Weighted Principal Support Vector Machines for Sufficient Dimension Reduction in Binary Classification

Yichao Wu

A joint work with Seung Jun Shin, Hao Helen Zhang and Yufeng Liu
Outline

1. Introduction
2. Weighted Principal Support Vector Machine
3. Kernel Weighted PSVM
4. Numerical Results
5. Summary
Outline

1. Introduction
2. Weighted Principal Support Vector Machine
3. Kernel Weighted PSVM
4. Numerical Results
5. Summary
For a given pair of \((Y, X) \in \mathbb{R} \times \mathbb{R}^p\),

- **Sufficient Dimension Reduction (SDR)** seeks a matrix \(B = (b_1, \cdots, b_d) \in \mathbb{R}^{p \times d}\) which satisfies

\[
Y \perp X | B^\top X. \tag{1}
\]
Central Subspace

- Dimension Reduction Subspace (DRS) is defined by $\text{span}(B) \subseteq \mathbb{R}^p$.

Central Subspace

\[ S_{Y|X} \] is the intersection of all DRSes.

- \( S_{Y|X} \) has a minimum dimension among all DRS and uniquely exists under very mild conditions. (Cook, 1998, Prop. 6.4)
- We assume \( S_{Y|X} = \text{span}(B) \).
- The dimension of \( S_{Y|X} \), \( d \), is called the structure dimension.
Estimation of $S_{Y|X}$

Seminal paper in Early 1990.


- Many other methods:
  - Sliced Average Variance Estimation (SAVE, 1991)
  - Principal Hessian Directions (pHd, 1992)
  - Contour Regression (2005)
  - Fourier-Transformation-Based Estimation (2005)
  - Directional Regression (2007)
  - Cumulative Sliced Regression (CUME; 2008)
  - and many others ...
Sliced Inverse Regression

Foundation of SIR

Under the linearity condition,

\[ E(Z|Y) \in S_Y|z = \Sigma^{1/2} S_Y|x. \]

where \( Z = \Sigma^{-1/2}\{X - E(X)\}. \)

- **Slice response** into \( H \) non-overlapping intervals, \( I_1, \cdots, I_H \),

\[ \hat{m}_h := E_n(Z|Y \in I_h) = \frac{1}{n_h} \sum_{Y \in I_h} z_i, \quad h = 1, \cdots, H. \]

- \( B \) is estimated by premultiplying first \( d \) leading eigenvectors of \( \sum_{h=1}^{H} \hat{m}_h \hat{m}_h^\top \) by \( \hat{\Sigma}^{-1/2} \).
If $Y \in \{-1, +1\}$ is binary:

- Only one possible choice to slice.

$$I_1 = \{i : y_i = -1\} \text{ and } I_2 = \{i : y_i = 1\}$$

- Associated $\bar{z}_1$ and $\bar{z}_2$ are linearly dependent since $\bar{z}_n = 0$.

$\Rightarrow$ SIR can estimate at most \textbf{ONE} direction.
Illustration to Wisconsin Diagnostic Breast Cancer Data

(a) SIR ($Y$ vs. $\hat{b}_1^\top X$)  

(b) SAVE ($\hat{b}_1^\top X$ vs. $\hat{b}_2^\top X$)
Outline

1 Introduction

2 Weighted Principal Support Vector Machine

3 Kernel Weighted PSVM

4 Numerical Results

5 Summary
For $(Y, X) \in \mathbb{R} \times \mathbb{R}^p$,

- PSVMs (Li et al., 2011; AOS) solve the following SVM-like problem:

$$
(a_0, c, b_0, c) = \arg\min_{a, b} \left\{ \frac{b^T \Sigma b}{\text{Var}(b^T X)} + \lambda \mathbb{E} \left[ 1 - \tilde{Y}_c (a + b^T (X - \mathbb{E}X)) \right]_+ \right\}.
$$

- $\tilde{Y}_c = \mathbb{1}\{Y \geq c\} - \mathbb{1}\{Y < c\}$ for a given constant $c$.
- $\Sigma = \text{cov}(X)$
- $[u]_+ = \max(0, u)$.

