Submitted 11/16; Revised 3/17; Published 4/17

Asymptotic behavior of Support Vector Machine for spiked population model

Hanwen Huang HUANGHW@UGA.EDU
Department of Epidemiology and Biostatistics
University of Georgia
Athens, GA 30602, USA

Editor: Xiaotong Shen

Abstract

For spiked population model, we investigate the large dimension N and large sample size M asymptotic behavior of the Support Vector Machine (SVM) classification method in the limit of $N, M \to \infty$ at fixed $\alpha = M/N$. We focus on the generalization performance by analytically evaluating the angle between the normal direction vectors of SVM separating hyperplane and corresponding Bayes optimal separating hyperplane. This is an analogous result to the one shown in Paul (2007) and Nadler (2008) for the angle between the sample eigenvector and the population eigenvector in random matrix theorem. We provide not just bound, but sharp prediction of the asymptotic behavior of SVM that can be determined by a set of nonlinear equations. Based on the analytical results, we propose a new method of selecting tuning parameter which significantly reduces the computational cost. A surprising finding is that SVM achieves its best performance at small value of the tuning parameter under spiked population model. These results are confirmed to be correct by comparing with those of numerical simulations on finite-size systems. We also apply our formulas to an actual dataset of breast cancer and find agreement between analytical derivations and numerical computations based on cross validation.

Keywords: Asymptotic behavior, Spiked population model, Support Vector Machine

1. Introduction

The Support Vector Machine (SVM) is a state-of-the-art powerful classification method proposed by Vapnik (Vapnik, 1995). It has been widely used in bioinformatics and many other disciplines and has achieved a lot of success. Like other classification methods, SVM may suffer from a loss of generalization ability in high dimensional situations as shown by Figure 1 which displays the application of SVM to a high dimensional two class toy example with class labels $+1$ and -1. The data have dimension $N = 100$, with $M_+ = 45$ data vectors from Class $+1$ represented as circles, and $M_- = 45$ data vectors from Class -1 represented as plus. The two distributions are nearly standard normal except that the mean in the first dimension is shifted to $+\mu$ and $-\mu$ for Class $+1$ and Class -1 respectively. Here $\mu = 1$. Figure 1 shows the projections of the data onto the two-dimensional subspace determined by the first dimension (dashed line) and the normal vector (solid line) of the SVM separating hyperplane. The angle between these two directions can be used to determine the generalization ability of the classifier. A classifier who has good generalization properties should have small angle. For the particular example shown in

Figure 1, the angle is 56.6°. Therefore, projection of a new data vector onto the SVM direction cannot be expected to provide effective discrimination. As mentioned by Marron et al. (2007), the reason is that the estimated SVM classifier is driven only by very particular aspects of the realization of the training data at hand. New data will have their own quite different quirks, which will bear no relation to these.

Hall et al. (2005) studied the High Dimensional Low Sample Size (HDLSS) asymptotics of SVM and shown that for fixed sample size $M = M_+ + M_-$, as $N \to \infty$ the angle depends on the signal size μ which is defined as half of the distance between the means of two distributions for this example. Assume that μ increases with N as N^γ, then if $\gamma{>}1/2$, SVM is strongly consistent, i.e., the angle approaches to 0°; if $\gamma{<}1/2$, SVM is strongly inconsistent, i.e., the angle approaches to 90°; if $\gamma = 1/2$, the angle is between 0° and 90°. Therefore the signal size μ has to be large enough in order to gain some prediction power.

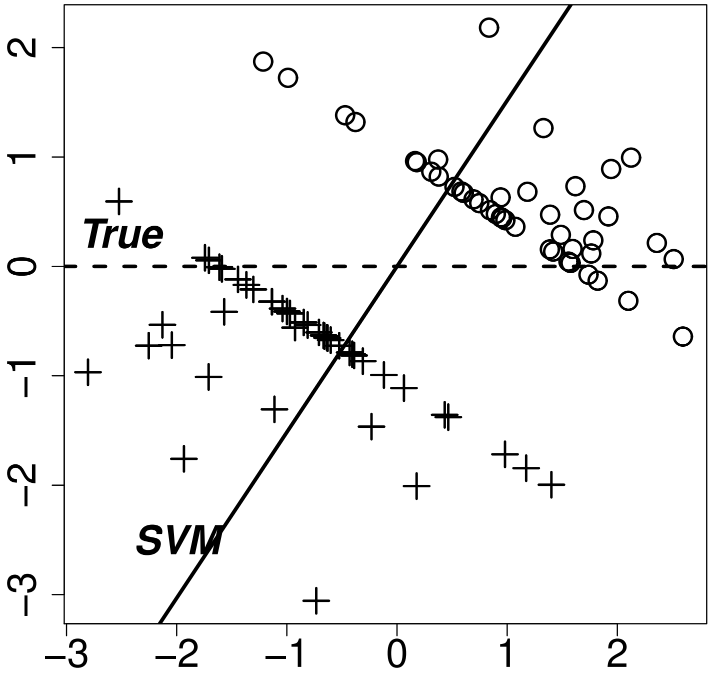

Figure 1: Toy examples, illustrating the performance of SVM on high dimensional data with $N = 100$ and sample size $M = M_+ + M_- = 90$. The circles denote the data from Class +1 and and the plus denote the data from Class −1. The dashed line represents the first dimension which is the true difference in the Gaussian means. The solid line represents the normal vector of SVM separating hyperplane. The angle between the solid and dashed lines is 56.6°.

Analogous conclusion has been drawn in the context of unsupervised learning for Principal Component Analysis (PCA). The study of sample covariance matrices is fundamental in multivariate analysis. It is well known that the sample covariance matrix is a consistent estimator of the population covariance matrix for fixed dimension N and sample size $M \to \infty$.

The PCA consistency in HDLSS context (fixed M and $N \to \infty$) was studied in Jung and Marron (2009); Jung et al. (2012) and it was shown that the asymptotic behavior of the Principal Component (PC) directions of sample covariance matrix depend on the size of the corresponding eigenvalues. Assume that the eigenvalue of the sample covariance matrix λ increases with N in the order of power γ, i.e. $\lambda \sim N^{\gamma}$. Then, if $\gamma > 1/2$, the corresponding estimated PC direction is strongly consistent, i.e. the angle between the estimated direction and its population counterpart is 0°; if $\gamma < 1/2$, the corresponding estimated PC direction is strongly inconsistent, i.e. the angle is 90°; if $\gamma = 1/2$, the angle is random and follows a certain distribution.

On the other hand, with the development of modern high-throughput technologies, it is not uncommon to have data where M is comparable in size to N, or substantially larger. There has been considerable effort to establish asymptotic results for sample eigenvalues and eigenvectors under the assumption that N and M grow at the same rate, that is, $M/N \to \alpha > 0$ (see review Bai (1999)). The limiting distribution of eigenvalues of the sample covariance matrix was derived in Marcenko and Pastur (1967). Johnstone (2001) studied the distribution of the largest eigenvalue in PCA. Baik and Silverstein (2006) investigated the convergence of the sample eigenvalues and eigenvectors under the spiked population. The degree of discrepancy in terms of the angle between the directions of sample and population eigenvectors was further derived in Paul (2007); Nadler (2008) for both $0 < \alpha < 1$ and $\alpha > 1$ situations. A phenomenon of retarded learning was observed that the angle goes through a critical phase transition from angle equal to 90° for $\lambda < \sqrt{\alpha}$ to angle less than 90° for $\lambda > \sqrt{\alpha}$. Therefore, one can only detect signals whose corresponding eigenvalues are larger than the critical value $\sqrt{\alpha}$ in PCA. More general results have been obtained by Hoyle and Rattray (2004) and Hoyle and Rattray (2007); Hoyle (2010) for general population covariance matrix.

