1993), or Ramm and Katsevich (1996).

For a continuous integrable function s, it can be easily proved that $\mathcal{R}s$ is continuous. If f and g are continuous densities with finite second moments, then f_{jk} and g_{jk} are continuous and integrable. Hence $H(\beta)$ is continuous in β when f and g are continuous and have finite second moments.

2.2 Asymptotics

Now we present the asymptotic results for $\widehat{\beta}_{\lambda,n}$. We state regularity conditions for the asymptotics first. Some remarks on the conditions then follow for exposition and clarification. Throughout this paper, we use $C_1, C_2, \ldots$ to denote positive constants independent of n.

(A1) The densities f and g are continuous and have finite second moments.

(A2) There exists $B(x_0, \delta_0)$, a ball centered at x_0 with radius $\delta_0 > 0$ such that $f(x) > C_1$ and $g(x) > C_1$ for every $x \in B(x_0, \delta_0)$.

(A3) For some $1 \le i^* \le d$,

$$\int_{\mathcal{X}} \{x_{i^*} \ge G_{i^*}^-\} x_{i^*} g(x)dx < \int_{\mathcal{X}} \{x_{i^*} \le F_{i^*}^+\} x_{i^*} f(x)dx$$

or

$$\int_{\mathcal{X}} \{x_{i^*} \leq G_{i^*}^+\} x_{i^*} g(x) dx > \int_{\mathcal{X}} \{x_{i^*} \geq F_{i^*}^-\} x_{i^*} f(x) dx.$$

Here $F_{i^*}^+, G_{i^*}^+ \in [-\infty, \infty]$ are upper bounds such that $\int_{\mathcal{X}} \{x_{i^*} \leq F_{i^*}^+\} f(x) dx = \min\left(1, \frac{\pi_-}{\pi_+}\right)$ and $\int_{\mathcal{X}} \{x_{i^*} \leq G_{i^*}^+\} g(x) dx = \min\left(1, \frac{\pi_+}{\pi_-}\right)$. Similarly, lower bounds $F_{i^*}^-$ and $G_{i^*}^-$ are defined as $\int_{\mathcal{X}} \{x_{i^*} \geq F_{i^*}^-\} f(x) dx = \min\left(1, \frac{\pi_-}{\pi_+}\right)$ and $\int_{\mathcal{X}} \{x_{i^*} \geq G_{i^*}^-\} g(x) dx = \min\left(1, \frac{\pi_+}{\pi_-}\right)$.

(A4) For an orthogonal transformation A_{j^*} that maps $\beta_+^*/\|\beta_+^*\|$ to the j^*-th unit vector e_{j^*} for some $1 \leq j^* \leq d$, there exist rectangles

$$\mathcal{D}^+ = \{x \in M^+ : l_i \leq (A_{j^*} x)_i \leq v_i \text{ with } l_i < v_i \text{ for } i \neq j^*\}$$

and

$$\mathcal{D}^- = \{x \in M^- : l_i \leq (A_{j^*} x)_i \leq v_i \text{ with } l_i < v_i \text{ for } i \neq j^*\}$$

such that $f(x) \geq C_2 > 0$ on $\mathcal{D}^+$, and $g(x) \geq C_3 > 0$ on $\mathcal{D}^-$, where $M^+ = \{x \in \mathcal{X} \mid \beta_0^* + x^\top \beta_+^* = 1\}$ and $M^- = \{x \in \mathcal{X} \mid \beta_0^* + x^\top \beta_+^* = -1\}$.

Remark 1

- *(A1) ensures that $H(\beta)$ is well-defined and continuous in β.*
- *When f and g are continuous, the condition that $f(x_0) > 0$ and $g(x_0) > 0$ for some x_0 implies (A2).*
- *The technical condition in (A3) is a minimal requirement to guarantee that β_+^*, the normal vector of the theoretically optimal hyperplane is not zero. Roughly speaking, it says that for at least one input variable, the mean values of the class conditional distributions f and g have to be different in order to avoid the degenerate case of $\beta_+^* = 0$. Some restriction of the supports through $F_{i^*}^+$, $G_{i^*}^+$, $F_{i^*}^-$ and $G_{i^*}^-$ is necessary in defining the mean values to adjust for potentially unequal class proportions. When $\pi_+ = \pi_-$, $F_{i^*}^+$ and $G_{i^*}^+$ can be taken to be $+\infty$ and $F_{i^*}^-$ and $G_{i^*}^-$ can be $-\infty$. In this case, (A3) simply states that the mean vectors for the two classes are different.*
- *(A4) is needed for the positive-definiteness of $H(\beta)$ around β^*. The condition means that there exist two subsets of the classification margins, M^+ and M^- on which the class densities f and g are bounded away from zero. For mathematical simplicity, the rectangular subsets $\mathcal{D}^+$ and $\mathcal{D}^-$ are defined as the mirror images of each other along the normal direction of the optimal hyperplane. This condition can be easily met when the supports of f and g are convex. Especially, if $\mathbb{R}^d$ is the support of f and g, it is trivially satisfied. (A4) requires that $\beta_+^* \neq 0$, which is implied by (A1) and (A3); see Lemma 4 for the proof. For the special case $d = 1$, M^+ and M^- consist of a point. $\mathcal{D}^+$ and $\mathcal{D}^-$ are the same as M^+ and M^-, respectively, and hence (A4) means that f and g are positive at those points in M^+ and M^-.*

Under the regularity conditions, we obtain a Bahadur-type representation of $\widehat{\beta}_{\lambda,n}$ (Theorem 1). The asymptotic normality of $\widehat{\beta}_{\lambda,n}$ follows immediately from the representation (Theorem 2). Consequently, we have the asymptotic normality of $h(x;\widehat{\beta}_{\lambda,n})$, the value of the SVM decision function at x (Corollary 3).

Theorem 1 *Suppose that (A1)-(A4) are met. For $\lambda = o(n^{-1/2})$, we have*

$$\sqrt{n}(\widehat{\beta}_{\lambda,n} - \beta^*) = -\frac{1}{\sqrt{n}} H(\beta^*)^{-1} \sum_{i=1}^{n} \psi(1 - Y^i h(X^i;\beta^*)) Y^i \widetilde{X}^i + o_{\mathbb{P}}(1).$$

Theorem 2 *Suppose (A1)-(A4) are satisfied. For $\lambda = o(n^{-1/2})$,*

$$\sqrt{n}(\widehat{\beta}_{\lambda,n} - \beta^*) \to N\Big(0, H(\beta^*)^{-1} G(\beta^*) H(\beta^*)^{-1}\Big)$$

in distribution, where

$$G(\beta) = \mathbb{E}\Big(\psi(1 - Y h(X;\beta)) \widetilde{X}\widetilde{X}^\top\Big).$$

Remark 2 *Since $\widehat{\beta}_{\lambda,n}$ is a consistent estimator of β^* as $n \to \infty$, $G(\beta^*)$ can be estimated by its empirical version with β^* replaced by $\widehat{\beta}_{\lambda,n}$. To estimate $H(\beta^*)$, one may consider the following nonparametric estimate:*

$$\frac{1}{n}\left[\sum_{i=1}^{n} p_b\Big(1 - Y^i h(X^i;\widehat{\beta}_{\lambda,n})\Big) \widetilde{X}^i (\widetilde{X}^i)^\top\right],$$

where $p_b(t) \equiv p(t/b)/b$, $p(t) \geq 0$ and $\int_{\mathbb{R}} p(t)dt = 1$. Note that $p_b(\cdot) \to \delta(\cdot)$ as $b \to 0$. However, estimation of $H(\beta^*)$ requires further investigation.

Corollary 3 *Under the same conditions as in Theorem 2,*

$$\sqrt{n}\Big(h(x;\widehat{\beta}_{\lambda,n}) - h(x;\beta^*)\Big) \to N\Big(0, \tilde{x}^\top H(\beta^*)^{-1} G(\beta^*) H(\beta^*)^{-1} \tilde{x}\Big)$$

in distribution.

Remark 3 *Corollary 3 can be used to construct a confidence bound for $h(x;\beta^*)$ based on an estimate $h(x;\widehat{\beta}_{\lambda,n})$, in particular, to judge whether $h(x;\beta^*)$ is close to zero or not given x. This may be useful if one wants to abstain from prediction at a new input x if it is close to the optimal classification boundary $h(x;\beta^*) = 0$.*

3. An Illustrative Example

In this section, we illustrate the relation between the Bayes decision boundary and the optimal hyperplane determined by (2) for two multivariate normal distributions in $\mathbb{R}^d$. Assume that f and g are multivariate normal densities with different mean vectors μ_f and μ_g and a common covariance matrix Σ. Suppose that $\pi_+ = \pi_- = 1/2$.

We verify the assumptions (A1)-(A4) so that Theorem 2 is applicable. For normal densities f and g, (A1) holds trivially, and (A2) is satisfied with

$$C_1 \;=\; |2\pi\Sigma|^{-1/2} \exp\left(- \sup_{\|x\| \le \delta_0} \left\{ (x-\mu_f)^\top \Sigma^{-1}(x-\mu_f), (x-\mu_g)^\top \Sigma^{-1}(x-\mu_g) \right\} \right)$$

for $\delta_0 > 0$. Since $\mu_f \neq \mu_g$, there exists $1 \le i^* \le d$ such that i^*-th elements of μ_f and μ_g are different. By taking $F^+_{i^*} = G^+_{i^*} = +\infty$ and $F^-_{i^*} = G^-_{i^*} = -\infty$, we can show that one of the inequalities in (A3) holds as mentioned in Remark 1. Since $\mathcal{D}^+$ and $\mathcal{D}^-$ can be taken to be bounded sets of the form in (A4) in $\mathbb{R}^{d-1}$, and the normal densities f and g are bounded away from zero on such $\mathcal{D}^+$ and $\mathcal{D}^-$, (A4) is satisfied. In particular, $\beta^*_+ \neq 0$ as implied by Lemma 4.