**Foundation of the PSVM**

Under linearity condition, $b_{0,c} \in \mathcal{S}_{Y|X}$ for any given $c$. 
Given a set of data \((X_i, Y_i), i = 1, \cdots, n:\)

1. For a given grid \(\min Y_i < c_1 < \cdots < c_H < \max Y_i\), solve a sequence of PSVMs for different values of \(c_h:\)

\[
(\hat{a}_{n,h}, \hat{b}_{n,h}) = \argmin_{a,b} b^\top \hat{\Sigma}_n b + \frac{\lambda}{n} \sum_{i=1}^{n} \left[ 1 - \tilde{Y}_{i,c_h}(a + b^\top (X_i - \bar{X}_n)) \right]_+.
\]

2. First \(k\) leading eigenvectors of

\[
\hat{M}_n^L = \sum_{h=1}^{H} \hat{b}_{n,h} \hat{b}_{n,h}^\top.
\]

estimate the basis set of \(S_Y|X\).
PSVM: Remarks

Pros:
- Outperforms SIR.
- Can be extended to kernel PSVM to handle nonlinear SDR.

Cons:
- Estimates only one direction if $Y$ is binary.
Weighted Principal Support Vector Machines

- Toward SDR with binary $Y$, WPSVM minimizes

$$
\Lambda_{\pi}(\theta) = \beta^\top \Sigma \beta + \lambda E \left\{ \pi(Y) \left[ 1 - Y \{\alpha + \beta^\top (X - EX)\} \right]_+ \right\}.
$$

- $\theta = (\alpha^\top, \beta^\top) \in \mathbb{R} \times \mathbb{R}^p$.
- $\pi(Y) = 1 - \pi$ if $Y = 1$ and $\pi$ otherwise for a given $\pi \in (0, 1)$.
- $Y$ itself is binary (no need $\tilde{Y}_c$).

- $\theta_{0,\pi} = (\alpha_{0,\pi}, \beta_{0,\pi})^\top = \operatorname{arg\,min}_\theta \Lambda_{\pi}(\theta)$.

Foundation of the Weighted PSVM

Under linearity condition, $\beta_{0,\pi} \in S_{Y\mid X}$ for any given $\pi \in (0, 1)$. 
Sample Estimation

Given \((X_i, Y_i) \in \mathbb{R}^p \times \{+1, -1\}, i = 1, \cdots, n:\)

1. For a given grid of \(\pi, 0 < \pi_1 < \cdots < \pi_H < 1\), solve a sequence of WPSVMs

\[
\hat{\Lambda}_{n,\pi_h}(\theta) = \beta^\top \hat{\Sigma}_n \beta + \frac{\lambda}{n} \sum_{i=1}^{n} \pi_h(Y_i)[1 - Y_i(\beta^\top (X_i - \bar{X}_n))]_+,
\]

and let \(\hat{\theta}_{n,h} = (\hat{\alpha}_{n,h}, \hat{\beta}_{n,h})^\top = \text{argmin}_\theta \hat{\Lambda}_{n,\pi_h}(\theta).\)

2. First \(k\) leading eigenvectors of the WPSVM candidate matrix

\[
\hat{M}_{WL}^{W} = \sum_{h=1}^{H} \hat{\beta}_{n,h} \hat{\beta}_{n,h}^\top
\]

estimate the basis set of \(S_{Y|X}.\)
Computation

- Let
  - $\eta = \hat{\Sigma}^{1/2} \beta$.
  - $U_i = \hat{\Sigma}^{-1/2} (X_i - \bar{X}_n)$.
- The WPSVM objective function $\hat{\Lambda}_{n,\pi_h}(\theta)$ becomes
  \[
  \eta^\top \eta + \frac{\lambda}{n} \sum_{i=1}^{n} \pi_h(Y_i) \left[ 1 - Y_i (\alpha + \eta^\top U_i) \right]^+. 
  \]
  \[\Rightarrow \text{ Equivalent to solve the linear WSVM w.r.t } (U_i, Y_i).\]
- Solve WSVM $H$ times for different weights of $\pi_h, h = 1, \cdots, H$. 

