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

Outline

Sufficient Dimension Reduction

For a given pair of $(Y, \mathbf{X}) \in \mathbb{R} \times \mathbb{R}^p$,

- Sufficient Dimension Reduction (SDR) seeks a matrix $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_d) \in \mathbb{R}^{p \times d}$ which satisfies

$$Y \perp \mathbf{X} | \mathbf{B}^\top \mathbf{X}. \tag{1}$$

Central Subspace

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

Central Subspace

Central Subspace, $\mathcal{S}_{Y|\mathbf{X}}$ is the intersection of all DRSes.

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

Estimation of $\mathcal{S}_{Y|\mathbf{X}}$

Seminal paper in Early 1990.

- K-C Li (1991) Sliced Inverse Regression for Dimension Reduction (with discussion). JASA, 86, 316–327.
- 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,

$$\mathrm{E}(\mathbf{Z}|Y) \in \mathcal{S}_{Y|\mathbf{Z}} = \mathbf{\Sigma}^{1/2}\mathcal{S}_{Y|\mathbf{X}}.$$

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

- Slice response into H non-overlapping intervals, $I_1, \cdots, I_H$,

$$\hat{\mathbf{m}}_h := E_n(\mathbf{Z}|Y \in I_h) = \frac{1}{n_h}\sum_{Y \in I_h} \mathbf{z}_i, \quad h = 1, \cdots, H.$$

- $\mathbf{B}$ is estimated by premultiplying first d leading eigenvectors of $\underline{\sum_{h=1}^{H} \hat{\mathbf{m}}_h \hat{\mathbf{m}}_h^\top}$ by $\hat{\mathbf{\Sigma}}^{-1/2}$.

SIR with Binary Response

If $Y \in \{-1, +1\}$ is binary:

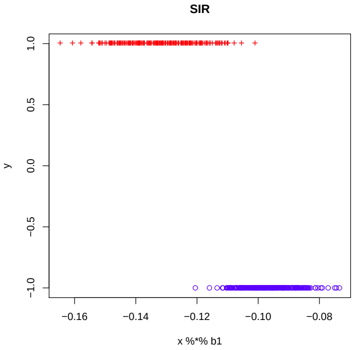
- Only one possible choice to slice.

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

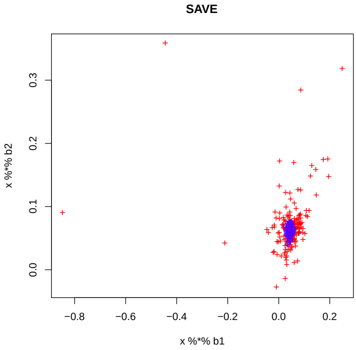
- Associated $\bar{\mathbf{z}}_1$ and $\bar{\mathbf{z}}_2$ are linearly dependent since $\bar{\mathbf{z}}_n = 0$.

$\Rightarrow$ SIR can estimate at most ONE direction.

Illustration to Wisconsin Diagnostic Breast Cancer Data



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



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

Outline

Principal Support Vector Machine

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

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

$$(a_{0,c}, \mathbf{b}_{0,c}) = \underset{a,\mathbf{b}}{\operatorname{argmin}} \underbrace{\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}}_{Var(\mathbf{b}^\top \mathbf{X})} + \lambda \mathrm{E}\big[1 - \tilde{Y}_c \underbrace{(a + \mathbf{b}^\top (\mathbf{X} - \mathrm{E}\mathbf{X}))}_{f(\mathbf{X})}\big]_+ .$$

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

Foundation of the PSVM

Under linearity condition, $\mathbf{b}_{0,c} \in \mathcal{S}_{Y|\mathbf{X}}$ for any given c.

PSVM: Sample Estimation

Given a set of data $(\mathbf{X}_i, Y_i), i = 1, \cdots, n$:

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:

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

2. First k leading eigenvectors of

$$\underline{\hat{\mathbf{M}}_n^L = \sum_{h=1}^{H} \hat{\mathbf{b}}_{n,h} \hat{\mathbf{b}}_{n,h}^\top .}$$

estimate the basis set of $\mathcal{S}_{Y|\mathbf{X}}$.