Denote the density and cumulative distribution function of $N(0,1)$ as ϕ and Φ, respectively. Note that β^* should satisfy the equation $S(\beta^*) = 0$, or

$$\Phi(a_f) = \Phi(a_g) \tag{5}$$

and

$$\mu_f \Phi(a_f) - \phi(a_f)\Sigma^{1/2}\omega^* = \mu_g \Phi(a_g) + \phi(a_g)\Sigma^{1/2}\omega^*, \tag{6}$$

where $a_f = \dfrac{1 - \beta^*_0 - \mu_f^\top \beta^*_+}{\|\Sigma^{1/2}\beta^*_+\|}$, $a_g = \dfrac{1 + \beta^*_0 + \mu_g^\top \beta^*_+}{\|\Sigma^{1/2}\beta^*_+\|}$ and $\omega^* = \Sigma^{1/2}\beta^*_+ / \|\Sigma^{1/2}\beta^*_+\|$. From (5) and the definition of a_f and a_g, we have $a^* \equiv a_f = a_g$. Hence

$$(\beta^*_+)^\top (\mu_f + \mu_g) = -2\beta^*_0. \tag{7}$$

It follows from (6) that

$$\beta^*_+ / \|\Sigma^{1/2}\beta^*_+\| = \frac{\Phi(a^*)}{2\phi(a^*)} \Sigma^{-1}(\mu_f - \mu_g). \tag{8}$$

First we show the existence of a proper constant a^* satisfying (8) and its relationship with a statistical distance between the two classes. Define $\Upsilon(a) = \phi(a)/\Phi(a)$ and let $d_\Sigma(u,v) = \{(u-v)^\top \Sigma^{-1}(u-v)\}^{1/2}$ denote the Mahalanobis distance between u and $v \in \mathbb{R}^d$. Since $\|\omega^*\| = 1$, we have $\Upsilon(a^*) = \|\Sigma^{-1/2}(\mu_f - \mu_g)\|/2$. Since $\Upsilon(a)$ is monotonically decreasing in a, there exists $a^* = \Upsilon^{-1}(d_\Sigma(\mu_f, \mu_g)/2)$ that depends only on μ_f, μ_g, and Σ. For illustration, when the Mahalanobis distances between the two normal distributions are 2 and 3, a^* is given by $\Upsilon^{-1}(1) \approx -0.303$ and $\Upsilon^{-1}(1.5) \approx -0.969$, respectively. The corresponding Bayes error rates are about 0.1587 and 0.06681. Figure 1 shows a graph of $\Upsilon(a)$ and a^* when $d_\Sigma(\mu_f, \mu_g)$=2 and 3.

Once a^* is properly determined, we can express the solution β^* explicitly by (7) and (8):

$$\beta^*_0 = -\frac{(\mu_f - \mu_g)^\top \Sigma^{-1}(\mu_f + \mu_g)}{2a^* d_\Sigma(\mu_f, \mu_g) + d_\Sigma(\mu_f, \mu_g)^2}$$

and

$$\beta^*_+ = \frac{2\Sigma^{-1}(\mu_f - \mu_g)}{2a^* d_\Sigma(\mu_f, \mu_g) + d_\Sigma(\mu_f, \mu_g)^2}.$$

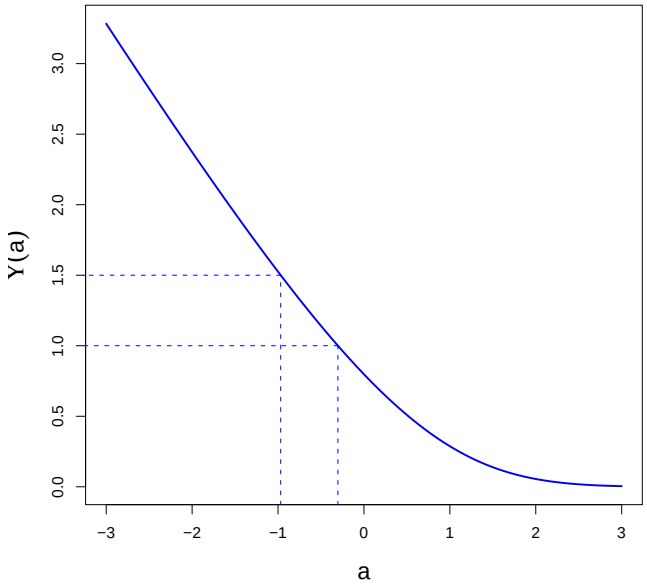


Figure 1: A plot of Υ function. The dashed lines correspond to the inverse mapping from the Mahalanobis distances of 2 and 3 to $a^* \approx -0.303$ and -0.969, respectively.

Thus the optimal hyperplane (3) is

$$\frac{2}{2a^* d_\Sigma(\mu_f, \mu_g) + d_\Sigma(\mu_f, \mu_g)^2} \left\{\Sigma^{-1}(\mu_f - \mu_g)\right\}^\top \left\{x - \frac{1}{2}(\mu_f + \mu_g)\right\} = 0,$$

which is equivalent to the Bayes decision boundary given by

$$\left\{\Sigma^{-1}(\mu_f - \mu_g)\right\}^\top \left\{x - \frac{1}{2}(\mu_f + \mu_g)\right\} = 0.$$

This shows that the linear SVM is equivalent to Fisher's linear discriminant analysis in this setting. In addition, $H(\beta^*)$ and $G(\beta^*)$ can be shown to be

$$G(\beta^*) = \frac{\Phi(a^*)}{2} \begin{bmatrix} 2 & (\mu_f + \mu_g)^\top \\ \mu_f + \mu_g & G_{22}(\beta^*) \end{bmatrix}$$

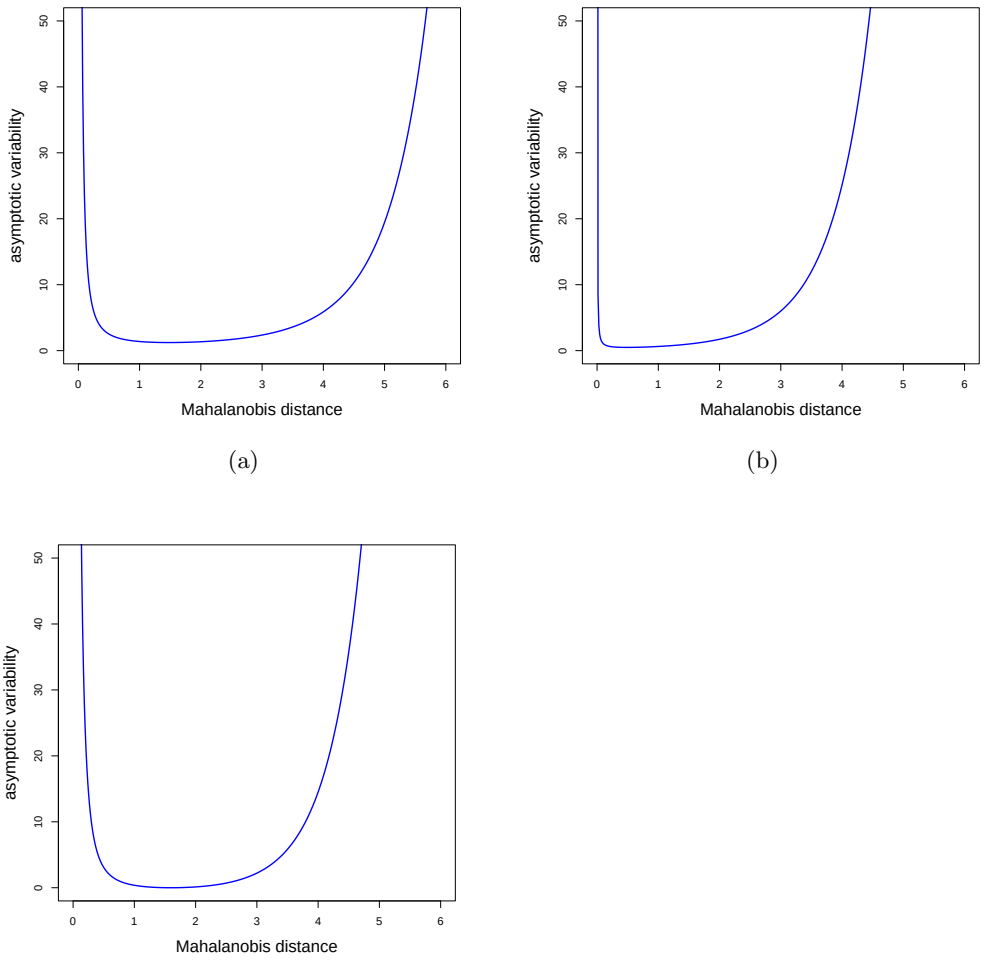
and

$$H(\beta^*) \quad = \quad \frac{\phi(a^*)}{4} \left(2a^* + d_\Sigma(\mu_f, \mu_g)\right) \begin{bmatrix} 2 & (\mu_f + \mu_g)^\top \\ \mu_f + \mu_g & H_{22}(\beta^*) \end{bmatrix},$$

where

$$G_{22}(\beta^*) = \mu_f\mu_f^\top + \mu_g\mu_g^\top + 2\Sigma - \left(\frac{a^*}{d_\Sigma(\mu_f,\mu_g)} + 1\right)(\mu_f-\mu_g)(\mu_f-\mu_g)^\top \text{ and}$$

$$H_{22}(\beta^*) = \mu_f\mu_f^\top + \mu_g\mu_g^\top + 2\Sigma + 2\left(\left(\frac{a^*}{d_\Sigma(\mu_f,\mu_g)}\right)^2 + \frac{a^*}{d_\Sigma(\mu_f,\mu_g)} - \frac{1}{d_\Sigma^2(\mu_f,\mu_g)}\right)(\mu_f-\mu_g)(\mu_f-\mu_g)^\top.$$



(a) (b)

(c)

Figure 2: The asymptotic variabilities of estimates of (a) the intercept, (b) the slope, and (c) their ratio for the optimal hyperplane as a function of the Mahalanobis distance.