In the present work, we study the analogous asymptotic results in the joint limit $N, M \to \infty$ with $M/N = \alpha$ in the supervised learning context for the SVM classification method. We focus on the generalization performance of SVM by deriving analytical results for the angle between the estimated direction and the true direction and investigating how this angle depends on μ, α and other model parameters. We consider a spiked population model and assume that the data from each class are generated from a purely noise model spiked with a few significant eigenvalues. We derive the analytical results using the replica method developed in statistical mechanics and also compare with numerical simulations on finite size systems. To the best of our knowledge, the present paper is the first that provides not just bounds, but sharp predictions of the asymptotic behavior of the SVM estimators in the limit $N, M \to \infty$ at fixed $M/N = \alpha$.

An immediate application of our analytical findings is for tuning parameter selection. SVM is required to solve problem of determining the tuning parameter τ that characterizes the strength of the penalty term. Cross validation (CV) is a practically useful strategy for handling this task; its basic concept is to evaluate the prediction error by examining the data under control. Smaller values of the CV error are expected to be better to express the generative model of the data. The minimum, if it exists, of the CV error when changing τ is thus considered to obtain an optimal value of τ. However, conducting CV through grid search for finding the minimizer of the CV error is rather computationally expensive especially for high dimensional data. Here we propose a new method of selecting optimal value of τ base on analytical evaluation for the angle between the estimated SVM direction and true direction which considerably reduces the computational cost. Under the spiked population assumption, smaller angle indicates smaller test error. A surprising finding is that SVM achieves its best performance at small value of the tuning parameter. All analytical results are confirmed by numerical experiments on finite-size systems and our formula is clarified to work well for moderate-size systems.

The rest of this paper is organized as follows: In Section 2, we state SVM in the context of spike population model. The analytical results for large N, M asymptotics are presented. In Section 3, we show the result of numerical experiments to support our analytical results. An application of the proposed tuning parameter selection method to the breast cancer data is also presented in this section. The last section is devoted to the conclusion.

2. Method

In the classification problem, we are given a training dataset consisting of M observations $(\mathbf{x}_i, y_i)$, for $i = 1, \cdots, M$. Here $\mathbf{x}_i \in R^N$ represents an input vector and $y_i \in \{+1, -1\}$ denotes the corresponding output class label. Each $(\mathbf{x}_i, y_i)$ is an independent random vector distributed according to a joint distribution function $p(\mathbf{x}, y)$. We assume that y has probability p_+ to be $+1$ and probability p_- to be -1 with $p_+ + p_- = 1$. Conditional on $y = +1, -1$, $\mathbf{x}$ follows multivariate distributions $p(\mathbf{x}|y = +1)$, $p(\mathbf{x}|y = -1)$ with mean $\boldsymbol{\mu}_+, \boldsymbol{\mu}_-$ and covariance matrices $\boldsymbol{\Sigma}_+, \boldsymbol{\Sigma}_-$, respectively. Without loss of generality, assume $\boldsymbol{\mu}_+ = -\boldsymbol{\mu}_- = \boldsymbol{\mu}$. Similar to linear discriminant analysis, we make an additional simplifying homoscedasticity assumption $\boldsymbol{\Sigma}_+ = \boldsymbol{\Sigma}_- = \boldsymbol{\Sigma}$. Here $\boldsymbol{\mu} \in R^N$ and $\boldsymbol{\Sigma}$ denotes the $N \times N$ matrix. Based on this setting, the data from two classes are generated from two multivariate distributions with the same covariance but different means. The signal size can be characterized by $\mu = \|\boldsymbol{\mu}\| = \sqrt{\sum_{j=1}^N \mu_j^2}$.

We consider a spiked covariance model here. For high dimensional data, typically only few components are biologically important. The remaining structures can be considered as i.i.d. background noise. Therefore, in high-dimensional settings, a collection of data can be modeled by a low-rank signal plus noise structure (Ma, 2013; Liu et al., 2008). We use a factor analysis model to explain correlations between a set of N variables by means of a smaller set of K causal factors. Specifically, we assume the following:

Assumption 1 *Each observation vector $\mathbf{x}$ from Class $+1$ can be viewed as an independent instantiation of the following generative model*

$$\mathbf{x} = \boldsymbol{\mu} + \sum_{m=1}^{K} \sigma \sqrt{\lambda_m} \mathbf{v}_m z_m + \boldsymbol{\epsilon}. \tag{1}$$

Here $\boldsymbol{\mu}$ is the mean vector, $\lambda_m>0$, $\mathbf{v}_m \in R^N$ are orthonormal vectors, i.e. $\mathbf{v}_m^T \mathbf{v}_m = 1$ and $\mathbf{v}_m^T \mathbf{v}_{m'} = 0$ for $m \neq m'$, $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}/\mu = \mathbf{v}_1$. The random variables $z_1, \cdots, z_K \overset{i.i.d}{\sim} N(0,1)$. The vector $\boldsymbol{\epsilon} = \{\epsilon_1, \cdots, \epsilon_N\}$ whose elements $\epsilon_j s$ are i.i.d random variables with $E(\epsilon_j) = 0$, $E(\epsilon_j^2) = \sigma^2$ and $E(\epsilon_j^3)<\infty$. The $\epsilon_j s$ are independent of $z_m s$. The $\mathbf{x}$ from Class -1 can be modeled in a similar way with $\boldsymbol{\mu}$ replaced by $-\boldsymbol{\mu}$.

In model (1), λ_m represents the strength of the m-th biological component, and σ^2 represents the level of background noise. The real biology is typically low-dimensional, i.e. $K \ll N$. Considering signal as one of the biological components, without loss of generality, we assume that $\boldsymbol{\mu}$ is in the same direction as $\mathbf{v}_1$, i.e. $\hat{\boldsymbol{\mu}} = \mathbf{v}_1$. Note that the eigenvalue λ_m is not necessarily decreasing in m and λ_1 is not necessarily the largest eigenvalue. From (1), the covariance matrix is

$$\mathbf{\Sigma} = \sigma^2 \mathbf{I}_N + \sum_{m=1}^{K} \sigma^2 \lambda_m \mathbf{v}_m \mathbf{v}_m^T, \tag{2}$$

where $\mathbf{I}_N$ is N-dimensional identity matrix. Although the ϵ_js are i.i.d, we didn't impose any parametric form for the distribution of ϵ_j which allows for very flexible covariance structures for $\mathbf{x}$, and thus the results are quite general. The requirement for the finite third order moment is to ensure Berry-Esseen central limit theorem applies. The Assumption 1 is also called spiked population model and has been used in many situations, see Baik and Silverstein (2006); Marcenko and Pastur (1967); Johnstone (2001) for examples. Such a population covariance is a finite rank perturbation of multiple of the identity matrix. In other words, all but finitely many eigenvalues of the population covariance matrix are the same. Examples of spiked data include speech recognition (Trevor Hastie, 1995), mathematical finance (LALOUX et al., 2000), wireless communications (Telatar, 1999), and physics of mixture (Sear and Cuesta, 2003).