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

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

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

- $\boldsymbol{\theta}_{0,\pi} = (\alpha_{0,\pi}, \boldsymbol{\beta}_{0,\pi})^\top = \mathrm{argmin}_{\boldsymbol{\theta}} \Lambda_\pi(\boldsymbol{\theta})$.

Foundation of the Weighted PSVM

Under linearity condition, $\boldsymbol{\beta}_{0,\pi} \in \mathcal{S}_{Y|\mathbf{X}}$ for any given $\pi \in (0, 1)$.

Sample Estimation

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

1. For a given grid of π, $0 < \pi_1 < \cdots < \pi_H < 1$, solve a sequence of WPSVMs

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

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

2. First k leading eigenvectors of the WPSVM candidate matrix

$$\underline{\hat{\mathbf{M}}_n^{WL} = \sum_{h=1}^{H} \hat{\boldsymbol{\beta}}_{n,h} \hat{\boldsymbol{\beta}}_{n,h}^\top}$$

   estimate the basis set of $\mathcal{S}_{Y|\mathbf{X}}$.

Computation

- Let
  - $\boldsymbol{\eta} = \hat{\boldsymbol{\Sigma}}_n^{1/2} \boldsymbol{\beta}$.
  - $\mathbf{U}_i = \hat{\boldsymbol{\Sigma}}_n^{-1/2} (\mathbf{X}_i - \bar{\mathbf{X}}_n)$.
- The WPSVM objective function $\hat{\Lambda}_{n,\pi_h}(\boldsymbol{\theta})$ becomes

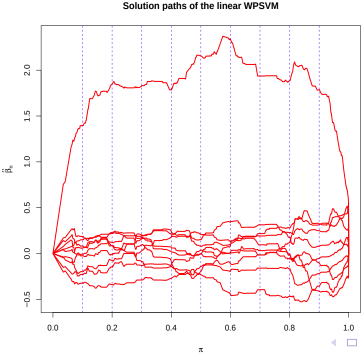
$$\boldsymbol{\eta}^\top \boldsymbol{\eta} + \frac{\lambda}{n} \sum_{i=1}^{n} \pi_h(Y_i) \left[1 - Y_i(\alpha + \boldsymbol{\eta}^\top \mathbf{U}_i)\right]_+ .$$

$\Rightarrow$ Equivalent to solve the linear WSVM w.r.t $(\mathbf{U}_i, Y_i)$.

- Solve WSVM H times for different weights of $\pi_h, h = 1, \cdots, H$.

π-path

- Wang et al. (2008, Biometrika) show that the WSVM solutions move piecewise-linearly as a function of π.
- Shin et al. (2012+, JCGS) implemented the π-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 $\hat{\boldsymbol{\theta}}_n$

Suppose $\boldsymbol{\Sigma}$ is positive definite,

$$\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0 \quad \text{in probability.}$$

Asymptotic Results (2)

Asymptotic Normality of $\hat{\boldsymbol{\theta}}_n$ (A Bahadur Representation)

Under some regularity conditions to ensure the existence of both Gradient vector $\mathbf{D}_{\boldsymbol{\theta}}$ and Hessian matrix $\mathbf{H}_{\boldsymbol{\theta}}$ of $\Lambda_\pi(\boldsymbol{\theta})$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -n^{-1/2}\mathbf{H}_{\boldsymbol{\theta}_0}^{-1} \sum_{i=1}^{n} \mathbf{D}_{\boldsymbol{\theta}_0}(\mathbf{Z}_i) + o_p(1),$$