For illustration, we consider the case when $d = 1$, $\mu_f + \mu_g = 0$, and $\sigma = 1$. The asymptotic variabilities of the intercept and the slope for the optimal decision boundary

are calculated according to Theorem 2. Figure 2 shows the asymptotic variabilities as a function of the Mahalanobis distance between the two normal distributions, $|\mu_f - \mu_g|$ in this case. Also, it depicts the asymptotic variance of the estimated classification boundary value $(-\hat{\beta}_0/\hat{\beta}_1)$ by using the delta method. Although the Mahalanobis distance roughly in the range of 1 to 4 would be of practical interest, the plots show a notable trend in the asymptotic variances as the distance varies. When the two classes get very close, the variances shoot up due to the difficulty in discriminating them. On the other hand, as the Mahalanobis distance increases, that is, the two classes become more separated, the variances become increasingly large. A possible explanation for the trend is that the intercept and the slope of the optimal hyperplane are determined by only a small fraction of data falling into the margins in this case.

4. Simulation Studies

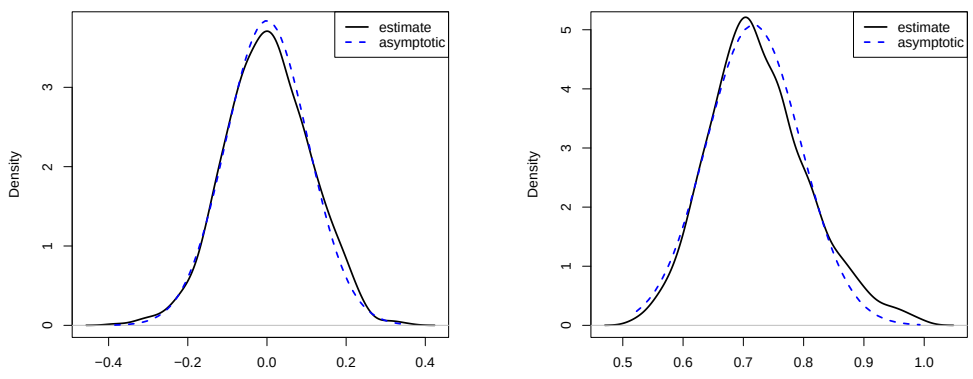
In this section, simulations are carried out to illustrate the asymptotic results and their potential for feature selection.

4.1 Bivariate Case

Theorem 2 is numerically illustrated with the multivariate normal setting in the previous section. Consider a bivariate case with mean vectors $\mu_f = (1,1)^\top$ and $\mu_g = (-1,-1)^\top$ and a common covariance matrix $\Sigma = I_2$. This example has $d_\Sigma(\mu_f, \mu_g) = 2\sqrt{2}$ and the corresponding Bayes error rate is 0.07865. Data were generated from the two normal distributions with an equal probability for each class. The total sample size was varied from 100 to 500. To see the direct effect of the hinge loss on the SVM coefficients without regularization as in the way the asymptotic properties in Section 2 are characterized ultimately, we estimated the coefficients of the linear SVM without the penalty term by linear programming. Such a simulation was repeated 1,000 times for each sample size, and Table 1 summarizes the results by showing the averages of the estimated coefficients of the SVM over 1,000 replicates. As expected, the averages get closer to the theoretically optimal coefficients β^* as the sample size grows. Moreover, the sampling distributions of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ approximate their theoretical counterparts for a large sample size as shown in Figure 3. The solid lines are the estimated density functions of $\hat{\beta}_0$ and $\hat{\beta}_1$ for $n = 500$, and the dotted lines are the corresponding asymptotic normal densities in Theorem 2.

Coefficients	Sample size n			Optimal values
	100	200	500	
β_0	0.0006	-0.0013	0.0022	0
β_1	0.7709	0.7450	0.7254	0.7169
β_2	0.7749	0.7459	0.7283	0.7169

Table 1: Averages of estimated and optimal coefficients over 1,000 replicates.



(a) (b)

Figure 3: Estimated sampling distributions of (a) $\hat{\beta}_0$ and (b) $\hat{\beta}_1$ with the asymptotic normal densities overlaid.

4.2 Feature Selection

Clearly the results we have established have implications to statistical inferences on the SVM. Among others, feature selection is of particular interest. By using the asymptotic variability of estimated coefficients, one can derive a new feature selection criterion based on the standardized coefficients. Such a criterion will take inherent statistical variability into account. More generally, this consideration of new criteria opens the possibility of casting feature selection for the SVM formally in the framework of hypothesis testing and extending standard variable selection procedures in regression to classification.

We investigate the possibility of using the standardized coefficients of $\hat{\beta}$ for selection of variables. For practical applications, one needs to construct a reasonable nonparametric estimator of the asymptotic variance-covariance matrix, whose entries are defined through line integrals. A similar technical issue arises in quantile regression. See Koenker (2005) for some suggested variance estimators in the setting.

For the sake of simplicity in the second set of simulation, we used the theoretical asymptotic variance in standardizing $\hat{\beta}$ and selected those variables with the absolute standardized coefficient exceeding a certain critical value. And we mainly monitored the type I error rate of falsely declaring the significance of a variable when it is not, over various settings of a mixture of two multivariate normal distributions. Different combinations of the sample size (n) and the number of variables (d) were tried. For a fixed even d, we set $\mu_f = (\mathbf{1}_{d/2}, \mathbf{0}_{d/2})^\top$, $\mu_g = \mathbf{0}_d^\top$, and $\Sigma = I_d$, where $\mathbf{1}_p$ and $\mathbf{0}_p$ indicate p-vectors of ones and zeros, respectively. Thus only the first half of the d variables have nonzero coefficients in the optimal hyperplane of the linear SVM. Table 2 shows the minima, median, and maxima of such type I error rates in selection of relevant variables over 200 replicates when the critical value was $z_{0.025} \approx 1.96$ (5% level of significance). If the asymptotic distributions were accurate, the

error rates would be close to the nominal level of 0.05. On the whole, the table suggests that when d is small, the error rates are very close to the nominal level even for small sample sizes, while for a large d, n has to be quite large for the asymptotic distributions to be valid. This pattern is clearly seen in Figure 4, which displays the median values of the type I error rates. In passing, we note that changing the proportion of relevant variables did not seem to affect the error rates, which are not shown here.

	Number of variables (d)			
n	6	12	18	24
250	[0.050, 0.060, 0.090]	[0.075, 0.108, 0.145]	[0.250, 0.295, 0.330]	[0.665, 0.698, 0.720]
500	[0.045, 0.080, 0.090]	[0.040, 0.068, 0.095]	[0.105, 0.130, 0.175]	[0.275, 0.293, 0.335]
750	[0.030, 0.055, 0.070]	[0.035, 0.065, 0.090]	[0.055, 0.095, 0.115]	[0.135, 0.185, 0.205]
1000	[0.050 ,0.065, 0.065]	[0.060, 0.068, 0.095]	[0.040, 0.075, 0.095]	[0.105, 0.135, 0.175]
1250	[0.065, 0.065, 0.070]	[0.035, 0.045, 0.050]	[0.055, 0.080, 0.105]	[0.070, 0.095, 0.125]
1500	[0.035, 0.050, 0.065]	[0.040, 0.058, 0.085]	[0.055, 0.075, 0.090]	[0.050, 0.095, 0.135]
1750	[0.030, 0.035, 0.060]	[0.035, 0.045, 0.075]	[0.040, 0.065, 0.095]	[0.055, 0.080, 0.120]
2000	[0.035, 0.040, 0.060]	[0.040, 0.065, 0.080]	[0.060, 0.070, 0.100]	[0.055, 0.075, 0.105]

Table 2: The minimum, median, and maximum values of the type I error rates of falsely flagging an irrelevant variable as relevant over 200 replicates by using the standardized SVM coefficients at 5% significance level.

We leave further development of asymptotic variance estimators for feature selection and comparison with risk based approaches such as the recursive feature elimination procedure as a future work.

5. Proofs

5.1 Technical Lemmas

Lemma 1 shows that there is a finite minimizer of $L(\beta)$, which is useful in proving the uniqueness of the minimizer in Lemma 6. In fact, the existence of the first moment of X is sufficient for Lemmas 1, 2, and 4. However, (A1) is needed for the existence and continuity of $H(\beta)$ in the proof of other lemmas and theorems.

Lemma 1 *Suppose that (A1) and (A2) are satisfied. Then* $L(\beta) \to \infty$ *as* $\|\beta\| \to \infty$ *and the existence of* β^* *is guaranteed.*

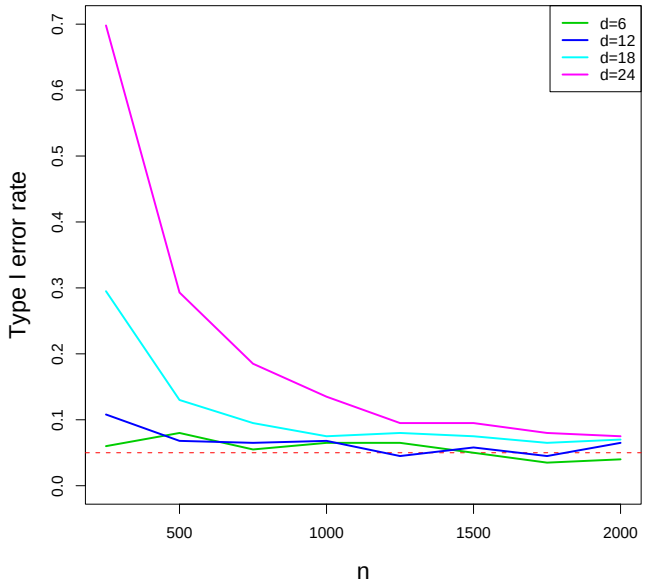


Figure 4: The median values of the type I error rates in variable selection depending on the sample size n and the number of variables d. The dotted line indicates the nominal level of 0.05.