The task of linear classification is to construct a hyperplane $\mathbf{x}^T\mathbf{w} = 0$ ($\mathbf{w} \in R^N$) so that the new data vector $\mathbf{x}$ is assigned to Class $+1$ when $\mathbf{x}^T\mathbf{w}>0$ and Class -1 otherwise. If the training data are linearly separable, SVM seeks to find this hyperplane such that the minimal distance between the hyperplane and the data point from each class is maximized. The hard-margin SVM solution can be formulated in terms of the following optimization problem

$$\begin{aligned} &\min_{\mathbf{w}} \left[\mathbf{w}^T\mathbf{w}\right] \\ \text{s.t.} \quad &\frac{y_i \mathbf{x}_i^T \mathbf{w}}{\sqrt{N}} \geq 1,\ i = 1, \cdots, M. \end{aligned} \tag{3}$$

To extend SVM to cases in which the data are not linearly separable, we introduce the slack variables ξ_i for $i = 1, \cdots, M$. The soft-margin SVM solution can be formulated in terms of the following optimization problem

$$\begin{aligned} &\min_{\mathbf{w}} \left[\mathbf{w}^T\mathbf{w} + \tau \sum_{i=1}^{M} \xi_i\right] \\ \text{s.t.} \quad &\frac{y_i \mathbf{x}_i^T \mathbf{w}}{\sqrt{N}} + \xi_i \geq 1,\ \xi_i \geq 0,\ i = 1, \cdots, M, \end{aligned} \tag{4}$$

where the tuning parameter τ determines the trade-off between increasing the margin-size and ensuring that the $\mathbf{x}_i$ lie on the correct side of the margin. For sufficiently large values of τ, the soft-margin SVM will behave identically to the hard-margin SVM. We will show below that, as $\tau \to \infty$, the asymptotic result of soft-margin SVM is the same as the asymptotic result of hard-margin SVM.

For the setting described in Assumption 1, the normal direction vector of the separating hyperplane based on Bayes optimal rule is in the same direction as $\boldsymbol{\mu}$. Therefore, the performance of any classification method can be evaluated by the angle between the normal direction vector of its separating hyperplane and $\boldsymbol{\mu}$. Propositions 1 and 2 provide the sharp prediction of the high-dimensional limiting angles for hard-margin SVM and soft-margin SVM respectively.

Proposition 1 *Under Assumption 1, in the limit $N, M \to \infty$, with fixed $\alpha = M/N$, denote θ the angle between $\boldsymbol{\mu}$ and $\mathbf{w}$ solved from the hard-margin SVM algorithm (3), then $\cos\theta$ converges to ρ that is determined by the following two nonlinear equations*

$$\frac{1-\rho^2}{1+\lambda_1\rho^2} = \alpha \int_{-\infty}^{z_c} Dz(z_c - z)^2, \tag{5}$$

$$\frac{\rho}{\sqrt{1+\lambda_1\rho^2}} = \alpha \int_{-\infty}^{z_c} Dz(z_c - z)\left(\frac{\mu}{\sigma} + \frac{\lambda_1\rho}{\sqrt{1+\lambda_1\rho^2}} z\right). \tag{6}$$

where z_c is an unknown parameter needs to be estimated, μ, σ, λ_1 are defined in (2), and the standard notation $Dz = \frac{dz}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$.

All the proofs are given in the supplementary materials. From equations (5) and (6), we can solve two unknown parameters ρ and z_c given α, μ, σ, and λ_1. It is interesting to note that the results do not depend on $\lambda_2, \cdots, \lambda_K$ which means that only the variance along the signal direction has influence on SVM performance. This observation is also confirmed by extensive simulations in Section 3.1. All the biological components in orthogonal directions have no impact. The nonlinear equations (5) and (6) have no closed form solution. We have to use some numerical algorithms to solve them. As expected, it can be easily checked from the numerical studies in Section 3 that $\cos(\theta)$ increases with α as well as the signal to noise ratio μ/σ, but decreases with λ_1.

Proposition 2 *Under Assumption 1, in the limit $N, M \to \infty$, with fixed $\alpha = M/N$, denote θ the angle between $\boldsymbol{\mu}$ and $\mathbf{w}$ solved from the soft-margin SVM algorithm (4), then $\cos\theta$ converges to ρ that is determined by the following three nonlinear equations*

$$\frac{1-\rho^2}{1+\lambda_1\rho^2} - \alpha q^2\hat{\tau}^2 \int_{-\infty}^{z_c - q\hat{\tau}} Dz - \alpha \int_{z_c - q\hat{\tau}}^{z_c} Dz(z_c - z)^2 = 0, \tag{7}$$

$$2q - 1 - \alpha q\hat{\tau} \int_{-\infty}^{z_c - q\hat{\tau}} Dzz - \alpha \int_{z_c - q\hat{\tau}}^{z_c} Dz(z_c - z)z = 0, \tag{8}$$

$$\frac{\rho F}{\sqrt{1+\lambda_1\rho^2}} - \frac{\alpha q\hat{\tau}\mu}{\sigma} \int_{-\infty}^{z_c - q\hat{\tau}} Dz - \frac{\alpha\mu}{\sigma} \int_{z_c - q\hat{\tau}}^{z_c} Dz(z_c - z) = 0, \tag{9}$$

where

$$F = 1 - \alpha\lambda_1 q\hat{\tau} \int_{-\infty}^{z_c - q\hat{\tau}} Dzz - \alpha\lambda_1 \int_{z_c - q\hat{\tau}}^{z_c} Dz(z_c - z)z,$$

and

$$z_c = \frac{1/\sqrt{q_0} - \mu\rho}{\sigma\sqrt{1+\lambda_1\rho}}, \qquad \hat{\tau} = \frac{\sigma\tau}{\sqrt{q_0}\sqrt{1+\lambda_1\rho}}.$$

Therefore, given $\alpha, \lambda_1, \mu, \sigma$, and τ, equations (7), (8), (9) can be used to solve three unknown parameters ρ, q_0, and q. The nonlinear equations (7), (8), and (9) have no closed form solution. We have to use some numerical algorithms to solve them. Under Assumption 1, if we further assume that $\boldsymbol{\epsilon}$ in (1) follows a normal distribution, then the SVM test error is $\varepsilon = \Phi\left(-\frac{\rho}{\sqrt{1+\lambda_1\rho^2}}\frac{\mu}{\sigma}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$.

It is interesting to note that, as $\tau \to \infty$, the two equations (7) and (9) are equivalent to (5) and (6) respectively. Therefore, for large τ, the behavior of soft-margin SVM is the same as hard-margin SVM. Our simulation studies in Section 3.2 will also confirm this.

For a given dataset, α, λ_1, μ, and σ can be estimated, therefore Proposition 2 allows us to select optimal tuning parameter τ by studying the dependence of ρ on τ for fixed $\alpha, \lambda_1, \mu, \sigma$.

We now discuss how to estimate λ_1, μ, and σ from the data. To estimate the background noise level σ^2, we use a robust variance estimate based on the full matrix of data values (Liu et al., 2008); that is, for the full set of $M \times N$ entries of the original $M \times N$ data matrix $\mathbf{X}$, we calculate the robust estimate of scale, the median absolute deviation from the median (MAD), to estimate σ as

$$\hat{\sigma} \;=\; \frac{\mathrm{MAD}_{\mathbf{X}}}{\mathrm{MAD}_{N(0,1)}}. \tag{10}$$

Here $\mathrm{MAD}_{\mathbf{X}} = \mathrm{median}(|x_{ij} - \mathrm{median}(\mathbf{X})|)$ and $\mathrm{MAD}_{N(0,1)} = \mathrm{median}(|r_i - \mathrm{median}(\mathbf{r})|)$, where $\mathbf{r}$ is a MN-dimensional vector whose elements are i.i.d. samples from $N(0,1)$ distribution.

