

# Sparse Learning for Support Vector Machines

*Hao Helen Zhang*

*North Carolina State University*

## Outline

1. Classification via SVM
2. High dimensional low sample size data
3. Variable selection for SVM
  - Regularization for sparsity
  - SCAD penalty for SVM
  - Algorithm & Examples
4. Variable selection for Multiclass SVM
  - Supnorm MSVM
  - Examples

## Classification Problems

- Training data:  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$
- Input vector  $\mathbf{x}_i \in \mathcal{R}^p$
- Class label  $y_i \in \{1, \dots, K\}$ . For binary case,  $y_i \in \{-1, +1\}$ .
- High dimensional setting:  $p \gg n$
- Acute leukemia data (Golub et al.1999):  $p = 7129, n = 72, k = 3$ .
- To learn a classification rule for making future predictions

$$\phi(\mathbf{x}) : \mathcal{R}^p \rightarrow \{1, \dots, K\}.$$

## Regularization Problems

Minimize the loss function with regularization:  $L(\mathbf{y}, f) + \lambda J(f)$

- $L(y, f(\mathbf{x}))$  is a convex upper bound of the classification error function
- Support Vector Machine:

$$\min_{f \in \mathcal{H}} \sum_{i=1}^n [1 - y_i f(\mathbf{x}_i)]_+ + \lambda \|f\|_{\mathcal{H}}^2.$$

- SVM asymptotically implements the Bayes rule (Lin, 2002)
- Linear SVM minimizes  $\sum_{i=1}^n [1 - y_i (b + \mathbf{w} \cdot \mathbf{x}_i)]_+ + \lambda \|\mathbf{w}\|^2$ .
- Other loss functions and their consistency (Zhang 2004)
  - exponential loss  $\exp(-yf)$  (AdaBoost), logistic regression
  - least squares loss  $(1 - yf)^2$  (Proximal SVM), modified least squares  $(1 - yf)_+^2$

## Linear SVM for High Dimensional Data

When  $p \gg n$ , linear classifiers often give better performance than nonlinear classifiers in many applications.

- Hall et al. (2005) gives “geometric representation of high dimensional low sample size data”: under a mild assumption for data distribution, when  $p \rightarrow \infty$ , all the pairwise distances between any two points are approximately identical to each other. Or, after proper scaling ( $p^{1/2}$ ), all the  $n$  points are asymptotically located at the vertices of a convex  $n$ -polyhedron.
- Similar observations made for text mining (Ju, Madigan, Scott 2002)



## SVM for Gene Expression Data

SVM has been successfully applied to cancer classification using microarray data.

- However, its decision rule utilizes all the genes without discrimination. The accuracy may suffer from the existence of redundant genes (Hastie et al. 2001, Guyon 2002)

Wish to identify informative genes, which are used to construct classifiers:

- better classification accuracy
- better interpretation of classifier and individual genes
  - to understand genetic signatures in cancers (Kitter, 1986)
  - to improve treatment strategies, e.g., drug discovery

## Existing Methods for Sparse Learning

Two types of approaches:

- Filtering approach: a preprocessing step to remove irrelevant genes before training classifiers
- Wrapper approach: simultaneous classification and feature selection

Filtering approach:

- gene ranking: select genes by their ranks, then train a classifier.
- dimension reduction approach: project the full data onto the first few principal directions and then conduct classification in the low dimensional subspace. (West 2003, “meta-genes”).

We focus on the regularization approach with shrinkage penalty for linear SVM

## Variable Selection in SVM

- 1-norm SVM: Bradley and Mangasarian (1998), Zhu et al. (2003)
  - Zhu et al. (2003) derived the whole solution path, facilitating the choice of tuning parameter
- recursive feature elimination (Guyon et al. 2000; Rakotomamonjy 2003)
- maximum entropy discrimination (Jebara and Jaakkola 2000)
- kernel scaling (Weston et al. 2000; Grandvalet and Canu 2002)
- block 1-norm regularization (Bach et al. 2004)

## Various Shrinkage Penalties

$L_q$  penalty: (Frank and Friedman, 1993)

$$J_q(|\mathbf{w}|) = \|\mathbf{w}\|_q^q = \sum_{j=1}^p |w_j|^q, \quad q \geq 0.$$