Yichao Wu
Weighted Principal Support Vector Machine
Wang et al. (2008, Biometrika) show that the WSVM solutions move piecewise-linearly as a function of $\pi$.

Shin et al. (2012+, JCGS) implemented the $\pi$-path algorithm in R while developing a two-dimensional solution surface for weighted SVMs.
Asymptotic Results (1)

- Standard approach based on M-estimation scheme.
- Similar to the results for the linear SVM:
  - Koo et al., 2008; JMLR
  - Jiang et al., 2008; JMLR

**Consistency of $\hat{\theta}_n$**

Suppose $\Sigma$ is positive definite,

$$\hat{\theta}_n \to \theta_0 \quad \text{in probability}.$$
Asymptotic Normality of \( \hat{\theta}_n \) (A Bahadur Representation)

Under some regularity conditions to ensure the existence of both Gradient vector \( D_\theta \) and Hessian matrix \( H_\theta \) of \( \Lambda_{\pi}(\theta) \),

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = -n^{-1/2}H_{\theta_0}^{-1} \sum_{i=1}^{n} D_{\theta_0}(Z_i) + o_p(1),
\]

where

\[
D_{\theta}(Z) = (0, 2\Sigma\beta)^\top - \lambda[\pi(Y)\tilde{X}Y\mathbb{1}\{\theta^\top\tilde{X}Y < 1\}] \quad \text{and} \quad H_\theta = 2\text{diag}(0, \Sigma) +
\lambda \sum_{y=-1,1} P(Y = y)\pi(y)f_{\beta^\top X|Y}(y - \alpha|y)E(\tilde{X}\tilde{X}^\top|\theta^\top\tilde{X} = y),
\]

with \( \tilde{X} = (1, X^\top)^\top \).
Asymptotic Results (3)

For a given grid of $\pi_1 < \cdots < \pi_H$, we define the population WPSVM kernel matrix

$$M_{0}^{WL} = \sum_{h=1}^{H} \beta_{0,h} \beta_{0,h}^{\top}.$$ 

Asymptotic Normality of $\hat{M}_n$

Suppose $\text{rank}(M_{0}^{WL}) = k$. Under the regularity conditions,

$$\sqrt{n} \left\{ \text{vec}(\hat{M}_n^{WL}) - \text{vec}(M_{0}^{WL}) \right\} \sim N(0, \Sigma_M),$$

where $\Sigma_M$ is explicitly provided.

- Asymptotic normality of eigenvectors of $\hat{M}_n^{WL}$ is followed by the normality of $\hat{M}_n$. (Bura & Pfeiffer, 2008)
Structure Dimensionality

$k$ Selection

We estimate $k$ as:

$$
\hat{k} = \arg\max_{k \in \{1, \ldots, p\}} \sum_{j=1}^{k} v_j - \rho \frac{k \log n}{\sqrt{n}} v_1,
$$

where $v_1 \geq \cdots \geq v_p$ are eigenvalues of $\hat{M}_n$. Then

$$
\lim_{n \to \infty} P(\hat{k} = k) = 1.
$$
Nonlinear SDR

Nonlinear SDR assumes

\[ Y \perp X | \phi(X). \]

- \( \phi : \mathbb{R}^p \mapsto \mathbb{R}^k \) is an arbitrary function of \( X \) which lives on \( \mathcal{H} \), a Hilbert space of functions of \( X \).
- SDR is achieved by estimating \( \phi \).
Kernel WPSVM: Objective Function

- Kernel WPSVM objective function is
  \[ \Lambda_{\pi}(\alpha, \psi) = \text{var}(\psi(X)) + \lambda E\{\pi(Y)[1 - Y(a + \psi(X) - E\psi(X))]_+}\]  

- Kernel WPSVM solves
  \[ (\alpha_{0,\pi}, \psi_{0,\pi}) = \arg\min_{\alpha \in \mathbb{R}, \psi \in \mathcal{H}} \Lambda_{\pi}(\alpha, \psi). \]
Foundation of the Kernel WPSVM

For a given $\pi$, $\psi_{0,\pi}$ has a version that is $\sigma\{\phi(X)\}$-measurable.