Proof. Without loss of generality, we may assume that $x_0 = 0$ in (A2) and $B(0, \delta_0) \subset \mathcal{X}$. For any $\varepsilon > 0$,

$$\begin{aligned}
L(\beta) &= \pi_+ \int_{\mathcal{X}} [1 - \widetilde{x}^\top \beta]_+ f(x)dx + \pi_- \int_{\mathcal{X}} [1 + h(x;\beta)]_+ g(x)dx \\
&\geq \pi_+ \int_{\mathcal{X}} \{h(x;\beta) \leq 0\}(1 - h(x;\beta))f(x)dx + \pi_- \int_{\mathcal{X}} \{h(x;\beta) \geq 0\}(1 + h(x;\beta))g(x)dx \\
&\geq \pi_+ \int_{\mathcal{X}} \{h(x;\beta) \leq 0\}(-h(x;\beta))f(x)dx + \pi_- \int_{\mathcal{X}} \{h(x;\beta) \geq 0\}h(x;\beta)g(x)dx \\
&\geq \int_{\mathcal{X}} |h(x;\beta)| \min\left(\pi_+ f(x), \pi_- g(x)\right) dx \\
&\geq C_1 \min(\pi_+, \pi_-) \int_{B(0,\delta_0)} |h(x;\beta)|dx \\
&= C_1 \min(\pi_+, \pi_-)\|\beta\| \int_{B(0,\delta_0)} |h(x;w)|dx \\
&\geq C_1 \min(\pi_+, \pi_-)\|\beta\| \mathrm{vol}\left(\{|h(x;w)| \geq \varepsilon\} \cap B(0,\delta_0)\right) \varepsilon,
\end{aligned}$$

where $w = \beta/\|\beta\|$ and $\mathrm{vol}(A)$ denotes the volume of a set A.

Note that $-1 \le w_0 \le 1$. For $0 \le w_0 < 1$ and $0 < \varepsilon < 1$,

$$\begin{aligned} & \mathrm{vol}\left(\{|h(x;w)| \ge \varepsilon\} \cap B(0,\delta_0)\right) \\ \ge\ & \mathrm{vol}\left(\{h(x;w) \ge \varepsilon\} \cap B(0,\delta_0)\right) \\ =\ & \mathrm{vol}\left(\left\{x^\top w_+/\sqrt{1-w_0^2} \ge (\varepsilon - w_0)/\sqrt{1-w_0^2}\right\} \cap B(0,\delta_0)\right) \\ \ge\ & \mathrm{vol}\left(\left\{x^\top w_+/\sqrt{1-w_0^2} \ge \varepsilon\right\} \cap B(0,\delta_0)\right) \equiv V(\delta_0,\varepsilon) \end{aligned}$$

since $(\varepsilon - w_0)/\sqrt{1-w_0^2} \le \varepsilon$. When $-1 < w_0 < 0$, we obtain

$$\mathrm{vol}_B(h(x;w) \le -\varepsilon) \ge V(\delta_0,\varepsilon)$$

in a similar way. Note that $V(\delta_0,\varepsilon)$ is independent of β and $V(\delta_0,\varepsilon) > 0$ for some $\varepsilon < \delta_0$. Consequently, $L(\beta) \ge C_1 \min(\pi_+,\pi_-)\|\beta\| V(\delta_0,\varepsilon)\varepsilon \to \infty$ as $\|\beta\| \to \infty$. The case $w_0 = \pm 1$ is trivial.

Since the hinge loss is convex, $L(\beta)$ is convex in β. Since $L(\beta) \to \infty$ as $\|\beta\| \to \infty$, the set, denoted by $\mathcal{M}$, of minimizers of $L(\beta)$ forms a bounded connected set. The existence of the solution β^* of $L(\beta)$ easily follows from this. ■

In Lemmas 2 and 3, we obtain explicit forms of $S(\beta)$ and $H(\beta)$ for non-constant decision functions.

Lemma 2 *Assume that (A1) is satisfied. If $\beta_+ \ne 0$, then we have*

$$\frac{\partial L(\beta)}{\partial \beta_j} = S(\beta)_j$$

for $0 \le j \le d$.

Proof. It suffices to show that

$$\frac{\partial}{\partial \beta_j} \int_{\mathcal{X}} [1 - h(x;\beta)]_+ f(x)dx = -\int_{\mathcal{X}} \{h(x;\beta) \le 1\}\widetilde{x}_j f(x)dx.$$

Define $\Delta(t) = [1 - h(x;\beta) - t\widetilde{x}_j]_+ - [1 - h(x;\beta)]_+$. Let $t > 0$.

First, consider the case $\widetilde{x}_j > 0$. Then,

$$\Delta(t) = \begin{cases} 0 & \text{if} \quad h(x;\beta) > 1 \\ h(x;\beta) - 1 & \text{if} \quad 1 - t\widetilde{x}_j < h(x;\beta) \le 1 \\ -t\widetilde{x}_j & \text{if} \quad h(x;\beta) \le 1 - t\widetilde{x}_j. \end{cases}$$

Observe that

$$\begin{aligned} \int_{\mathcal{X}} \Delta(t)\{\widetilde{x}_j > 0\} f(x)dx \ =\ & \int_{\mathcal{X}} \{1 - t\widetilde{x}_j < h(x;\beta) \le 1\}(h(x;\beta) - 1) f(x)dx \\ & -t\int_{\mathcal{X}} \{h(x;\beta) \le 1 - t\widetilde{x}_j, \widetilde{x}_j > 0\}\widetilde{x}_j f(x)dx \end{aligned}$$

and that

$$\left|\frac{1}{t}\int_{\mathcal{X}}\{1-t\widetilde{x}_j < h(x;\beta) \leq 1\}(h(x;\beta)-1)f(x)dx\right| \leq \int_{\mathcal{X}}\{1-t\widetilde{x}_j < h(x;\beta) \leq 1\}\widetilde{x}_j f(x)dx.$$

By Dominated Convergence Theorem,

$$\lim_{t\downarrow 0}\int_{\mathcal{X}}\{1-t\widetilde{x}_j < h(x;\beta) \leq 1\}\widetilde{x}_j f(x)dx = \int_{\mathcal{X}}\{h(x;\beta)=1\}\widetilde{x}_j f(x)dx = 0$$

and

$$\lim_{t\downarrow 0}\int_{\mathcal{X}}\{h(x;\beta) \leq 1-t\widetilde{x}_j, \widetilde{x}_j > 0\}\widetilde{x}_j f(x)dx = \int_{\mathcal{X}}\{h(x;\beta) \leq 1, \widetilde{x}_j > 0\}\widetilde{x}_j f(x)dx.$$

Hence

$$\lim_{t\downarrow 0}\frac{1}{t}\int_{\mathcal{X}}\Delta(t)\{\widetilde{x}_j > 0\}f(x)dx = -\int_{\mathcal{X}}\{h(x;\beta) \leq 1, \widetilde{x}_j > 0\}\widetilde{x}_j f(x)dx. \tag{9}$$

Now assume that $\widetilde{x}_j < 0$. Then,

$$\Delta(t) = \begin{cases} 0 & \text{if} \quad h(x;\beta) > 1-t\widetilde{x}_j \\ 1-h(x;\beta)-t\widetilde{x}_j & \text{if} \quad 1 < h(x;\beta) \leq 1-t\widetilde{x}_j \\ -t\widetilde{x}_j & \text{if} \quad h(x;\beta) \leq 1. \end{cases}$$

In a similar fashion, one can show that

$$\lim_{t\downarrow 0}\frac{1}{t}\int_{\mathcal{X}}\Delta(t)\{\widetilde{x}_j < 0\}f(x)dx = -\int_{\mathcal{X}}\{h(x;\beta) \leq 1, \widetilde{x}_j < 0\}\widetilde{x}_j f(x)dx. \tag{10}$$

Combining (9) and (10), we have shown that

$$\lim_{t\downarrow 0}\frac{1}{t}\int_{\mathcal{X}}\Delta(t)f(x)dx = -\int_{\mathcal{X}}\{h(x;\beta) \leq 1\}\widetilde{x}_j f(x)dx.$$

The proof for the case $t < 0$ is similar. ■

The proof of Lemma 3 is based on the following identity

$$\int \delta(Dt+E)T(t)dt = \frac{1}{|D|}T(-E/D) \tag{11}$$

for constants D and E. This identity follows from the fact that $\delta(at) = \delta(t)/|a|$ and $\int \delta(t-a)T(t)dt = T(a)$ for a constant a.

Lemma 3 *Suppose that (A1) is satisfied. Under the condition that $\beta_+ \neq 0$, we have*

$$\frac{\partial^2 L(\beta)}{\partial\beta_j\partial\beta_k} = H(\beta)_{jk},$$

for $0 \leq j,k \leq d$.