where

$$\mathbf{D}_{\boldsymbol{\theta}}(\mathbf{Z}) = (0, 2\boldsymbol{\Sigma}\boldsymbol{\beta})^\top - \lambda[\pi(Y)\tilde{\mathbf{X}}Y\mathbb{1}\{\boldsymbol{\theta}^\top\tilde{\mathbf{X}}Y < 1\}] \text{ and}$$
$$\mathbf{H}_{\boldsymbol{\theta}} = 2diag(0, \boldsymbol{\Sigma}) + \lambda \sum_{y=-1,1} P(Y=y)\pi(y) f_{\boldsymbol{\beta}^\top\mathbf{X}|Y}(y-\alpha|y)\mathrm{E}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top|\boldsymbol{\theta}^\top\tilde{\mathbf{X}} = y),$$

with $\tilde{\mathbf{X}} = (1, \mathbf{X}^\top)^\top$.

Asymptotic Results (3)

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

$$\mathbf{M}_0^{WL} = \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h}\boldsymbol{\beta}_{0,h}^\top.$$

Asymptotic Normality of $\hat{\mathbf{M}}_n$

Suppose $rank(\mathbf{M}_0^{WL}) = k$. Under the regularity conditions,

$$\sqrt{n}\left\{\text{vec}(\hat{\mathbf{M}}_n^{WL}) - \text{vec}(\mathbf{M}_0^{WL})\right\} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{M}}),$$

where $\boldsymbol{\Sigma}_{\mathbf{M}}$ is explicitly provided.

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

Structure Dimensionality

k Selection

We estimate k as:

$$\hat{k} = \underset{k \in \{1, \cdots, p\}}{\operatorname{argmax}} \sum_{j=1}^{k} v_j - \rho \frac{k \log n}{\sqrt{n}} v_1,$$

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

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

Outline

Nonlinear SDR

Nonlinear SDR assumes

$$Y \perp \mathbf{X} | \phi(\mathbf{X}).$$

- $\phi : \mathbb{R}^p \mapsto \mathbb{R}^k$ is an arbitrary function of $\mathbf{X}$ which lives on $\mathcal{H}$, a Hilbert space of functions of $\mathbf{X}$.
- SDR is achieved by estimating ϕ.

Kernel WPSVM: Objective Function

- Kernel WPSVM objective function is

$$\Lambda_\pi(\alpha,\psi) = \mathsf{var}(\psi(\mathbf{X})) + \lambda \mathrm{E}\left\{\pi(Y)[1 - Y(a + \psi(\mathbf{X}) - \mathrm{E}\psi(\mathbf{X}))]_+\right\}$$

- Kernel WPSVM solves

$$(\alpha_{0,\pi},\psi_{0,\pi}) = \underset{\alpha\in\mathbb{R},\psi\in\mathcal{H}}{\operatorname{argmin}} \Lambda_\pi(\alpha,\psi).$$

Kernel WPSVM: Foundation

Foundation of the Kernel WPSVM

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

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

$$\boldsymbol{\beta}_{0,\pi} \in \mathcal{S}_{Y|\mathbf{X}} = \text{span}(\mathbf{B}) \;\Leftrightarrow\; \boldsymbol{\beta}_{0,\pi}^\top \mathbf{X} \text{ is a linear function of } \mathbf{B}^\top \mathbf{X}.$$

Kernel WPSVM: Sample Estimation

- Use Reproducing Kernel Hilbert Space.
- Using a linear operator $\Sigma : \langle \psi_1, \Sigma\psi_2 \rangle_{\mathcal{H}} = \mathrm{cov}\{\psi_1(\mathbf{X}), \psi_2(\mathbf{X})\}$,

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

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

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

  as a basis set.
- By proposition 2 in Li et al. (2011), $\omega_j(\mathbf{X}), j = 1, \cdots, d$ can be readily obtained by eigen-decomposition of $(\mathbf{I}_n - \mathbf{J}_n)\mathbf{K}_n(\mathbf{I}_n - \mathbf{J}_n)$.
- We chose $d \approx n/4$.