To estimate λ_1, we use the results from Baik and Silverstein (2006) which shows that in the limit of $M, N \to \infty$, with fixed $\alpha = M/N$, the sample eigenvalue $\tilde{\lambda}_1$ satisfies

$$\tilde{\lambda}_1 \;\xrightarrow{a.s.}\; \begin{cases} (\lambda_1+1)\left(1+\frac{1}{\alpha\lambda_1}\right)-1, & \text{for} \quad \lambda_1 > \sqrt{1/\alpha}, \\ (1+\sqrt{1/\alpha})^2 - 1, & \text{for} \quad \lambda_1 \le \sqrt{1/\alpha}. \end{cases} \tag{11}$$

Therefore, for any finite α, $\tilde{\lambda}_1$ is not a consistent estimator of λ_1. We use equation (11) to estimate λ_1 as

$$\hat{\lambda}_1 \;=\; \begin{cases} \frac{1}{2}\left(\tilde{\lambda}_1 - \frac{1}{\alpha} + \sqrt{\left(\tilde{\lambda}_1 - \frac{1}{\alpha}\right)^2 - \frac{4}{\alpha}}\right), & \text{for} \quad \tilde{\lambda}_1 > 1/\alpha + 2\sqrt{1/\alpha}, \\ \sqrt{1/\alpha}, & \text{for} \quad \tilde{\lambda}_1 \le 1/\alpha + 2\sqrt{1/\alpha}. \end{cases} \tag{12}$$

To estimate μ, let M_+, M_1 denote the sample sizes of Class $+1$ and Class -1 respectively. Define $\boldsymbol{\mu}_c = \bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-$, where $\bar{\mathbf{x}}_+$ and $\bar{\mathbf{x}}_-$ represent the sample means for Class $+1$ and Class -1 respectively. The following Proposition describes the relationship between $\boldsymbol{\mu}_c$ and μ.

Proposition 3 *Under Assumption 1, in the limit $N, M \to \infty$, with fixed $\alpha = M/N$, $r_+ = M_+/M$, and $r_- = M_-/M$, then $\|\boldsymbol{\mu}_c\|^2$ converges to*

$$4\mu^2 + \frac{\sigma^2}{\alpha r_+ r_-}. \tag{13}$$

Therefore, we estimate μ as

$$\hat{\mu} = \frac{1}{2}\sqrt{\|\hat{\boldsymbol{\mu}}_c\|^2 - \frac{\hat{\sigma}^2}{\alpha r_+ r_-}},$$

where $\hat{\sigma}$ is given from (10) and $\hat{\boldsymbol{\mu}}_c = \frac{1}{M_+}\sum_{i=1}^{M_+}\mathbf{x}_i - \frac{1}{M_-}\sum_{i=1}^{M_-}\mathbf{x}_i$ is the sample estimation of $\boldsymbol{\mu}_c$.

3. Numerical Results

3.1 Hard-margin SVM

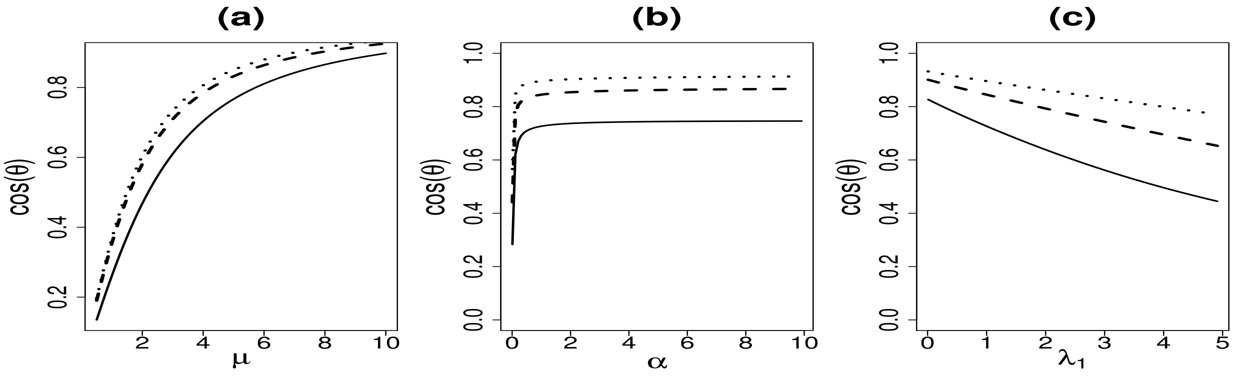

Figure 2: (a) Dependence of $\cos(\theta)$ on μ for fixed $\sigma = 1$, $\lambda_1 = 1$ and $\alpha = 0.1$ (solid), 0.5 (dashed), 1.5 (dotted); (b) Dependence of $\cos(\theta)$ on α for fixed $\lambda_1 = 1$ and $\mu = 3$ (solid), 5 (dashed), 7 (dotted); (c) Dependence of $\cos(\theta)$ on λ_1 for fixed $\alpha = 1$ and $\mu = 3$ (solid), 5 (dashed), 7 (dotted).

Figure 2 shows the dependence of $\cos(\theta)$ on the parameters μ, α, and λ_1 based on numerical solutions of equations (5) and (6). Here θ represents the angle between the directions of SVM separating hyperplane and Bayes optimal separating hyperplane. For spiked population model (1), the normal vector of Bayes optimal separating hyperplane lies in the direction of $\boldsymbol{\mu}$. Discrimination methods whose normal vector $\mathbf{w}/\|\mathbf{w}\|$ lies close to this direction should have good "generalization" properties, i.e., new data will be discriminated as well as possible. Figure 2(a) shows that, for fixed α and λ_1, the classification performance is improved as we increase the signal size μ. Figure 2(b) shows that, for fixed μ and λ_1, $\cos(\theta)$ increases with α, indicating that the classification performance is improved by adding more samples to the training data. For $\alpha<1/2$, the increasing is faster; for $\alpha>2$, the increasing becomes slower and saturated. This indicates that, for HDLSS situations, increasing training data can improve the prediction power dramatically; while for situations when sample size is twice as big as the dimension, adding more samples can not gain too much power. Figure 2(c) shows that, for fixed μ and α, $\cos(\theta)$ decreases with λ_1 as expected.

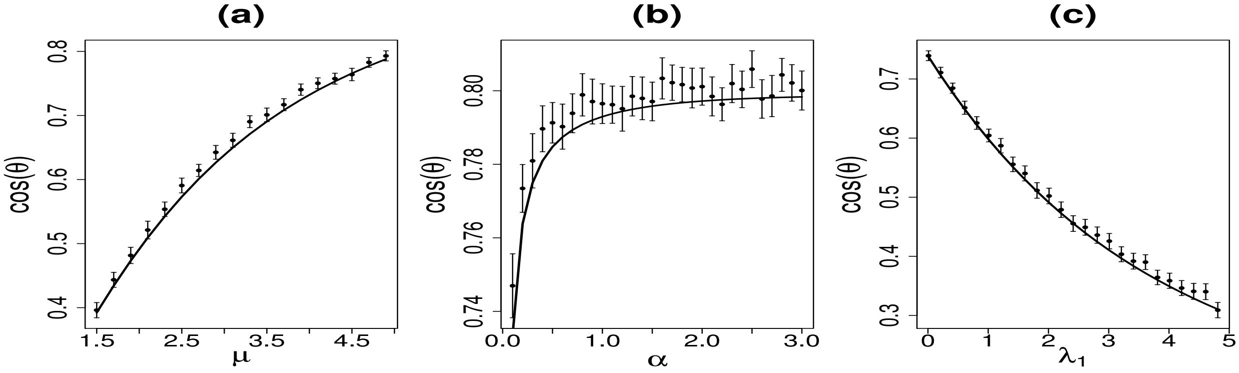

Figure 3: Comparison of analytical calculations with simulation experiments. The solid curves represent the theoretical results, the dots and bars represent the mean and standard error of the estimated $\cos\theta$ by applying SVM algorithm (3) to 100 simulated data sets for each parameter setting. In simulations, the dimension $N = 100$, the background noise $\sigma = 1$. The other parameters are: (a) $\alpha = 1, \lambda_1 = 2$; (b) $\mu = 5, \lambda_1 = 2$; (c) $\alpha = 1, \mu = 2$.