- Roughly speaking, $\psi_{0,\pi}$ is a function of $\phi$.
- It is a nonlinear-generalization of linear SDR:

$$\beta_{0,\pi} \in S_{Y|X} = \text{span}(B) \iff \beta_{0,\pi}^\top X \text{ is a linear function of } B^\top X.$$
Use Reproducing Kernel Hilbert Space.

Using a linear operator $\Sigma : \langle \psi_1, \Sigma \psi_2 \rangle_\mathcal{H} = \text{cov}\{\psi_1(X), \psi_2(X)\}$,

$$\Lambda_\pi(\alpha, \psi) = \langle \psi, \Sigma \psi \rangle_\mathcal{H} + \lambda \mathbb{E} \{\pi(Y)[1 - Y(a + \psi(X) - \mathbb{E}\psi(X))]_+\}.$$

Li et al. (2011) proposed to use the first $d$ leading eigenfunctions of the operator $\Sigma_n : \mathcal{H} \mapsto \mathcal{H}$ such that

$$\langle \psi_1, \Sigma_n \psi_2 \rangle_\mathcal{H} = \text{cov}_n(\psi_1(X), \psi_2(X)),$$

as a basis set.

By proposition 2 in Li et al. (2011), $\omega_j(X), j = 1, \cdots, d$ can be readily obtained by eigen-decomposition of $(I_n - J_n)K_n(I_n - J_n)$.

We chose $d \approx n/4$. 
Kernel WPSVM: Sample Estimation

- Sample version of $\Lambda_\pi(\alpha, \psi)$ is

$$\hat{\Lambda}_{n,\pi}(\alpha, \gamma) = \gamma^T \Omega^T \Omega \gamma + \lambda \sum_{i=1}^{n} \pi(Y_i) \left[ 1 - Y_i \{ \alpha + \gamma^T \Omega_i \} \right]_+.$$ 

- $\omega_1, \cdots, \omega_d$ be the first $d$ leading eigenfunctions of the operator $\Sigma_n$. Then,

$$\Omega = \begin{bmatrix} \omega_1^*(X_1) & \cdots & \omega_d^*(X_1) \\ \vdots & \ddots & \vdots \\ \omega_1^*(X_n) & \cdots & \omega_d^*(X_n) \end{bmatrix}$$

where $\omega_j^*(X) = \omega_j(X) - n^{-1} \sum_{i=1}^{n} \omega_j(X_i)$. 

Yichao Wu

Weighted Principal Support Vector Machine
Kernel WPSVM: Dual Problem

Dual Formulation

\[ \hat{\nu} = \arg\max_{\nu_1, \ldots, \nu_n} \sum_{i=1}^{n} \nu_i - \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_i \nu_j Y_i Y_j P_{(i,j)}^{(\Omega)} \]

subject to

i) \[ 0 \leq \nu_i \leq \lambda \pi(Y_i), i = 1, \ldots, n \]

ii) \[ \sum_{i=1}^{n} \nu_i Y_i = 0 \]

where \( P_{(i,j)}^{(\Omega)} \) is the \((i,j)\)th element of \( P_{\Omega} = \Omega(\Omega^\top \Omega)^{-1} \Omega^\top \).

The kernel WPSVM solution is given by

\[ \hat{\gamma}_n = \frac{\lambda}{2} \sum_{i=1}^{n} \hat{\nu}_i Y_i \{ (\Omega^\top \Omega)^{-1} \Omega_i \}. \]
1. For a given grid $\pi_1 < \cdots < \pi_H$, we compute a sequence of kernel WPSVM solutions:

$$(\hat{\alpha}_{n,h}, \hat{\gamma}_{n,h}) = \arg\min_{\alpha, \gamma} \hat{\Lambda}_{n,\pi_h}(\alpha, \gamma).$$

2. Corresponding kernel matrix is

$$\sum_{h=1}^{H} \hat{\gamma}_{n,h} \hat{\gamma}_{n,h}^\top.$$ (2)

3. Let $\hat{V}_n = (\hat{v}_1, \cdots, \hat{v}_k)$ denote the first $k$ leading eigenvectors of (2),

$$\hat{\phi}(x) = \hat{V}_n^\top(\omega_1(x), \cdots, \omega_d(x)).$$
Simulation - Set Up

- \( \mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^\top \sim N_p(0, \mathbf{I}), \ i = 1, \cdots, n \) where \((n, p) = (500, 10)\).