Kernel WPSVM: Sample Estimation

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

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

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

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega_1^*(\mathbf{X}_1) & \cdots & \omega_d^*(\mathbf{X}_1) \\ \vdots & \ddots & \vdots \\ \omega_1^*(\mathbf{X}_n) & \cdots & \omega_d^*(\mathbf{X}_n) \end{bmatrix}$$

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

Kernel WPSVM: Dual Problem

Dual Formulation

$$\hat{\boldsymbol{\nu}} = \underset{\nu_1,\cdots,\nu_n}{\operatorname{argmax}} \sum_{i=1}^{n} \nu_i - \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_i \nu_j Y_i Y_j P_{\boldsymbol{\Omega}}^{(i,j)}$$

subject to

i) $0 \leq \nu_i \leq \lambda \pi(Y_i), i = 1, \cdots, n$

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

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

- The kernel WPSVM solution is given by

$$\hat{\gamma}_n = \frac{\lambda}{2} \sum_{i=1}^{n} \hat{\nu}_i Y_i \{(\boldsymbol{\Omega}^\top \boldsymbol{\Omega})^{-1} \boldsymbol{\Omega}_i\}.$$

Kernel WPSVM: Summary

1. For a given gird $\pi_1 < \cdots < \pi_H$, we compute a sequence of kernel WPSVM solutions:

$$(\hat{\alpha}_{n,h}, \hat{\boldsymbol{\gamma}}_{n,h}) = \underset{\alpha,\boldsymbol{\gamma}}{\operatorname{argmin}}\, \hat{\Lambda}_{n,\pi_h}(\alpha, \boldsymbol{\gamma}).$$

2. Corresponding kernel matrix is

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

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

$$\hat{\phi}(\mathbf{x}) = \hat{\mathbf{V}}_n^{\top}(\omega_1(\mathbf{x}), \cdots, \omega_d(\mathbf{x})).$$

Outline

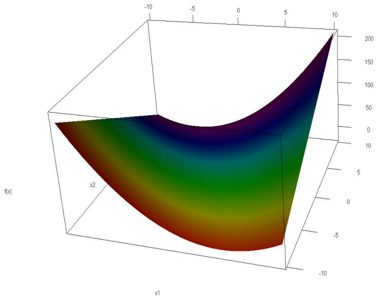
Simulation -Set Up

- $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^\top \sim N_p(\mathbf{0}, \mathbf{I})$, $i = 1, \cdots, n$ where $(n, p) = (500, 10)$.
- We consider 5 Models:
  - Model I: $Y = sign\{X_1/[0.5 + (X_2 + 1)^2] + 0.2\epsilon\}$.
  - Model II: $Y = sign\{(X_1 + 0.5)(X_2 - 0.5)^2 + 0.2\epsilon\}$.
  - Model III: $Y = sign\{\sin(X_1)/e^{X_2} + 0.2\epsilon\}$.
  - Model IV: $Y = sign\{X_1(X_1 + X_2 + 1) + 0.2\epsilon\}$.
  - Model V: $Y = sign\{(X_1^2 + X_2^2)^{1/2} \log(X_1^2 + X_2^2)^{1/2} + 0.2\epsilon\}$.
- $\mathbf{B} = (\mathbf{e}_1, \mathbf{e}_2)$ s.t. $\mathbf{e}_i^\top \mathbf{X} = X_i, i = 1, 2$ $(k = 2)$.
- Performance is measured by

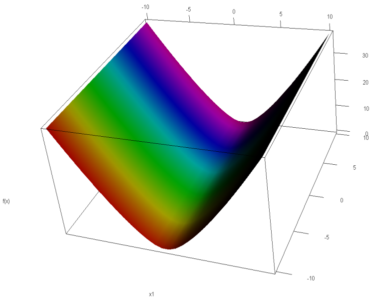
$$\|\mathbf{P}_{\hat{\mathbf{B}}} - \mathbf{P}_{\mathbf{B}}\|_F,$$

where $\mathbf{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top$ and $\|\cdot\|_F$ denotes Frobenius norm.

True Classification Function



(c) Model IV



(d) Model V

Figure: Surface plots of the Model IV and V.