Proof. Define

$$\Psi(\beta) = \int_{\mathcal{X}} \{x^\top \beta_+ < 1 - \beta_0\} s(x) dx$$

for a continuous and integrable function s defined on $\mathcal{X}$. Without loss of generality, we may assume that $\beta_1 \neq 0$. It is sufficient to show that for $0 \leq j, k \leq d$,

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k} \int_{\mathcal{X}} [1 - h(x;\beta)]_+ f(x) dx = \int_{\mathcal{X}} \delta(1 - h(x;\beta)) \widetilde{x}_j \widetilde{x}_k f(x) dx.$$

Define $\mathcal{X}_{-j} = \{(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d) \ : \ (x_1, \ldots, x_d) \in \mathcal{X}\}$ and $\mathcal{X}_j = \{x_j \ : (x_1, \ldots, x_d) \in \mathcal{X}\}$. Observe that

$$\frac{\partial \Psi(\beta)}{\partial \beta_0} = -\frac{1}{|\beta_1|} \int_{\mathcal{X}_{-1}} s\left(\frac{1 - h(x;\beta) + \beta_1 x_1}{\beta_1}, x_2, \ldots, x_d\right) dx_{-1} \tag{12}$$

and that for $k \neq 1$,

$$\frac{\partial \Psi(\beta)}{\partial \beta_k} = -\frac{1}{|\beta_1|} \int_{\mathcal{X}_{-1}} x_k s\left(\frac{1 - h(x;\beta) + \beta_1 x_1}{\beta_1}, x_2, \ldots, x_d\right) dx_{-1}. \tag{13}$$

If $\beta_p = 0$ for any $p \neq 1$, then

$$\frac{\partial \Psi(\beta)}{\partial \beta_1} = -\frac{1 - \beta_0}{\beta_1 |\beta_1|} \int_{\mathcal{X}_{-1}} s\left(\frac{1 - \beta_0}{\beta_1}, x_2, \ldots, x_d\right) dx_{-1}. \tag{14}$$

If there is $p \neq 1$ with $\beta_p \neq 0$, then we have

$$\frac{\partial \Psi(\beta)}{\partial \beta_1} = -\frac{1}{|\beta_p|} \int_{\mathcal{X}_{-p}} x_1 s\left(x_1, \ldots, x_{p-1}, \frac{1 - h(x;\beta) + \beta_p x_p}{\beta_p}, x_{p+1}, \ldots, x_d\right) dx_{-p}. \tag{15}$$

Choose $D = \beta_1$, $E = \widetilde{x}^\top \beta - \beta_1 x_1 - 1$, $t = x_1$ in (11). It follows from (13) and (15) that

$$\begin{aligned} \frac{\partial \Psi(\beta)}{\partial \beta_k} &= -\frac{1}{|\beta_1|} \int_{\mathcal{X}_{-1}} x_k s\left(\frac{1 - h(x;\beta) + \beta_1 x_1}{\beta_1}, x_2, \ldots, x_d\right) dx_{-1} \\ &= -\int_{\mathcal{X}_{-1}} \int_{\mathcal{X}_1} x_k s(x) \delta(h(x;\beta) - 1) dx_1 dx_{-1} \\ &= -\int_{\mathcal{X}} \delta(1 - h(x;\beta)) x_k s(x) dx. \end{aligned} \tag{16}$$

Similarly, we have

$$\frac{\partial \Psi(\beta)}{\partial \beta_0} = -\int_{\mathcal{X}} \delta(1 - h(x;\beta)) s(x) dx \tag{17}$$

by (12).

Choosing $D = \beta_1$, $E = \beta_0 - 1$, $t = x_1$ in (11), we have

$$\int_{\mathcal{X}_1} \delta(\beta_1 x_1 - 1 + \beta_0) x_1 s(x) dx_1 = \frac{1}{|\beta_1|} \left(\frac{1 - \beta_0}{\beta_1}\right) s\left(\frac{1 - \beta_0}{\beta_1}, x_2, \ldots, x_d\right)$$

by (14). This implies (16) for $k = 1$. The desired result now follows from (16) and (17). ■

The following lemma asserts that the optimal decision function is not a constant under the condition that the centers of two classes are separated.

Lemma 4 *Suppose that (A1) is satisfied. Then (A3) implies that* $\beta_+^* \neq 0$.

Proof. Suppose that

$$\int_{\mathcal{X}} \{x_{i^*} \geq G_{i^*}^-\} x_{i^*} g(x) dx < \int_{\mathcal{X}} \{x_{i^*} \leq F_{i^*}^+\} x_{i^*} f(x) dx \tag{18}$$

in (A3). We will show that

$$\min_{\beta_0} L(\beta_0, 0, \ldots 0) > \min_{\beta_0, \beta_{i^*} > 0} L(\beta_0, 0, \ldots, 0, \beta_{i^*}, 0, \ldots, 0), \tag{19}$$

implying that $\beta_+^* \neq 0$. Henceforth, we will suppress β's that are equal to zero in $L(\beta)$. The population minimizer $(\beta_0^*, \beta_{i^*}^*)$ is given by the minimizer of

$$L(\beta_0, \beta_{i^*}) \;=\; \pi_+ \int_{\mathcal{X}} [1 - \beta_0 - \beta_{i^*} x_{i^*}]_+ f(x) dx + \pi_- \int_{\mathcal{X}} [1 + \beta_0 + \beta_{i^*} x_{i^*}]_+ g(x) dx.$$

First, consider the case $\beta_{i^*} = 0$.

$$L(\beta_0) \;=\; \begin{cases} \pi_-(1 + \beta_0), & \beta_0 > 1 \\ 1 + (\pi_- - \pi_+)\beta_0, & -1 \leq \beta_0 \leq 1 \\ \pi_+(1 - \beta_0), & \beta_0 < -1 \end{cases}$$

with its minimum

$$\min_{\beta_0} L(\beta_0) = 2 \min(\pi_+, \pi_-). \tag{20}$$

Now consider the case $\beta_{i^*} > 0$, where

$$\begin{aligned} L(\beta_0, \beta_{i^*}) \;=\; & \pi_+ \int_{\mathcal{X}} \left\{ x_{i^*} \leq \frac{1 - \beta_0}{\beta_{i^*}} \right\} (1 - \beta_0 - \beta_{i^*} x_{i^*}) f(x) dx \\ & + \pi_- \int_{\mathcal{X}} \left\{ x_{i^*} \geq \frac{-1 - \beta_0}{\beta_{i^*}} \right\} (1 + \beta_0 + \beta_{i^*} x_{i^*}) g(x) dx. \end{aligned}$$

Let $\widetilde{\beta}_0$ denote the minimizer of $L(\beta_0, \beta_{i^*})$ for a given β_{i^*}. Note that $\partial L(\beta_0, \beta_{i^*})/\partial \beta_0$ is given as

$$\begin{aligned} & \frac{\partial L(\beta_0, \beta_{i^*})}{\partial \beta_0} \\ = \; & -\pi_+ \int_{\mathcal{X}} \left\{ x_{i^*} \leq \frac{1 - \beta_0}{\beta_{i^*}} \right\} f(x) dx + \pi_- \int_{\mathcal{X}} \left\{ x_{i^*} \geq \frac{-1 - \beta_0}{\beta_{i^*}} \right\} g(x) dx, \end{aligned} \tag{21}$$

which is monotonic increasing in β_0 with $\lim_{\beta_0 \to -\infty} \frac{\partial L(\beta_0, \beta_{i^*})}{\partial \beta_0} \to -\pi_+$ and $\lim_{\beta_0 \to \infty} \frac{\partial L(\beta_0, \beta_{i^*})}{\partial \beta_0}$ $\to \pi_-$. Hence $\widetilde{\beta}_0$ exists for a given β_{i^*}.

When $\pi_- < \pi_+$, we can easily check that $F_{i^*}^+ < \infty$ and $G_{i^*}^- = -\infty$. $F_{i^*}^+$ and $G_{i^*}^-$ may not be determined uniquely, meaning that there may exist an interval with probability zero. There is no significant change in the proof under the assumption that $F_{i^*}^+$ and $G_{i^*}^-$ are unique. Note that

$$\frac{1 - \widetilde{\beta}_0}{\beta_{i^*}} \leq F_{i^*}^+$$

by definition of $F_{i^*}^+$ and (21). Then,

$$\frac{-1-\widetilde{\beta}_0}{\beta_{i^*}} \leq F_{i^*}^+ - \frac{2}{\beta_{i^*}} \to -\infty \quad \text{as} \quad \beta_{i^*} \to 0,$$

and

$$\frac{1-\widetilde{\beta}_0}{\beta_{i^*}} \to F_{i^*}^+ \quad \text{as} \quad \beta_{i^*} \to 0.$$

¿From (18),

$$\pi_- \int_{\mathcal{X}} x_{i^*} g(x)dx < \pi_+ \int_{\mathcal{X}} \{x_{i^*} \leq F_{i^*}^+\} x_{i^*} f(x)dx. \tag{22}$$

Now consider the minimum of $L(\widetilde{\beta}_0, \beta_{i^*})$ with respect to $\beta_{i^*} > 0$. From (21),

$$\begin{aligned} & L(\widetilde{\beta}_0, \beta_{i^*}) \qquad (23) \\ = \; & \pi_+ \int_{\mathcal{X}} \left\{ x_{i^*} \leq \frac{1-\widetilde{\beta}_0}{\beta_{i^*}} \right\} (1 - \widetilde{\beta}_0 - \beta_{i^*} x_{i^*}) f(x)dx \\ & + \pi_- \int_{\mathcal{X}} \left\{ x_{i^*} \geq \frac{-1-\widetilde{\beta}_0}{\beta_{i^*}} \right\} (1 + \widetilde{\beta}_0 + \beta_{i^*} x_{i^*}) g(x)dx \\ = \; & 2\pi_- \int_{\mathcal{X}} \left\{ x_{i^*} \geq \frac{-1-\widetilde{\beta}_0}{\beta_{i^*}} \right\} g(x)dx \\ & + \beta_{i^*} \left(\pi_- \int_{\mathcal{X}} \left\{ x_{i^*} \geq \frac{-1-\widetilde{\beta}_0}{\beta_{i^*}} \right\} x_{i^*} g(x)dx - \pi_+ \int_{\mathcal{X}} \left\{ x_{i^*} \leq \frac{1-\widetilde{\beta}_0}{\beta_{i^*}} \right\} x_{i^*} f(x)dx \right) \end{aligned}$$

By (22), it can be easily seen that the second term in (23) is negative for sufficiently small $\beta_{i^*} > 0$, implying

$$L(\widetilde{\beta}_0, \beta_{i^*}) < 2\pi_- \quad \text{for some} \quad \beta_{i^*} > 0.$$

If $\pi_- > \pi_+$, then $F_{i^*}^+ = \infty$ and $G_{i^*}^- > -\infty$. Similarly, it can be checked that

$$L(\widetilde{\beta}_0, \beta_{i^*}) < 2\pi_+ \quad \text{for some} \quad \beta_{i^*} > 0.$$

Suppose that $\pi_- = \pi_+$. Then it can be verified that

$$\frac{1-\widetilde{\beta}_0}{\beta_{i^*}} \to \infty \quad \text{as} \quad \beta_{i^*} \to 0,$$

and

$$\frac{-1-\widetilde{\beta}_0}{\beta_{i^*}} \to -\infty \quad \text{as} \quad \beta_{i^*} \to 0.$$

In this case, $L(\widetilde{\beta}_0, \beta_{i^*}) < 1$.