- We consider 5 Models:
  - Model I: \( Y = \text{sign}\{X_1/[0.5 + (X_2 + 1)^2] + 0.2\epsilon\} \).
  - Model II: \( Y = \text{sign}\{(X_1 + 0.5)(X_2 - 0.5)^2 + 0.2\epsilon\} \).
  - Model III: \( Y = \text{sign}\{\sin(X_1)/e^{X_2} + 0.2\epsilon\} \).
  - Model IV: \( Y = \text{sign}\{X_1(X_1 + X_2 + 1) + 0.2\epsilon\} \).
  - Model V: \( Y = \text{sign}\{(X_1^2 + X_2^2)^{1/2} \log(X_1^2 + X_2^2)^{1/2} + 0.2\epsilon\} \).

- \( \mathbf{B} = (\mathbf{e}_1, \mathbf{e}_2) \) s.t. \( \mathbf{e}_i^\top \mathbf{X} = X_i, i = 1, 2 (k = 2) \).

- Performance is measured by

\[
\| \mathbf{P}_{\hat{\mathbf{B}}} - \mathbf{P}_\mathbf{B} \|_F,
\]

where \( \mathbf{P}_\mathbf{A} = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \) and \( \| \cdot \|_F \) denotes Frobenius norm.
Introduction
Weighted Principal Support Vector Machine
Kernel Weighted PSVM
Numerical Results
Summary

True Classification Function

Figure: Surface plots of the Model IV and V.
Table: Averaged F-distance measures over 100 independent repetitions with associated standard deviations in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>SAVE</th>
<th>pHd</th>
<th>Fourier</th>
<th>IHT</th>
<th>LWPSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1.285</td>
<td>1.542</td>
<td>1.289</td>
<td>1.316</td>
<td>0.695</td>
</tr>
<tr>
<td></td>
<td>(.161)</td>
<td>(.193)</td>
<td>(.156)</td>
<td>(.254)</td>
<td>(.171)</td>
</tr>
<tr>
<td>II</td>
<td>1.265</td>
<td>1.383</td>
<td>1.205</td>
<td>1.140</td>
<td>0.896</td>
</tr>
<tr>
<td></td>
<td>(.187)</td>
<td>(.186)</td>
<td>(.214)</td>
<td>(.199)</td>
<td>(.198)</td>
</tr>
<tr>
<td>III</td>
<td>1.255</td>
<td>1.491</td>
<td>1.282</td>
<td>1.295</td>
<td>0.688</td>
</tr>
<tr>
<td></td>
<td>(.186)</td>
<td>(.198)</td>
<td>(.163)</td>
<td>(.232)</td>
<td>(.180)</td>
</tr>
<tr>
<td>IV</td>
<td>0.771</td>
<td>0.680</td>
<td>0.469</td>
<td>0.474</td>
<td>0.482</td>
</tr>
<tr>
<td></td>
<td>(.272)</td>
<td>(.194)</td>
<td>(.103)</td>
<td>(.105)</td>
<td>(.101)</td>
</tr>
<tr>
<td>V</td>
<td>0.273</td>
<td>0.283</td>
<td>0.492</td>
<td>1.424</td>
<td>1.530</td>
</tr>
<tr>
<td></td>
<td>(.052)</td>
<td>(.053)</td>
<td>(.241)</td>
<td>(.011)</td>
<td>(.171)</td>
</tr>
</tbody>
</table>
### Results - Structure Dimensionality