- $J_0(|\mathbf{w}|) = \sum_{j=1}^p I(w_j \neq 0)$  ( $L_0$  penalty)
- $J_1(|\mathbf{w}|) = \sum_{j=1}^p |w_j|$  (LASSO penalty)
- $J_2(|\mathbf{w}|) = \sum_{j=1}^p w_j^2$  (ridge penalty; standard SVM)
- $J_\infty(|\mathbf{w}|) = \max_j |w_j|$  (supnorm penalty)

Sparsity/Thresholding occurs only when  $q \leq 1$ . (Tibshirani 1996)

## Penalized Least Squares (Orthonormal Design)

Consider a regression model

$$\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

with  $y_i$ 's conditionally independent given  $X$ . Assume the columns of  $X$  are orthonormal. Define  $\mathbf{z} = X^T \mathbf{y}$ ,  $\hat{\mathbf{y}} = XX^T \mathbf{y}$ . Then

$$\frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^p J_q(|\theta_j|) = \frac{1}{2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \frac{1}{2} \sum_{j=1}^p (z_j - \theta_j)^2 + \lambda \sum_{j=1}^p J_q(|\theta_j|)$$

Only need component-wise minimization:

$$\min \frac{1}{2} (z - \theta)^2 + \lambda J_q(|\theta|)$$

## Characteristics of Penalty Functions

simple regression:  $\min \frac{1}{2}(z - \theta)^2 + \lambda J_q |\theta|$

- $L_0$  penalty: the solution  $\hat{\theta} = zI(|z| > \lambda)$ 
  - implements best subset selection
  - hard to optimize in practice, neither continuous nor convex.
- $L_1$  penalty: the solution  $\hat{\theta} = \text{sign}(z)(|z| - \lambda)_+$ 
  - penalizing each nonzero component by a constant factor
  - sparsity:  $\hat{\theta} = 0$  if  $|z| \leq \lambda$ .
- $L_2$  penalty:  $\hat{\theta} = \frac{z}{1 + 2\lambda}$ .
  - penalizing each component by a fraction of its magnitude