Hence, under (18), we have shown that

$$L(\widetilde{\beta}_0, \beta_{i^*}) < 1 \quad \text{for some} \quad \beta_{i^*} > 0.$$

This, together with (20), implies (19). For the second case in (A3), the same arguments hold with $\beta_{i^*} < 0$. ■

Note that (A1) implies that H is well-defined and continuous in its argument. (A4) ensures that $H(\beta)$ is positive definite around β^* and thus we have a lower bound result in Lemma 5.

Lemma 5 *Under (A1), (A3) and (A4),*

$$\beta^\top H(\beta^*)\beta \geq C_4\|\beta\|^2,$$

where C_4 may depend on β^*.

Proof. Since the proof for the case $d = 1$ is trivial, we consider the case $d \geq 2$ only. Observe that

$$\begin{aligned}
\beta^\top H(\beta^*)\beta &= \mathbb{E}\Big(\delta(1 - Yh(X;\beta^*))h^2(X;\beta)\Big) \\
&= \pi_+ \int_{\mathcal{X}} \delta(1 - h(x;\beta^*))h^2(x;\beta)f(x)dx + \pi_- \int_{\mathcal{X}} \delta(1 + h(x;\beta^*))h^2(x;\beta)g(x)dx \\
&= \pi_+(\mathcal{R}h^2 f)(1 - \beta_0^*, \beta_+^*) + \pi_-(\mathcal{R}h^2 g)(1 + \beta_0^*, -\beta_+^*) \\
&= \pi_+(\mathcal{R}h^2 f)(1 - \beta_0^*, \beta_+^*) + \pi_-(\mathcal{R}h^2 g)(-1 - \beta_0^*, \beta_+^*).
\end{aligned}$$

The last equality follows from the homogeneity of Radon transform.

Recall that $\beta_+^* \neq 0$ by Lemma 4. Without loss of generality, we assume that $j^* = 1$ in (A4). Let $A_1\beta_+ = a = (a_1, \ldots, a_d)$ and $z = (A_1 x)/\|\beta_+^*\|$. Given $u = (u_2, \ldots, u_d)$, let $u^* = \Big((1 - \beta_0^*)/\|\beta_+^*\|, u\Big)$. Define $\mathcal{Z} = \{z = (A_1 x)/\|\beta_+^*\| : x \in \mathcal{X}\}$ and $\mathcal{U} = \{u : u_j = \|\beta_+^*\|z_j \text{ for } j = 2, \ldots, d, \text{ and } z \in \mathcal{Z}\}$. Note that $\det A_1 = 1$, $du = \|\beta_+^*\|^{d-1} dz_2 \ldots dz_d$,

$$\|\beta_+^*\| A_1^\top z\Big|_{z_1 = (1-\beta_0^*)/\|\beta_+^*\|^2} = A_1^\top \begin{pmatrix} (1 - \beta_0^*)/\|\beta_+^*\| \\ \|\beta_+^*\|z_2 \\ \vdots \\ \|\beta_+^*\|z_d \end{pmatrix} = A_1^\top u^*$$

and

$$\begin{aligned}
\beta_0 + \|\beta_+^*\| a^\top z\Big|_{z_1 = (1-\beta_0^*)/\|\beta_+^*\|^2} &= \beta_0 + \|\beta_+^*\|\Big(a_1(1 - \beta_0^*)/\|\beta_+^*\|^2 + \sum_{j=2}^d a_j z_j\Big) \\
&= \beta_0 + a_1(1 - \beta_0^*)/\|\beta_+^*\| + \|\beta_+^*\| \sum_{j=2}^d a_j z_j.
\end{aligned}$$

Using the transformation A_1, we have

$$\begin{aligned}
&(\mathcal{R}h^2 f)(1-\beta_0^*,\beta_+^*)\\
=&\int_{\mathcal{Z}}\delta\left(1-\beta_0^*-\|\beta_+^*\|(A_1\beta_+^*)^\top z\right)h^2\left(\|\beta_+^*\|A_1^\top z;\beta\right)f\left(\|\beta_+^*\|A_1^\top z\right)\|\beta_+^*\|^d dz\\
=&\int_{\mathcal{Z}}\delta\left(1-\beta_0^*-\|\beta_+^*\|^2 e_1^\top z\right)\left(\beta_0+\|\beta_+^*\|(A_1\beta_+)^\top z\right)^2 f\left(\|\beta_+^*\|A_1^\top z\right)\|\beta_+^*\|^d dz\\
=&\int_{\mathcal{Z}}\delta\left(1-\beta_0^*-\|\beta_+^*\|^2 z_1\right)\left(\beta_0+\|\beta_+^*\|a^\top z\right)^2 f\left(\|\beta_+^*\|A_1^\top z\right)\|\beta_+^*\|^d dz\\
=&\frac{1}{\|\beta_+^*\|}\int_{\mathcal{U}}\Big(\beta_0+a_1(1-\beta_0^*)/\|\beta_+^*\|+\sum_{j=2}^d a_j u_j\Big)^2 f(A_1^\top u^*)du.
\end{aligned}$$

The last equality follows from the identity (11). Let $\mathcal{D}_*^+=\left\{u: A_1^\top u^*\in\mathcal{D}^+\right\}$. By (A4), there exists a constant $C_2>0$ and a rectangle $\mathcal{D}_*^+$ on which $f(A_1^\top u^*)\geq C_2$ for $u\in\mathcal{D}_*^+$. Then

$$\begin{aligned}
&(\mathcal{R}h^2 f)(1-\beta_0^*,\beta_+^*)\\
\geq&\frac{1}{\|\beta_+^*\|}\int_{\mathcal{D}_*^+}\Big(\beta_0+a_1(1-\beta_0^*)/\|\beta_+^*\|+\sum_{j=2}^d a_j u_j\Big)^2 f(A_1^\top u^*)du\\
\geq&\frac{1}{\|\beta_+^*\|}\cdot C_2\int_{\mathcal{D}_*^+}\Big(\beta_0+a_1(1-\beta_0^*)/\|\beta_+^*\|+\sum_{j=2}^d a_j u_j\Big)^2 du\\
=&\frac{1}{\|\beta_+^*\|}\cdot C_2\cdot\mathrm{vol}(\mathcal{D}_*^+)\mathbb{E}^u\Big(\beta_0+a_1(1-\beta_0^*)/\|\beta_+^*\|+\sum_{j=2}^d a_j U_j\Big)^2\\
=&\frac{1}{\|\beta_+^*\|}\cdot C_2\cdot\mathrm{vol}(\mathcal{D}_*^+)\Big\{\Big(\beta_0+a_1(1-\beta_0^*)/\|\beta_+^*\|+\mathbb{E}^u\sum_{j=2}^d a_j U_j\Big)^2+\mathbb{V}^u(\sum_{j=2}^d a_j U_j)\Big\},
\end{aligned}$$

where U_j for $j=2,\ldots,d$ are independent and uniform random variables defined on $\mathcal{D}_*^+$, and $\mathbb{E}^u$ and $\mathbb{V}^u$ denote the expectation and variance with respect to the uniform distribution.

Letting $\bar{\mu}_i=(l_i+v_i)/2$, we have

$$\begin{aligned}
&(\mathcal{R}h^2 f)(1-\beta_0^*,\beta_+^*) \qquad (24)\\
\geq&\frac{1}{\|\beta_+^*\|}\cdot C_2\cdot\mathrm{vol}(\mathcal{D}_*^+)\Big\{\Big(\beta_0+a_1(1-\beta_0^*)/\|\beta_+^*\|+\sum_{j=2}^d a_j\bar{\mu}_j\Big)^2+\min_{2\leq j\leq d}\mathbb{V}^u(U_j)\sum_{j=2}^d a_j^2\Big\}.
\end{aligned}$$

Similarly, it can be shown that

$$\begin{aligned}
&(\mathcal{R}h^2 g)(-1-\beta_0^*,\beta_+^*) \qquad (25)\\
\geq&\frac{1}{\|\beta_+^*\|}\cdot C_3\cdot\mathrm{vol}(\mathcal{D}_*^-)\Big\{\Big(\beta_0-a_1(1+\beta_0^*)/\|\beta_+^*\|+\sum_{j=2}^d a_j\bar{\mu}_j\Big)^2+\min_{2\leq j\leq d}\mathbb{V}^u(U_j)\sum_{j=2}^d a_j^2\Big\},
\end{aligned}$$

where $\mathcal{D}_*^-=\Big\{u: A_1^\top\Big((-1-\beta_0^*)/\|\beta_+^*\|,u\Big)\in\mathcal{D}^-\Big\}$.