<table>
<thead>
<tr>
<th>Model</th>
<th>( k )</th>
<th>( n )</th>
<th>( p = 10 ) ( p = 20 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAVE</td>
</tr>
<tr>
<td>( f'_1 )</td>
<td>1</td>
<td>500</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>92</td>
</tr>
<tr>
<td>( f_1 )</td>
<td>2</td>
<td>500</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>15</td>
</tr>
<tr>
<td>( f'_2 )</td>
<td>1</td>
<td>500</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>93</td>
</tr>
<tr>
<td>( f_2 )</td>
<td>2</td>
<td>500</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>13</td>
</tr>
<tr>
<td>( f'_3 )</td>
<td>1</td>
<td>500</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>91</td>
</tr>
<tr>
<td>( f_3 )</td>
<td>2</td>
<td>500</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>16</td>
</tr>
</tbody>
</table>

**Table:** Empirical probabilities (in percentage) of correctly estimating true \( k \) based on 100 independent repetitions.

**SAVE:** the permutation test (Cook and Yin, 2001).
Results - Kernel WPSVM

(a) Original

(b) SAVE

(c) Linear WPSVM

**Figure:** Nonlinear SDR results for a random data set from Model V.
Results - Kernel WPSVM

(a) Kernel WPSVM(\(\hat{\phi}_1(X)\) vs. \(\hat{\phi}_2(X)\))

(b) \(\hat{\phi}_1(X)\) vs. \((X_1^2 + X_2^2)^{1/2}\)

Figure: Kernel WPSVM results for a random data set from Model V.
Two-sample Hotelling’s $T^2$ test statistics:

\[
T_n^2 = (\bar{X}_+ - \bar{X}_-)^\top \left\{ \hat{\Sigma}_n \left( \frac{1}{n_+} + \frac{1}{n_-} \right) \right\}^{-1} (\bar{X}_+ - \bar{X}_-).
\]

**Table:** Averaged $T_n^2$ computed from the first two estimated sufficient predictors over 100 independent repetitions.

<table>
<thead>
<tr>
<th>Model</th>
<th>SAVE</th>
<th>pHd</th>
<th>FCN</th>
<th>IHT</th>
<th>LWPSVM</th>
<th>KWPSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>76.0</td>
<td>74.0</td>
<td>104.0</td>
<td>97.2</td>
<td>103.8</td>
<td>581.7</td>
</tr>
<tr>
<td></td>
<td>(30.7)</td>
<td>(20.6)</td>
<td>(25.7)</td>
<td>(24.5)</td>
<td>(25.7)</td>
<td>(71.9)</td>
</tr>
<tr>
<td>V</td>
<td>1.2</td>
<td>1.1</td>
<td>4.0</td>
<td>8.7</td>
<td>8.8</td>
<td>626.0</td>
</tr>
<tr>
<td></td>
<td>(1.2)</td>
<td>(1.1)</td>
<td>(4.2)</td>
<td>(4.5)</td>
<td>(4.6)</td>
<td>(78.1)</td>
</tr>
</tbody>
</table>
WDBC data - SDR results

(a) $k$-selection

(b) Linear WPSVM

(c) Kernel WPSVM
3-NN test error rate for the raw data: 7.7% (1.2)