To examine the validity of our analysis and to determine the finite-size effect, Figure 3 provides the comparison with numerical simulations on finite size systems. Similar to Figure 2, we consider the dependence of $\cos(\theta)$ on three parameters μ, α and λ_1 in the plots in Figure 3 (a), (b), and (c) respectively. Here the dimension of the simulated data $N = 100$ and the data are generated according to Assumption 1 with ϵ_j follows i.i.d standard normal distribution. We repeat simulation 100 times for each parameter setting. The mean and standard errors over 100 replications are presented. From Figure 3, we can see that our analytical curves show fairly good agreement with the simulation experiment. Thus our analytical formulas (5) and (6) provide reliable estimates even for moderate system sizes. The benefit of these formulas is their computational ease. We also find that the simulation results for SVM estimators are independent of the choices of orthogonal components $\lambda_{m\geq 2}$ which further confirms that the analytical results described by Proposition 1 are correct.

3.2 Soft-margin SVM

Figure 4 shows the dependence of $\cos(\theta)$ on the parameters μ, α, and λ_1 based on the solution of nonlinear equations (7), (8), and (9) for fixed $\tau = 1$ and $\sigma = 1$. Similar to Figure 2, the $\cos\theta$ increases with μ and α but decreases with λ_1.

To study the the influence of the tuning parameter τ on the performance of the soft-margin SVM classification method (4). Figure 5 shows the dependence of $\cos(\theta)$ as function

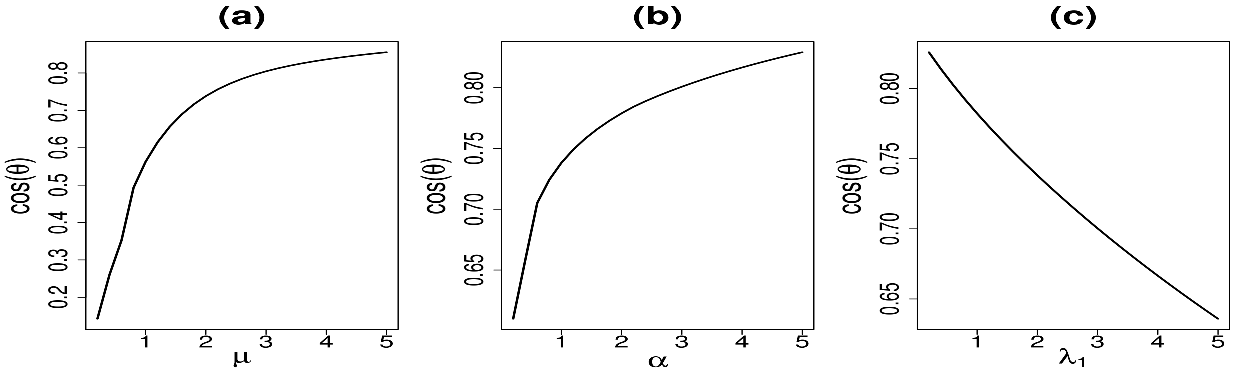

Figure 4: Dependence of $\cos(\theta)$ on μ, α, and λ_1 for $\tau = 1$ and $\sigma = 1$. (a) Dependence of $\cos(\theta)$ on μ for fixed $\lambda_1 = 2$ and $\alpha = 1$; (b) Dependence of $\cos(\theta)$ on α for fixed $\lambda_1 = 2$ and $\mu = 2$; (c) Dependence of $\cos(\theta)$ on λ_1 for fixed $\alpha = 1$ and $\mu = 2$.

of $\log\tau$ for fixed α, μ, λ_1, and σ. Both the analytical solution based on Proposition 2 and numerical experiment based on simulated finite dimensional data are provided and they excellently agree with each other. In simulation, we randomly generate a training set and a test set for the given parameter setting, the test error can be obtained by applying the classifier built from the training set to the test set. The results from the summary over 100 replications are given in Figure 5. From the upper panel, it is interesting to note that $\cos\theta$ reaches a maximum value as one decreases the tuning parameter τ to a threshold value. After that value, further decreasing τ cannot change $\cos\theta$. On the other hand, if we increase τ, $\cos\theta$ will approach the value determined by the hard-margin SVM method as shown by the dashed line in the upper panel of Figure 5. These observations are further confirmed by the dependence of test error ε on $\log\tau$ as shown in the lower panel. The test error reaches a minimum value if we decrease τ to the same threshold value as for $\cos\theta$. From equations (7), (8), and (9), it can be derived that the limiting value of $\cos\theta$ as $\tau \to 0$ is

$$\rho_c = \cos\theta_c = \sqrt{\frac{\alpha\left(\frac{\mu}{\sigma}\right)^2}{1+\alpha\left(\frac{\mu}{\sigma}\right)^2}} \tag{14}$$

which is independent of λ_1. This finding of λ_1 independence is also confirmed by numerical simulations with data on finite size systems. Therefore, if $\boldsymbol{\epsilon}$ in (1) follows a normal distribution, then the best test error we can achieve using the soft-margin SVM classification method (4) is $\Phi\left(-\frac{\rho_c}{\sqrt{1+\lambda_1\rho_c^2}}\frac{\mu}{\sigma}\right)$.

Koo et al. (2008) studied the asymptotic behavior of the coefficients of the linear SVM in the limit of $M \to \infty$ with N fixed. They established a Bahadur type representation of the coefficients and derived their asymptotic normality and statistical variability. Denote $\mathbf{w}^\star$ the minimizer of the population version of the SVM loss function. It was shown in Koo et al. (2008) that the SVM solution $\hat{\mathbf{w}}$ converges to $\mathbf{w}^\star$ which is in the same direction as $\boldsymbol{\mu}$

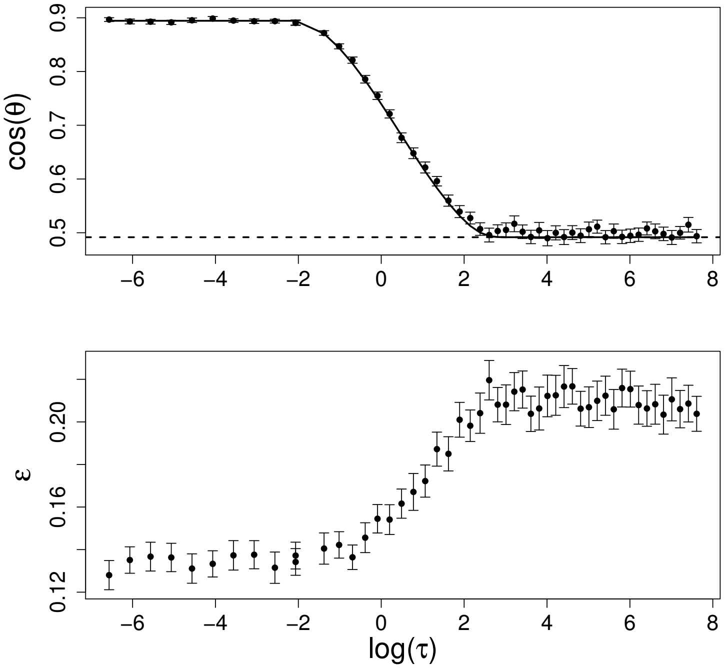

Figure 5: Upper panel: compare theoretical result with simulation experiment for the dependence of $\cos(\theta)$ on tuning parameter $\log(\tau)$ for fixed $\alpha = 1$, $\mu = 2$, $\sigma = 2$, and $\lambda_1 = 2$. The solid line is the theoretical curve, the dots and bars represent the mean and standard error based on 100 simulated data sets at each parameter setting. In simulation, the dimension $N = 100$. The dashed line represents the value based on the hard-margin SVM solution from equations (5) and (6). Lower panel: dependence of the test error ε on $\log(\tau)$ based on simulations.

under the spiked population setting (1). Therefore $\rho \to 1$ as $M \to \infty$ with N fixed. This can be confirmed in (14) by letting $\alpha \to \infty$ on the right hand side. On the other hand, if we let $N \to \infty$ with M fixed, from (14) we get $\rho_c \to 0$ if $\mu/\sqrt{N} \to 0$ and $\rho_c \to 1$ if $\mu/\sqrt{N} \to \infty$. This confirms the results of Hall et al. (2005) for HDLSS setting. Therefore, our asymptotic results are more general with both traditional and HDLSS asymptotics as special cases.