Combining (24)-(25) and letting $C_5 = 2/\|\beta_+^*\| \min\Big(\pi_+ C_2 \cdot \mathrm{vol}(\mathcal{D}_*^+), \pi_- C_3 \cdot \mathrm{vol}(\mathcal{D}_*^-)\Big)$ and $C_6 = \min\Big(1, \min_{2\le j\le d} \mathbb{V}^u(U_j)\Big)$, we have

$$
\begin{aligned}
& \beta^\top H(\beta^*)\beta \\
\ge\ & \frac{\pi_+}{\|\beta_+^*\|} \cdot C_2 \cdot \mathrm{vol}(\mathcal{D}_*^+)\Big\{\Big(\beta_0 + a_1(1-\beta_0^*)/\|\beta_+^*\| + \sum_{j=2}^d a_j\bar{\mu}_j\Big)^2 + \min_{2\le j\le d}\mathbb{V}^u(U_j)\sum_{j=2}^d a_j^2\Big\} \\
& + \frac{\pi_-}{\|\beta_+^*\|} \cdot C_3 \cdot \mathrm{vol}(\mathcal{D}_*^-)\Big\{\Big(\beta_0 - a_1(1+\beta_0^*)/\|\beta_+^*\| + \sum_{j=2}^d a_j\bar{\mu}_j\Big)^2 + \min_{2\le j\le d}\mathbb{V}^u(U_j)\sum_{j=2}^d a_j^2\Big\} \\
\ge\ & C_5C_6\Big\{\Big(\beta_0 + a_1(1-\beta_0^*)/\|\beta_+^*\| + \sum_{j=2}^d a_j\bar{\mu}_j\Big)^2 \\
& + \Big(\beta_0 - a_1(1+\beta_0^*)/\|\beta_+^*\| + \sum_{j=2}^d a_j\bar{\mu}_j\Big)^2 + 2\sum_{j=2}^d a_j^2\Big\}/2 \\
=\ & C_5C_6\Big\{\Big(\beta_0 + a_1(1-\beta_0^*)/\|\beta_+^*\|\Big)^2 + \Big(\beta_0 - a_1(1+\beta_0^*)/\|\beta_+^*\|\Big)^2 \\
& + 4\Big(\beta_0 - a_1\beta_0^*/\|\beta_+^*\|\Big)\sum_{j=2}^d a_j\bar{\mu}_j + 2\Big(\sum_{j=2}^d a_j\bar{\mu}_j\Big)^2 + 2\sum_{j=2}^d a_j^2\Big\}/2.
\end{aligned}
$$

Note that

$$
\begin{aligned}
& \Big(\sum_{j=2}^d a_j\bar{\mu}_j\Big)^2 + 2\Big(\beta_0 - a_1\beta_0^*/\|\beta_+^*\|\Big)\sum_{j=2}^d a_j\bar{\mu}_j \\
=\ & \Big(\sum_{j=2}^d a_j\bar{\mu}_j + \beta_0 - a_1\beta_0^*/\|\beta_+^*\|\Big)^2 - \Big(\beta_0 - a_1\beta_0^*/\|\beta_+^*\|\Big)^2
\end{aligned}
$$

and

$$
\Big(\beta_0 + a_1(1-\beta_0^*)/\|\beta_+^*\|\Big)^2 + \Big(\beta_0 - a_1(1+\beta_0^*)/\|\beta_+^*\|\Big)^2 - 2\Big(\beta_0 - a_1\beta_0^*/\|\beta_+^*\|\Big)^2 = 2a_1^2/\|\beta_+^*\|^2.
$$

Thus, the lower bound of $\beta^\top H(\beta^*)\beta$ except for the constant C_5C_6 allows the following quadratic form in terms of $\beta_0, a_1, \ldots, a_d$. Let

$$
Q(\beta_0, a_1, \ldots, a_d) = a_1^2/\|\beta_+^*\|^2 + \Big(\sum_{j=2}^d a_j\bar{\mu}_j + \beta_0 - a_1\beta_0^*/\|\beta_+^*\|\Big)^2 + \sum_{j=2}^d a_j^2.
$$

Obviously $Q(\beta_0, a_1, \ldots, a_d) \ge 0$ and $Q(\beta_0, a_1, \ldots, a_d) = 0$ implies that $a_1 = \ldots = a_d = 0$ and $\beta_0 = 0$. Therefore Q is positive definite. Letting $\nu_1 > 0$ be the smallest eigenvalue of the matrix corresponding to Q, we have proved

$$
\beta^\top H(\beta^*)\beta \quad \ge \quad C_5C_6\nu_1(\beta_0^2 + \sum_{j=1}^d a_j^2) = C_5C_6\nu_1(\beta_0^2 + \sum_{j=1}^d \beta_j^2).
$$

The last equality follows from the fact that $A_1\beta_+ = a$ and the transformation A_1 preserves the norm. With the choice of $C_4 = C_5C_6\nu_1 > 0$, the result follows. ■

Lemma 6 *Suppose that (A1)-(A4) are met. Then $L(\beta)$ has a unique minimizer.*

Proof. By Lemma 1, we may choose any minimizer $\beta^* \in \mathcal{M}$. By Lemma 4 and 5, $H(\beta)$ is positive definite at β^*. Then $L(\beta)$ is locally strictly convex at β^*, so that $L(\beta)$ has a local minimum at β^*. Hence the minimizer of $L(\beta)$ is unique. ■

5.2 Proof of Theorems 1 and 2

For fixed $\theta \in \mathbb{R}^{d+1}$, define

$$\Lambda_n(\theta) = n\Big(l_{\lambda,n}(\beta^* + \theta/\sqrt{n}) - l_{\lambda,n}(\beta^*)\Big)$$

and

$$\Gamma_n(\theta) = \mathbb{E}\Lambda_n(\theta).$$

Observe that

$$\Gamma_n(\theta) = n\left(L(\beta^* + \theta/\sqrt{n}) - L(\beta^*)\right) + \frac{\lambda}{2}\left(\|\theta_+\|^2 + 2\sqrt{n}{\beta^*_+}^\top\theta_+\right).$$

By Taylor series expansion of L around β^*, we have

$$\Gamma_n(\theta) = \frac{1}{2}\theta^\top H(\widetilde{\beta})\theta + \frac{\lambda}{2}\left(\|\theta_+\|^2 + 2\sqrt{n}{\beta^*_+}^\top\theta_+\right),$$

where $\widetilde{\beta} = \beta^* + (t/\sqrt{n})\theta$ for some $0 < t < 1$. Define $D_{jk}(\alpha) = H(\beta^* + \alpha)_{jk} - H(\beta^*)_{jk}$ for $0 \le j, k \le d$. Since $H(\beta)$ is continuous in β, there exists $\delta_1 > 0$ such that $|D_{jk}(\alpha)| < \varepsilon_1$ if $\|\alpha\| < \delta_1$ for any $\varepsilon_1 > 0$ and all $0 \le j, k \le d$. Then, as $n \to \infty$,

$$\frac{1}{2}\theta^\top H(\widetilde{\beta})\theta = \frac{1}{2}\theta^\top H(\beta^*)\theta + o(1).$$

It is because for sufficiently large n such that $\|(t/\sqrt{n})\theta\| < \delta_1$,

$$\begin{aligned}\left|\theta^\top\Big(H(\widetilde{\beta}) - H(\beta^*)\Big)\theta\right| &\le \sum_{j,k}|\theta_j||\theta_k|\left|D_{jk}\left(\frac{t}{\sqrt{n}}\theta\right)\right| \\ &\le \varepsilon_1\sum_{j,k}|\theta_j||\theta_k| \le 2\varepsilon_1\|\theta\|^2.\end{aligned}$$

Together with the assumption that $\lambda = o(n^{-1/2})$, we have

$$\Gamma_n(\theta) = \frac{1}{2}\theta^\top H(\beta^*)\theta + o(1).$$

Define $W_n = -\sum_{i=1}^n \zeta^i Y^i \widetilde{X}^i$ where $\zeta^i = \Big\{Y^i h(X^i;\beta^*) \le 1\Big\}$. Then $\frac{1}{\sqrt{n}}W_n$ follows asymptotically $N\Big(0, nG(\beta^*)\Big)$ by central limit theorem. Note that

$$\mathbb{E}\left(\zeta^i Y^i \widetilde{X}^i\right) = 0 \quad \text{and} \quad \mathbb{E}\left(\zeta^i Y^i \widetilde{X}^i (\zeta^i Y^i \widetilde{X}^i)^\top\right) = \mathbb{E}\left(\zeta^i \widetilde{X}^i (\widetilde{X}^i)^\top\right). \tag{26}$$

Recall that β^* is characterized by $S(\beta^*) = 0$ implying the first part of (26). If we define

$$R_{i,n}(\theta) = \Big[1 - Y^i h(X^i; \beta^* + \theta/\sqrt{n})\Big]_+ - \Big[1 - Y^i h(X^i; \beta^*)\Big]_+ + \zeta^i Y^i h(X^i; \theta/\sqrt{n}),$$

then we see that

$$\Lambda_n(\theta) = \Gamma_n(\theta) + W_n^\top \theta/\sqrt{n} + \sum_{i=1}^n \Big(R_{i,n}(\theta) - \mathbb{E}R_{i,n}(\theta)\Big)$$

and

$$\Big|R_{i,n}(\theta)\Big| \le \Big|h(X^i;\theta)/\sqrt{n}\Big| U\left(\Big|h(X^i;\theta)/\sqrt{n}\Big|\right), \tag{27}$$

where

$$U(t) = \Big\{\Big|1 - Y^i h(X^i;\beta^*)\Big| \le t\Big\} \qquad \text{for} \quad t \in \mathbb{R}.$$

To verify (27), let $\zeta = \{a \le 1\}$ and $R = [1-z]_+ - [1-a]_+ + \zeta(z-a)$. If $a > 1$, then $R = (1-z)\{z \le 1\}$; otherwise, $R = (z-1)\{z > 1\}$. Hence,

$$\begin{aligned} R &= (1-z)\{a>1, z\le 1\} + (z-1)\{a<1, z>1\} \\ &\le |z-a|\{a>1, z\le 1\} + |z-a|\{a<1, z>1\} \\ &= |z-a|\Big(\{a>1, z\le 1\} + \{a<1, z>1\}\Big) \\ &\le |z-a|\{|1-a| \le |z-a|\}. \end{aligned} \tag{28}$$

Choosing $z = Y^i h(X^i; \beta^* + \theta/\sqrt{n})$ and $a = Y^i h(X^i;\beta^*)$ in (28) yields (27).