Results - Linear WPSVM

Table: Averaged F-distance measures over 100 independent repetitions with associated standard deviations in parentheses.

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

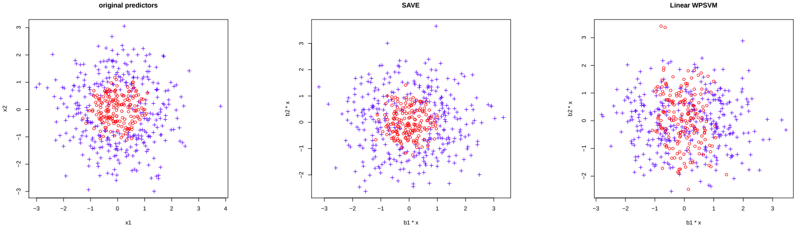
Results - Structure Dimensionality

Model	k	n	$p = 10$		$p = 20$	
			SAVE	WPSVM	SAVE	WPSVM
f_1'	1	500	91	84	82	86
		1000	92	95	93	92
f_1	2	500	7	66	6	40
		1000	15	98	16	74
f_2'	1	500	80	95	41	71
		1000	93	93	88	85
f_2	2	500	17	42	15	19
		1000	13	72	17	54
f_3'	1	500	87	89	86	91
		1000	91	98	90	81
f_3	2	500	15	56	7	44
		1000	16	86	11	74

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

SAVE: the permutation test (Cook and Yin, 2001).

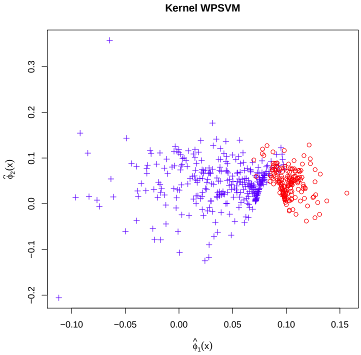
Results - Kernel WPSVM



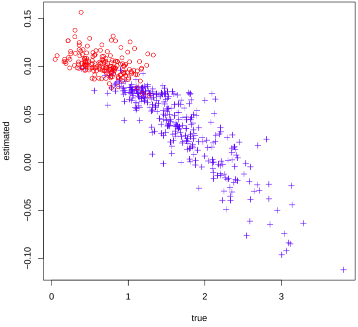
(a) Original (b) SAVE (c) Linear WPSVM

Figure: Nonlinear SDR results for a random data set from Model V.

Results - Kernel WPSVM



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



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

Figure: Kernel WPSVM results for a random data set from Model V.

Results - Kernel WPSVM

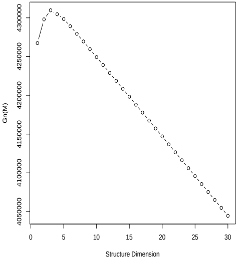
Two-sample Hotelling's T^2 test statistics:

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

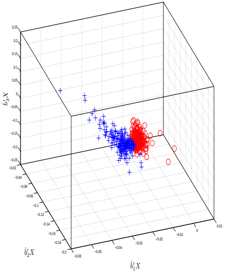
Table: Averaged T_n^2 computed from the first two estimated sufficient predictors over 100 independent repetitions.

Model	SAVE	pHd	FCN	IHT	LWPSVM	KWPSVM
IV	76.0	74.0	104.0	97.2	103.8	581.7
	(30.7)	(20.6)	(25.7)	(24.5)	(25.7)	(71.9)
V	1.2	1.1	4.0	8.7	8.8	626.0
	(1.2)	(1.1)	(4.2)	(4.5)	(4.6)	(78.1)

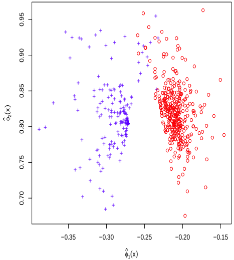
WDBC data - SDR results



(a) k-selection



(b) Linear WPSVM



(c) Kernel WPSVM

WDBC data - Classification Accuracy

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

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

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

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 π-path algorithm.