The analytical results in Figure 5 are based on the true values for α, λ_1, μ and σ which ultimately need to be estimated from the given data. In Figure 6 we provide the comparison between the results using the true values and the results using the estimated values for μ, α, λ_1 and σ. For each simulated data, we first estimate μ, α, λ_1 and σ and then use them to derive theoretical results. Figure 6 indicates that the influence of moderate estimation errors in the parameters is small.

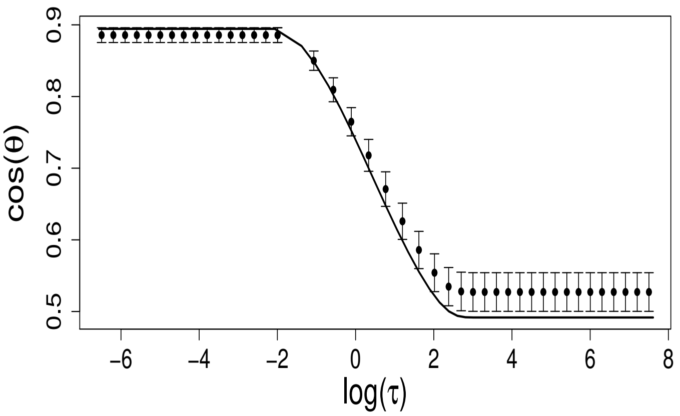

Figure 6: Comparison between the results using the true values and the results using the estimated values for parameters. Here the true parameter values are $\alpha = 1$, $\mu = 2$, $\sigma = 1$, and $\lambda_1 = 2$. The solid curve represents the results derived using the true values. The dots and bars represent the means and standard errors of the $\cos\theta$ values derived using the estimated parameters for 100 simulated data sets.

Although Figure 5 suggests that, for spiked population model, the best performance of SVM is achieved at the smallest value τ, in practice, using too tiny τ could cause difficulties in numerically solving the optimization problem. In order to provide a practical recommendation for the tuning parameter, we need to estimate the threshold value τ_c at which the limiting value $\cos\theta_c$ is almost achieved, i.e. the elbow point in Figure 5. More precisely, τ_c is defined as $\tau_c = \max\{\tau : \rho(\tau) = \rho_c\}$. In practice, we can compute τ_c by numerically finding the largest τ that can give $\cos\theta = \rho_c$ and use it as a guideline for choosing τ. Figure 7 displays the change of $\log\tau_c$ as functions of the parameters μ, α, λ_1. It is shown that $\log\tau_c$ decreases with all three parameters.

3.3 Check the model assumptions

The key assumptions for deriving the results in Propositions 1 and 2 are homoscedasticity (equal covariance) and spiked covariance condition (1). In this section, we use simulation to

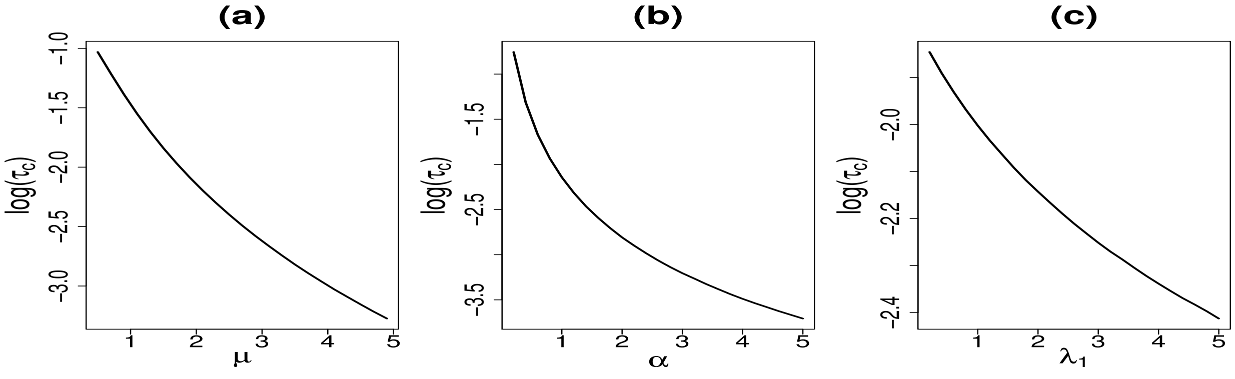

Figure 7: (a) Dependence of $\log(\tau_c)$ on μ for fixed $\lambda_1 = 2$ and $\alpha = 1$; (b) Dependence of $\log(\tau_c)$ on α for fixed $\lambda_1 = 2$ and $\mu = 2$; (c) Dependence of $\log(\tau_c)$ on λ_1 for fixed $\alpha = 1$ and $\mu = 2$.

study the validity of our method in situations where these assumptions are not true. Figure 8 is for situation where the two covariance matrices from the positive and negative classes are different. In simulation, we first generate M samples from $N(0, \sigma^2 \mathbf{I}_p)$ distribution. Then $M/2$ of them are shifted by μ in x_1 direction to form the positive class and the remaining $M/2$ are shifted by $-\mu$ in x_1 direction to form the negative class. Both classes are further divided into two subclasses with sample size $M/4$ for each. In the positive class, the two subclasses are separated by shifting in x_2 direction by μ and $-\mu$ respectively. Similarly, the two subclasses in the negative class are separated by shifting in x_3 direction by μ and $-\mu$ respectively. The data generated in this way satisfies the spiked assumption but the two classes have different covariances. Figure 8 shows that the theoretical estimation and direct computation agree fairly well with each other. We have tried several different settings for μ, α and σ and got similar results. Therefore, our method is fairly robust to homoscedasticity as long as the spiked condition (1) holds for the covariance matrices of both classes.

Figures 9, 10, and 11 are for situations where the spiked assumption is violated. In simulation, we first generate M samples from $N(0, \mathbf{\Sigma})$ distribution. Then $M/2$ of them are shifted by μ in x_1 direction to form the positive class and the remaining $M/2$ are shifted by $-\mu$ in x_1 direction to form the negative class. The covariance matrix $\mathbf{\Sigma}$ is diagonal with the ith eigenvalue equal to $i\sigma^2$ for $i \leq K$ and σ^2 for $i{>}K$. The spiked condition requires $K \ll p$. If we increase K, the spiked condition can be violated. Figures 9, 10, and 11 show the results for $K = 0.1N$, $K = 0.3N$, and $K = 0.5N$ respectively. For situations where the number of uncommon eigenvalues K is less than 10% of the total number of variables N, our method can provide quite accurate estimation for $\cos\theta$ and also bigger $\cos\theta$ corresponds to smaller test error as illustrated in Figure 9. For situations where K is 30% of N, our method can still provide reasonable estimation for $\cos\theta$, but $\cos\theta$ cannot be used as a criterion for choosing τ because bigger $\cos\theta$ does not always correspond to