<table>
<thead>
<tr>
<th>$k$</th>
<th>SAVE</th>
<th>pHd</th>
<th>FCN</th>
<th>IHT</th>
<th>LWPSVM</th>
<th>KWPSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.3</td>
<td>39.7</td>
<td>12.7</td>
<td>23.0</td>
<td>5.3</td>
<td>8.7</td>
</tr>
<tr>
<td></td>
<td>(2.2)</td>
<td>(4.5)</td>
<td>(2.3)</td>
<td>(3.1)</td>
<td>(1.1)</td>
<td>(2.0)</td>
</tr>
<tr>
<td>2</td>
<td>13.5</td>
<td>34.5</td>
<td>9.2</td>
<td>13.0</td>
<td>5.2</td>
<td>8.5</td>
</tr>
<tr>
<td></td>
<td>(1.8)</td>
<td>(4.8)</td>
<td>(2.0)</td>
<td>(2.8)</td>
<td>(1.1)</td>
<td>(2.0)</td>
</tr>
<tr>
<td>3</td>
<td>12.0</td>
<td>30.3</td>
<td>8.0</td>
<td>6.9</td>
<td>5.4</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>(1.7)</td>
<td>(5.0)</td>
<td>(1.8)</td>
<td>(1.5)</td>
<td>(1.2)</td>
<td>(2.0)</td>
</tr>
<tr>
<td>4</td>
<td>11.9</td>
<td>27.0</td>
<td>7.5</td>
<td>5.7</td>
<td>5.4</td>
<td>7.8</td>
</tr>
<tr>
<td></td>
<td>(1.6)</td>
<td>(4.6)</td>
<td>(1.9)</td>
<td>(1.4)</td>
<td>(1.3)</td>
<td>(1.9)</td>
</tr>
<tr>
<td>5</td>
<td>12.1</td>
<td>25.2</td>
<td>7.2</td>
<td>5.8</td>
<td>5.5</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>(1.8)</td>
<td>(4.2)</td>
<td>(1.8)</td>
<td>(1.3)</td>
<td>(1.5)</td>
<td>(1.9)</td>
</tr>
</tbody>
</table>

Table: Averaged test error rates (in percentage) of the kNN classifier ($\kappa = 3$) over 100 random partitions for the WDBC data with respect to the first $k$ sufficient predictors ($k = 1, 2, 3, 4, 5$), which are estimated by different SDR methods. Corresponding standard deviations are given in parentheses.
Outline

1. Introduction
2. Weighted Principal Support Vector Machine
3. Kernel Weighted PSVM
4. Numerical Results
5. Summary
Summary

- Most existing SDR methods suffer if $Y$ is binary.
- The proposed WPSVM preserves all the merits of the PSVM and performs very well in binary classification.
- Computational efficiency can be improved by employing the $\pi$-path algorithm.
Thank you!!!
Selected References

For any \( a \in \mathbb{R}^p \), \( E(a^T X | B^T X) \) is a linear function of \( B^T X \).

\[ \iff E(X | B^T X) = P_\Sigma(B)X = B(B^T \Sigma B)^{-1}B^T \Sigma X \]

- Common and essential assumption in SDR.
- Hard to check since \( B \) is unknown.
- Holds if \( X \) is elliptically symmetric. (eg. \( X \) is multivariate normal)
- Approximately holds if \( p \) gets large for fixed \( d \). (Hall and Li, 1993)
- Assumption is only for the marginal distribution of \( X \).
1. Randomly split the data into the training and testing sets.

2. Apply the WPSVM to the training set and compute its candidate matrix, $\hat{M}^{tr}_n$.

3. For a given $\rho$,
   
   3.a Compute $\hat{k}_{tr} = \arg\max_{k \in \{1, \ldots, p\}} = G_n(k; \rho, \hat{M}^{tr}_n)$.
   
   3.b Transform training predictors $\tilde{X}^{tr}_{j'} = (\hat{V}^{tr}_n)^\top X^{tr}_{j'}$ where $\hat{V}^{tr}_n = (\hat{v}^{tr}_{1}, \ldots, \hat{v}^{tr}_{\hat{k}_{tr}})$ are the first $\hat{k}_{tr}$ leading eigenvectors of $\hat{M}^{tr}_n$.

   3.c For each $\pi_h, h = 1, \ldots, H$, apply the WSVM to
   
   $\{(\tilde{X}^{tr}_{j'}, Y^{tr}_{j'}) : j' = 1, \ldots, n_{tr}\}$ to predict $Y^{ts}_{j'}$.

   3.d Denoting the predicted label $\hat{Y}^{ts}_{j'}$, compute the total cost on the test data set.

   $$TC(\rho) = \sum_{h=1}^{H} \left\{ \sum_{j'=1}^{n_{ts}} \pi_h(Y^{ts}_{j'}) \cdot 1(\hat{Y}^{ts}_{j'} \neq Y^{ts}_{j'}) \right\}.$$  

4. Repeat 3.a–d to select $\rho^*$ which minimizes $TC(\rho)$. 