Since cross-product terms in $\mathbb{E}(\sum_i (R_{i,n} - \mathbb{E}R_{i,n}))^2$ cancel out, we obtain from (27) that for each fixed θ,

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}\left(|R_{i,n}(\theta) - \mathbb{E}R_{i,n}(\theta)|^2\right) &\le \sum_{i=1}^n \mathbb{E}\left(R_{i,n}(\theta)^2\right) \\ &\le \sum_{i=1}^n \mathbb{E}\left(\big(h(X^i;\theta)/\sqrt{n}\big)^2 U\left(|h(X^i;\theta)/\sqrt{n}|\right)\right) \\ &\le \sum_{i=1}^n \mathbb{E}\left((1+\|X^i\|^2)\|\theta\|^2/n\; U\left(\sqrt{1+\|X^i\|^2}\|\theta\|/\sqrt{n}\right)\right) \\ &= \|\theta\|^2 \mathbb{E}\left((1+\|X\|^2)\; U\left(\sqrt{1+\|X\|^2}\|\theta\|/\sqrt{n}\right)\right). \end{aligned}$$

(A1) implies that $\mathbb{E}\left(\|X\|^2\right) < \infty$. Hence, for any $\varepsilon > 0$, choose C_7 such that $\mathbb{E}\Big((1 + \|X\|^2)\{\|X\| > C_7\}\Big) < \varepsilon/2$. Then

$$\begin{aligned} & \mathbb{E}\left((1+\|X\|^2)\; U\left(\sqrt{1+\|X\|^2}\|\theta\|/\sqrt{n}\right)\right) \\ \le\; & \mathbb{E}\Big((1+\|X\|^2)\{\|X\| > C_7\}\Big) + (1+C_7^2)\mathbb{P}\left(U\left(\sqrt{1+C_7^2}\|\theta\|/\sqrt{n}\right)\right) \end{aligned}$$

By (A1), the distribution of X is not degenerate, which in turn implies that $\lim_{t\downarrow 0}\mathbb{P}(U(t)) = 0$. We can take a large N such that $\mathbb{P}\left(U\left(\sqrt{1+C_7^2}\|\theta\|/\sqrt{n}\right)\right) < \varepsilon/(2(1+C_7^2))$ for $n \geq N$. This proves that

$$\sum_{i=1}^{n}\mathbb{E}\left(|R_{i,n}(\theta) - \mathbb{E}R_{i,n}(\theta)|^2\right) \to 0$$

as $n \to \infty$. Thus, for each fixed θ,

$$\Lambda_n(\theta) = \frac{1}{2}\theta^\top H(\beta^*)\theta + W_n^\top\theta/\sqrt{n} + o_{\mathbb{P}}(1).$$

Let $\eta_n = -H(\beta^*)^{-1}W_n/\sqrt{n}$. By Convexity Lemma in Pollard (1991), we have

$$\Lambda_n(\theta) = \frac{1}{2}(\theta-\eta_n)^\top H(\beta^*)(\theta-\eta_n) - \frac{1}{2}\eta_n^\top H(\beta^*)\eta_n + r_n(\theta),$$

where, for each compact set K in $\mathbb{R}^{d+1}$,

$$\sup_{\theta\in K}|r_n(\theta)| \to 0 \qquad \text{in probability.}$$

Because η_n converges in distribution, there exists a compact set K containing B_ε, where B_ε is a closed ball with center η_n and radius ε with probability arbitrarily close to one. Hence we have

$$\Delta_n = \sup_{\theta\in B_\varepsilon}|r_n(\theta)| \to 0 \qquad \text{in probability.} \tag{29}$$

For examination of the behavior of $\Lambda_n(\theta)$ outside B_ε, consider $\theta = \eta_n + \gamma v$, with $\gamma > \varepsilon$ and v, a unit vector and a boundary point $\theta^* = \eta_n + \varepsilon v$. By Lemma 5, convexity of Λ_n, and the definition of Δ_n, we have

$$\begin{aligned}
\frac{\varepsilon}{\gamma}\Lambda_n(\theta) + \left(1-\frac{\varepsilon}{\gamma}\right)\Lambda_n(\eta_n) &\geq \Lambda_n(\theta^*) \\
&\geq \frac{1}{2}(\theta^*-\eta_n)^\top H(\beta^*)(\theta^*-\eta_n) - \frac{1}{2}\eta_n^\top H(\beta^*)\eta_n - \Delta_n \\
&\geq \frac{C_4}{2}\varepsilon^2 + \Lambda_n(\eta_n) - 2\Delta_n,
\end{aligned}$$

implying that

$$\inf_{\|\theta-\eta_n\|>\varepsilon}\Lambda_n(\theta) \geq \Lambda_n(\eta_n) + \left(\frac{C_4}{2}\varepsilon^2 - 2\Delta_n\right).$$

By (29), we can take Δ_n so that $2\Delta_n < C_4\varepsilon^2/4$ with probability tending to one. So the minimum of Λ_n cannot occur at any θ with $\|\theta-\eta_n\| > \varepsilon$. Hence, for each $\varepsilon > 0$ and $\widehat{\theta}_{\lambda,n} = \sqrt{n}(\widehat{\beta}_{\lambda,n} - \beta^*)$,

$$\mathbb{P}\Big(\|\widehat{\theta}_{\lambda,n} - \eta_n\| > \varepsilon\Big) \to 0.$$

This completes the proof of Theorems 1 and 2. ■

6. Discussion

In this paper, we have investigated asymptotic properties of the coefficients of variables in the SVM solution for nonseparable linear classification. More specifically, we have established a Bahadur type representation of the coefficients and their asymptotic normality using Radon transformation of the second moments of the variables. The representation shows how the statistical behavior of the coefficients is determined by the margins of the optimal hyperplane and the underlying probability distribution. Shedding a new statistical light on the SVM, these results provide an insight into its asymptotic behavior and can be used to improve our statistical practice with the SVM in various aspects.

There are several issues yet to be investigated. The asymptotic results that we have obtained so far pertain only to the linear SVM in nonseparable cases. Although it may be of more theoretical consideration than practical, a similar analysis of the linear SVM in the separable case is anticipated, which will ultimately lead to a unified theory for separable as well as nonseparable cases. The separable case would require a slightly different treatment than the nonseparable case because the regularization parameter λ needs to remain positive in the limit to guarantee the uniqueness of the solution. An extension of the SVM asymptotics to the nonlinear case is another direction of interest. In this case, the minimizer defined by the SVM is not a vector of coefficients of a fixed length but a function in a reproducing kernel Hilbert space. So, the study of asymptotic properties of the minimizer in the function space essentially requires investigation of its pointwise behavior or its functionals in general as the sample size grows. A general theory in Shen (1997) on asymptotic normality and efficiency of substitution estimates for smooth functionals is relevant. In particular, Theorem 2 in Shen (1997) provides the asymptotic normality of the penalized sieve MLE, characterization of which bears a close resemblance with function estimation for the nonlinear case. However, the theory was developed under the assumption of the differentiability of the loss, and it has to be modified for proper theoretical analysis of the SVM. As in the approach for the linear case presented in this paper, one may get around the non-differentiability issue of the hinge loss by imposing appropriate regularity conditions to ensure that the minimizer is unique and the expected loss is differentiable and locally quadratic around the minimizer.

Consideration of these extensions will lead to a more complete picture of the asymptotic behavior of the SVM solution.

Acknowledgments

The authors are grateful to Wolodymyr Madych for personal communications on Radon transform. This research was supported by a Korea Research Foundation Grant funded by the Korean government (MOEHRD, Basic Research Promotion Fund) (KRF-2005-070-C00020).

References

R. R. Bahadur. A note on quantiles in large samples. *Annals of Mathematical Statistics*, 37:577–580, 1966.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statististical Association*, 101:138–156, 2006.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. Technical Report, 2004.

P. Chaudhuri. Nonparametric estimates of regression quantiles and their local Bahadur representation. *The Annals of Statistics*, 19:760–777, 1991.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

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

S. R. Deans. *The Radon Transform and Some of Its Applications*. Krieger Publishing Company, Florida, 1993.

I. Guyon and A. Elisseeff. Introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

A. B. Ishak and B. Ghattas. An efficient method for variable selection using svm-based criteria. Preprint, Institut de Mathématiques de Luminy, 2005.

R. Koenker. *Quantile Regression (Econometric Society Monographs)*. Cambridge University Press, 2005.

Y. Lin. Some asymptotic properties of the support vector machine. Technical report 1029, Department of Statistics, University of Wisconsin-Madison, 2000.

Y. Lin. A note on margin-based loss functions in classification. *Statististics and Probability Letters*, 68:73–82, 2002.

F. Natterer. *The Mathematics of Computerized Tomography*. Wiley & Sons, New York, 1986.

D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7:186–199, 1991.

A. G. Ramm and A. I. Katsevich. *The Radon Transform and Local Tomography*. CRC Press, Boca Raton, 1996.

B. Schölkopf and A. Smola. *Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

J. C. Scovel and I. Steinwart. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35:575–607, 2006.

X. Shen. On methods of sieves and penalization. *Annals of Statistics*, 25:2555–2591, 1997.

I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Transactions on Information Theory*, 51:128–142, 2005.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1996.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–84, 2004.