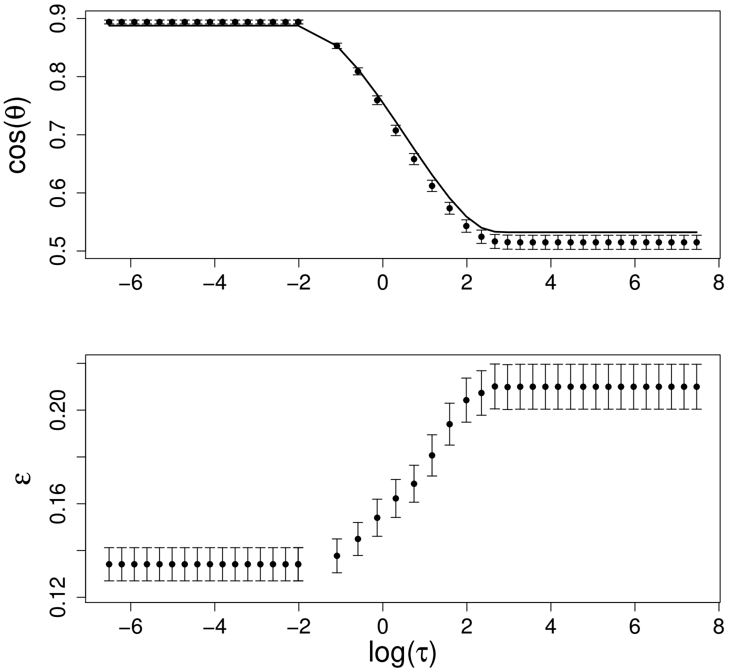

Figure 8: Comparison of theoretical prediction with direct computation for simulated data using parameters $\alpha = 1, \mu = 2, \sigma = 2$, and $\lambda_1 = 2$. In simulation, the covariances are different between the positive class and the negative class.

smaller test error as illustrated in Figure 10. For situations where K is 50% of N, our method cannot provide estimation for $\cos\theta$. Moreover, $\cos\theta$ and ε behavior in an opposite way, i.e. smaller $\cos\theta$ corresponds to smaller test error as illustrated in Figure 11.

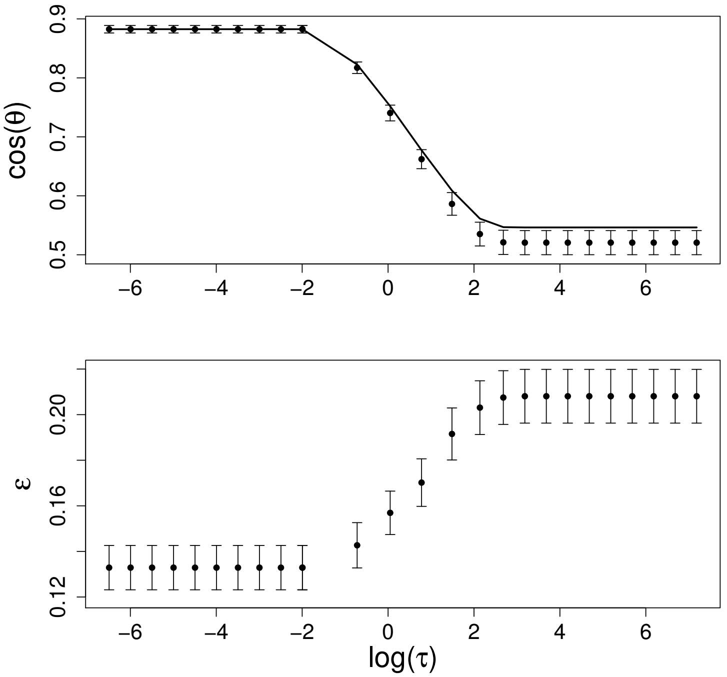

Figure 9: Comparison of theoretical prediction with direct computation for simulated data using parameters $\alpha = 1, \mu = 2, \sigma = 2$, and $\lambda_1 = 2$. In simulation, the number of the uncommon eigenvalues K is equal to 10% of the total number of variables N.

In summary, our simulations indicate that the proposed method depends on the spiked assumption but is not sensitive to the homoscedasticity violation. The spiked assumption is based on factor analysis which is one of the most useful tools for modeling common dependence among all the variables. In genetics, factor analysis modeling appeared to be useful tools to investigate the dependence structure in high-dimensional microarray data. It can fit the data with covariance matrix governed by linkage disequilibrium patterns (Rochat et al., 2007). For data set which cannot be modeled using spiked population, our results indicate that further exploring the data structure is useful for understanding the classification performance.

3.4 Real Data

We apply our methods to a breast cancer dataset from The Cancer Genome Atlas Research Network (Network, 2008) which include two subtypes: LumA and LumB. As in Liu et al. (2008), we filter the genes using the ratio of the sample standard deviation and sample mean of each gene. After gene filtering, the dataset contained 235 patients with 56 genes. Among the 235 samples, there are 154 LumA samples and 81 LumB samples.

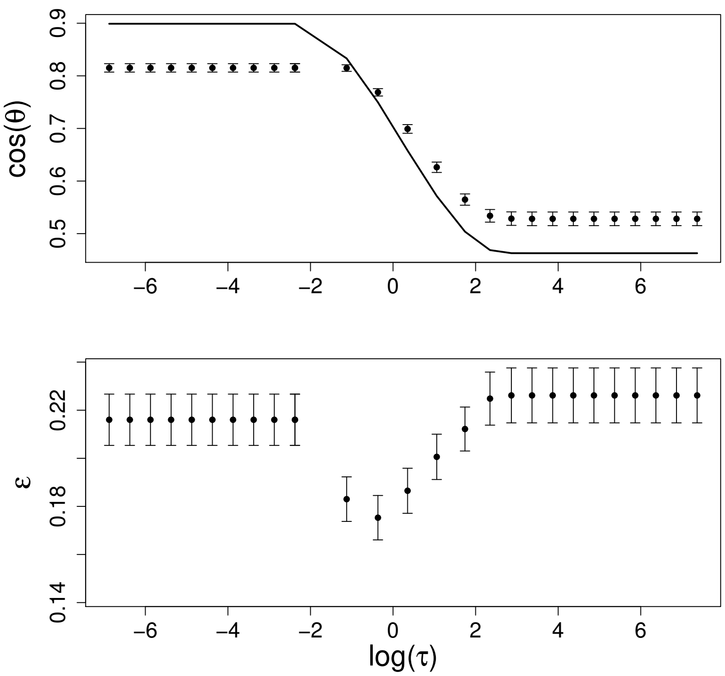

Figure 10: Comparison of theoretical prediction with direct computation for simulated data using parameters $\alpha = 1, \mu = 2, \sigma = 2$, and $\lambda_1 = 2$. In simulation, the number of the uncommon eigenvalues K is equal to 30% of the total number of variables N.

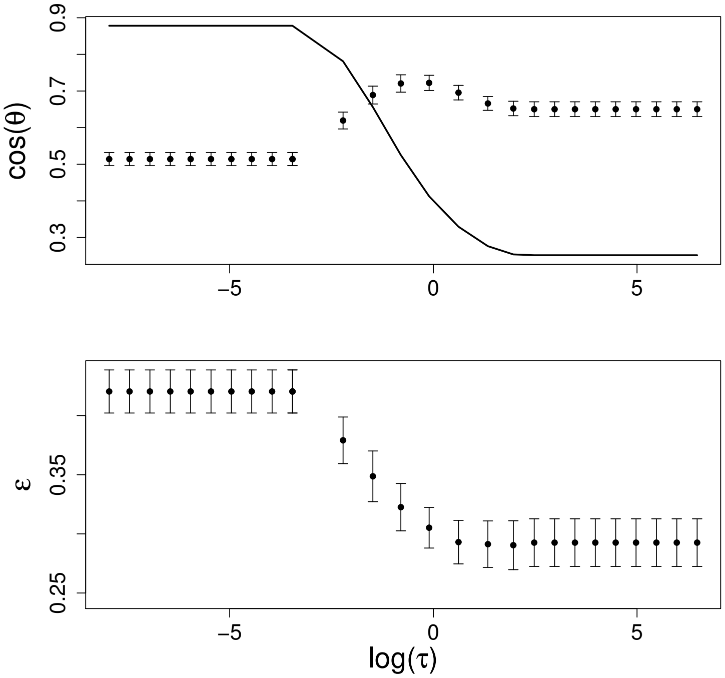

Figure 11: Comparison of theoretical prediction with direct computation for simulated data using parameters $\alpha = 1, \mu = 2, \sigma = 2$, and $\lambda_1 = 2$. In simulation, the number of the uncommon eigenvalues K is equal to 50% of the total number of variables N.