Thank you!!!

Selected References

- Li, K.-C. (1991), "Sliced inverse regression for dimension reduction (with discussion)," Journal of the American Statistical Association, 86, 316–342.
- Li, B., Artemiou, A., and Li, L. (2011), "Principal Support Vector Machines for Linear and Nonlinear Sufficient Dimension Reduction," Annals of Statistics, 39, 3182–3210.
- Wang, J., Shen, X., and Liu, Y. (2008), "Probability estimation for large-margin classifier," Biometrika, 95, 149–167.
- Shin, S.J., Wu, Y., and Zhang, H.H. (2013), "Two dimensional solution surface of the weighted support vector machines," Journal of Computational and Graphical Statistics, to appear.
- Koo, J.-Y., Lee, Y., Kim, Y., and Park, C. (2008) "A Bahadur Representation of the Linear Support Vector Machine," Journal of Machine Learning Research, 9(Jul):1343–1368.
- Jiang, B., Zhang, X., and Cai, T. (2008) "Estimating the Confidence Interval for Prediction Errors of Support Vector Machine Classifiers", Journal of Machine Learning Research, 9(Mar):521–540.
- Bura, E. and Pfeiffer, C. (2008), "On the distribution of the left singular vectors of a random matrix and its applications," Statistics and Probability Letters, 78, 2275–2280.

Linearity Condition

- For any $\mathbf{a} \in \mathbb{R}^p$, $\mathrm{E}(\mathbf{a}^\top \mathbf{X} | \mathbf{B}^\top \mathbf{X})$ is a linear function of $\mathbf{B}^\top \mathbf{X}$.

$\Leftrightarrow$ $\mathrm{E}(\mathbf{X} | \mathbf{B}^\top \mathbf{X}) = P_{\mathbf{\Sigma}}(\mathbf{B})\mathbf{X} = \mathbf{B}(\mathbf{B}^\top \mathbf{\Sigma} \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{\Sigma} \mathbf{X}$

- Common and essential assumption in SDR.
- Hard to check since $\mathbf{B}$ is unknown.
- Holds if $\mathbf{X}$ is elliptically symmetric. (eg. $\mathbf{X}$ is multivariate normal)
- Approximately holds if p gets large for fixed d. (Hall and Li, 1993)
- Assumption is only for the marginal distribution of $\mathbf{X}$.

ρ Selection

1. Randomly split the data into the training and testing sets.
2. Apply the WPSVM to the training set and compute its candidate matrix, $\widehat{\mathbf{M}}_n^{\text{tr}}$.
3. For a given ρ,
   - 3.a Compute $\hat{k}_{\text{tr}} = \text{argmax}_{k \in \{1, \cdots, p\}} = G_n(k; \rho, \widehat{\mathbf{M}}_n^{\text{tr}})$.
   - 3.b Transform training predictors $\tilde{\mathbf{X}}_{j'}^{\text{tr}} = (\widehat{\mathbf{V}}_n^{\text{tr}})^\top \mathbf{X}_{j'}^{\text{tr}}$ where $\widehat{\mathbf{V}}_n^{\text{tr}} = (\widehat{\mathbf{v}}_1^{\text{tr}}, \cdots, \widehat{\mathbf{v}}_{\hat{k}_{\text{tr}}}^{\text{tr}})$ are the first $\hat{k}_{\text{tr}}$ leading eigenvectors of $\widehat{\mathbf{M}}_n^{\text{tr}}$.
   - 3.c For each $\pi_h, h = 1, \cdots, H$, apply the WSVM to $\{(\tilde{\mathbf{X}}_{j'}^{\text{tr}}, Y_{j'}^{\text{tr}}) : j' = 1, \cdots, n_{\text{tr}}\}$ to predict $Y_{j'}^{\text{ts}}$.
   - 3.d Denoting the predicted label $\hat{Y}_{j'}^{\text{ts}}$, compute the total cost on the test data set.

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

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