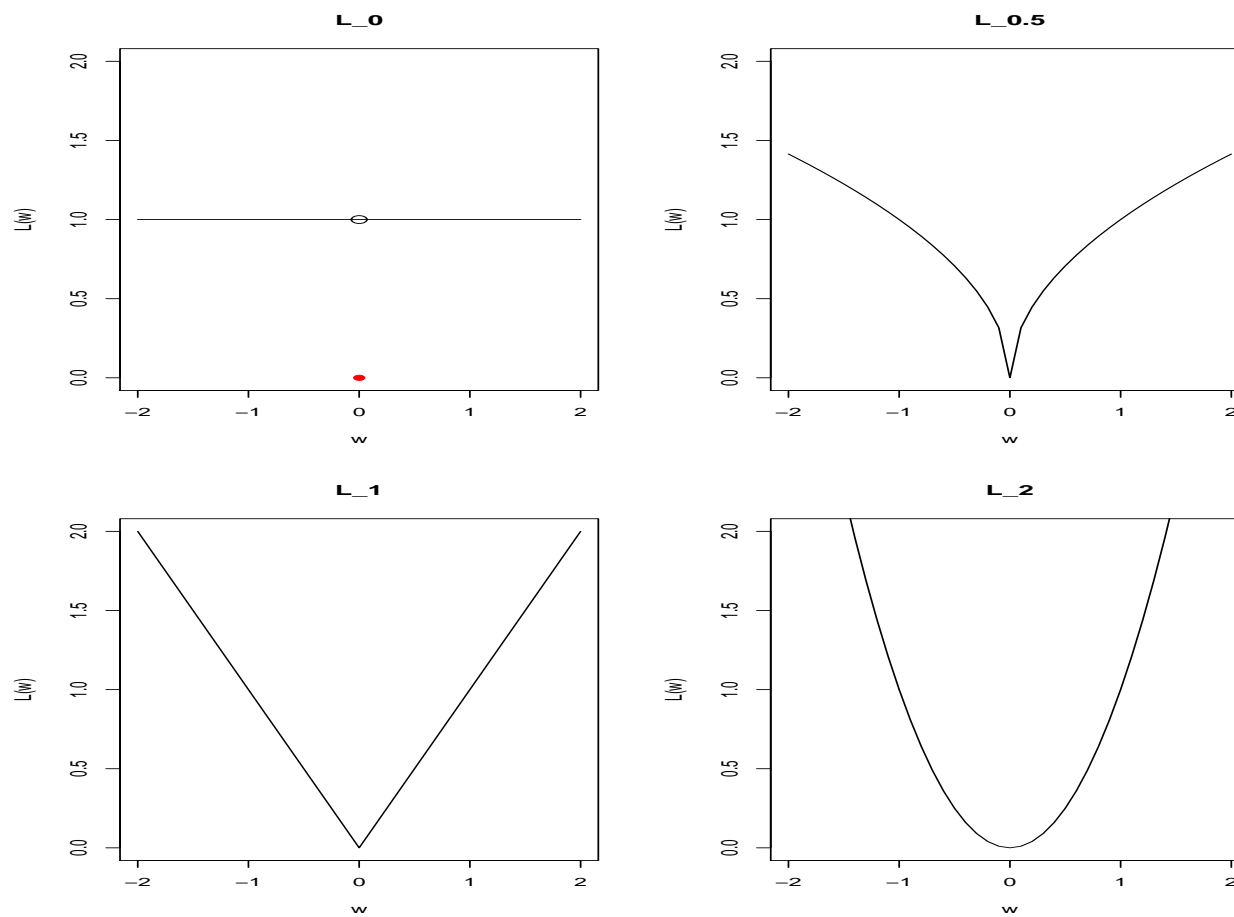


Figure 1: Hard-thresholding penalty  $L_0$ ; soft-thresholding penalties  $L_{0.5}$ ,  $L_1$ ; and  $L_2$  penalty.

## Smoothly Clipped Absolute Deviation (SCAD) Penalty

Fan and Li (2002)

$$p_{\lambda}(w) = \begin{cases} \lambda|w| & \text{if } |w| \leq \lambda, \\ -\frac{(|w|^2 - 2a\lambda|w| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |w| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |w| > a\lambda, \end{cases}$$

where  $a > 2$  and  $\lambda > 0$  are tuning parameters.

- A quadratic spline function with two knots at  $\lambda$  and  $a\lambda$ .
- Except being singular at the origin, the function  $p_{\lambda}(w)$  has a continuous first-order derivative. Not convex.
- imposing a constant penalty on large coefficients.