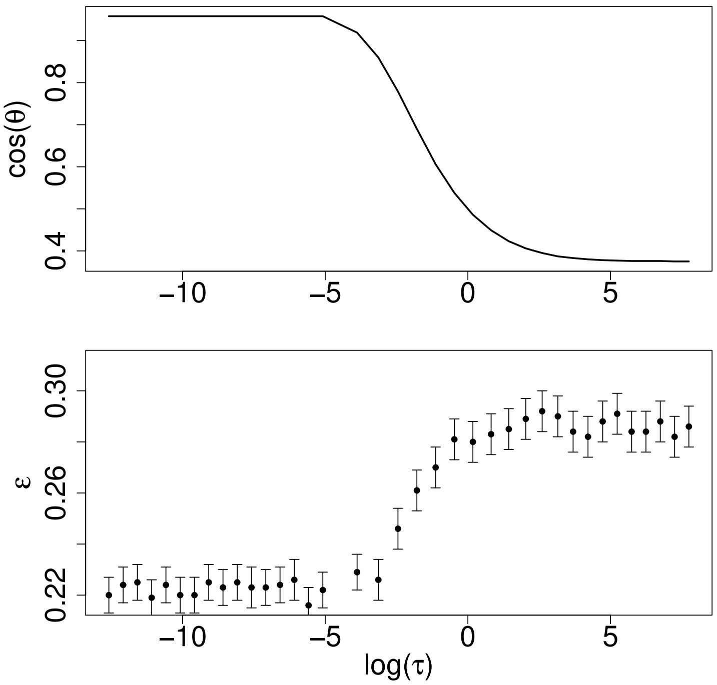

Figure 12: Upper panel: theoretical prediction of the dependence of $\cos(\theta)$ on tuning parameter $\log(\tau)$ based on the solutions from equations (7), (8) and (9) using parameters estimated from the breast cancer data. Lower panel: dependence of cross-validation error ε on tuning parameter $\log(\tau)$. The dots and bars represent the mean and standard error of the cross validation error based on 100 random splittings of the breast cancer data.

We consider LumA as Class +1 and LumB as Class -1. Assume the data are generated based on model (1), using the method discussed in Section 2, we obtain the following parameter estimations: $\hat{\mu} = 3.80$, $\hat{\sigma} = 2.32$, $\hat{\lambda}_1 = 4.06$, $\alpha = 4.20$, $N = 56$, $M = 235$, $M_+ = 154$, $M_- = 81$. The upper panel of Figure 12 shows the analytical curve for the dependence of $\cos\theta$ on τ. It shows that if we choose τ less than 6.19×10^{-3}, we can get the smallest angle. The lower panel of Figure 12 shows the dependence of the cross validation errors as a function of τ. The cross validation errors are computed by randomly splitting the data into two parts, 90% for training and 10% for test. The mean and standard error over 100 random splitting are reported in the lower panel of Figure 12. It shows that the cross validation error can achieve minimum value if τ is less than around 5×10^{-3}. The two results are consistent with each other and similar to the previous simulation results as shown in Figure 5. This indicates that model (1) is a reasonable assumption for this data set.

4. Conclusion

In this study, we examine the asymptotic behavior of SVM in the limit of $N, M \to \infty$ with fixed $\alpha = M/N$. We investigate the estimators of both the hard-margin SVM and the soft-margin SVM methods. Our focus is on the angle between the direction of the estimated separating hyperplane and the Bayes optimal separating hyperplane. Under spiked population model assumption, we analytically evaluate the relation between this angle and the SVM tuning parameter. On the basis of this finding, a new method of selecting tuning parameter is developed for analyzing high dimensional data which significantly reduces the computational cost. The analytical calculations are compared with numerical simulations on finite-size systems and the agreement between the numerical data and the analytical result is fairly good, and thus, our formulas are validated. Although the asymptotic results that we have obtained apply only to the spiked population model, they have shed a new light on the asymptotic behavior of SVM and can also improve the practical use of SVM in various aspects. For situations where the spiked model cannot be applied, one possible solution is to use the generalized spiked population model proposed in Bai and Yao (2012) to re-derive our results. This is one of our future research topics.

It is shown in Figure 1 that a lot of data points are piling up on the two boundaries. This is a phenomenon called data piling which has been studied in Marron et al. (2007) in more details. The reason is that the hinge loss function used in SVM is not continuous differentiable. The consequence of data piling is that the generalization performance is adversely affected. To overcome this problem, Marron et al. (2007) proposed a new classification method call Distance Weighted Discrimination (DWD) which does not have data piling problem. Simulation studies have shown that DWD typically yields better classification performance than SVM in high dimensions, but deeper theoretical evidence is strongly desired. It will be interesting to study the asymptotic property of DWD and compare it with SVM from a analytical point of view. This is another direction that we will pursue in future. The same technique can also be used in other popular classifier that currently heavily relies on cross validation. Examples include the hybrid of DWD and SVM proposed in Qiao and Zhang (2015) and the Large-Margin Unified Machines proposed in Liu et al. (2011).

Acknowledgments

The author would like to thank the Action Editor Professor Xiaotong Shen and three reviewers for their constructive comments and suggestions, which led to substantial improvements of the presentation of this paper.

References

Z. D. Bai. Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statistica Sinica*, 9:611–677, 1999.

Zhidong Bai and Jianfeng Yao. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167 – 177, 2012.

Jinho Baik and Jack W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382 – 1408, 2006.

Peter Hall, J. Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005.

D. C. Hoyle and M. Rattray. Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Phys. Rev. E*, 69:026124, 2004.

D. C. Hoyle and M. Rattray. Statistical mechanics of learning multiple orthogonal signals: Asymptotic theory and fluctuation effects. *Phys. Rev. E*, 75:016101, 2007.

David C Hoyle. Statistical mechanics of learning orthogonal signals for general covariance models. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(04):P04009, 2010.

Iain M. Johnstone. On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

S. K. Jung and J. S. Marron. PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37:4104–4130, 2009.

Sungkyu Jung, Arusharka Sen, and J.S. Marron. Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis*, 109:190 – 203, 2012.

Ja-Yong Koo, Yoonkyung Lee, Yuwon Kim, and Changyi Park. A bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, 9:1343–1368, June 2008.

LAURENT LALOUX, PIERRE CIZEAU, MARC POTTERS, and JEAN-PHILIPPE BOUCHAUD. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 03(03):391–397, 2000.

Yufeng Liu, David Neil Hayes, Andrew Nobel, and J. S Marron. Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293, 2008.

Yufeng Liu, Hao Helen Zhang, and Yichao Wu. Hard or soft classification? large-margin unified machines. *Journal of the American Statistical Association*, 106(493):166–177, 2011.

Zongming Ma. Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, 41(2):772–801, 04 2013.

V. A. Marcenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, April 1967.

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

Boaz Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, 2008.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068, 2008.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642, 2007.

Xingye Qiao and Lingsong Zhang. Flexible high-dimensional classification machines and their asymptotic properties. *Journal of Machine Learning Research*, 16:1547–1572, 2015.

R. H. Rochat, L. de las Fuentes, G. Stormo, V. G. Davila-Roman, and C. Charles Gu. A novel method combining linkage disequilibrium information and imputed functional knowledge for tagsnp selection. *Hum Hered*, 64(4):243–249, 2007.

Richard P. Sear and José A. Cuesta. Instabilities in complex mixtures with a large number of components. *Phys. Rev. Lett.*, 91:245701, Dec 2003.

E. Telatar. Capacity of multi-antenna Gaussian channels. *Eur. Trans. Telecomm. ETT*, 10 (6):585–596, November 1999.

Robert Tibshirani Trevor Hastie, Andreas Buja. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, NY, 1995.