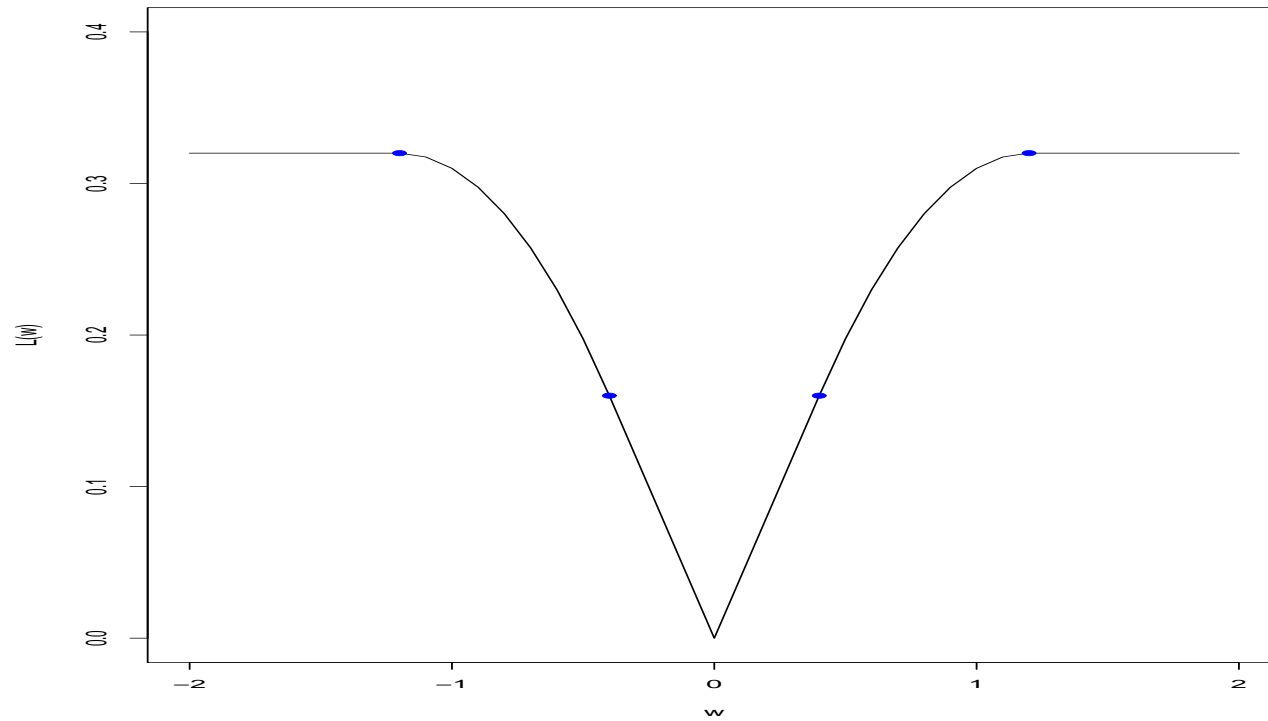


Figure 2: The SCAD penalty function with  $\lambda = 0.4$  and  $a = 3$ .

## SCAD vs $L_1$ Penalty

Consider  $\min \frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|)$

- Recall  $L_1$  penalty solution  $\hat{\theta} = \text{sign}(z)(|z| - \lambda)_+$ 
  - the estimate can be biased for large coefficients
- SCAD penalty overcomes the biased problem of the  $L_1$  penalty.

$$\hat{\theta} = \begin{cases} \text{sign}(|z| - \lambda)_+ & \text{if } |z| \leq 2\lambda; \\ [(a - 1)z - \text{sign}(z)a\lambda]/(a - 2), & \text{if } 2\lambda < |z| \leq a\lambda; \\ z, & \text{if } |z| > a\lambda. \end{cases}$$

- soft-thresholding at the origin; similar to  $L_0$  for large coefficients
- Under mild regularity conditions, the penalized likelihood estimator: root-n consistent and has oracle property (Fan and Li, 2002)
- With  $\lambda_n$  properly chosen,  $L_q$  ( $q < 1$ ) penalty has oracle property

## SVM with SCAD Penalty

$$\min_{b, \mathbf{w}} \frac{1}{n} \sum_{i=1}^n [1 - y_i(b + \mathbf{w} \cdot \mathbf{x}_i)]_+ + \sum_{j=1}^d p_\lambda(|w_j|). \quad (1)$$

- $\lambda$  balances the trade-off between data fitting and model parsimony.
- If  $\lambda$  is too small, the procedure tends to overfit the training data and gives a classifier with little sparsity.
- If  $\lambda$  is too large, the produced classifier can be very sparse but having a poor discriminating power.

Fan and Li (2001) showed that the Bayes risks are not sensitive to the choice of  $a$  and suggested using  $a = 3.7$ .

## Local Quadratic Approximation (LQA)

Generalization of Newton's method for unconstrained optimization: finds a step away from current point by minimizing a quadratic approximation of the problem.

$$[1 - y_i(b + \mathbf{w} \cdot \mathbf{x}_i)]_+ = \frac{1 - y_i(b + \mathbf{w} \cdot \mathbf{x}_i)}{2} + \frac{|y_i - (b + \mathbf{w} \cdot \mathbf{x}_i)|}{2}.$$

Given initial values  $(b_0, \mathbf{w}_0)$ ,

$$|y_i - (b + \mathbf{w} \cdot \mathbf{x}_i)| \approx \frac{1}{2} \frac{[y_i - (b + \mathbf{w} \cdot \mathbf{x}_i)]^2}{|y_i - (b_0 + \mathbf{w}_0 \cdot \mathbf{x}_i)|} + \frac{1}{2} |y_i - (b_0 + \mathbf{w}_0 \cdot \mathbf{x}_i)|.$$

For the SCAD penalty  $p_\lambda(|w_j|)$ , if  $w_{j0} \neq 0$ , we use

$$p_\lambda(|w_j|) \approx p_\lambda(|w_{j0}|) + \frac{p'_\lambda(|w_{j0}|)}{2|w_{j0}|} (w_j^2 - w_{j0}^2).$$

Approximations have same gradients as original functions at  $(b_0, \mathbf{w}_0)$ .

## Algorithms

At each step, we need to minimize some quadratic function

$$\min_{b, \mathbf{w}} \tilde{A}(b, \mathbf{w}) = \frac{1}{2} \begin{pmatrix} b \\ \mathbf{w} \end{pmatrix}^T Q \begin{pmatrix} b \\ \mathbf{w} \end{pmatrix} - P \begin{pmatrix} b \\ \mathbf{w} \end{pmatrix}.$$

Equivalent to solve the linear equation system

$$Q \begin{pmatrix} \hat{b} \\ \hat{\mathbf{w}} \end{pmatrix} = P.$$

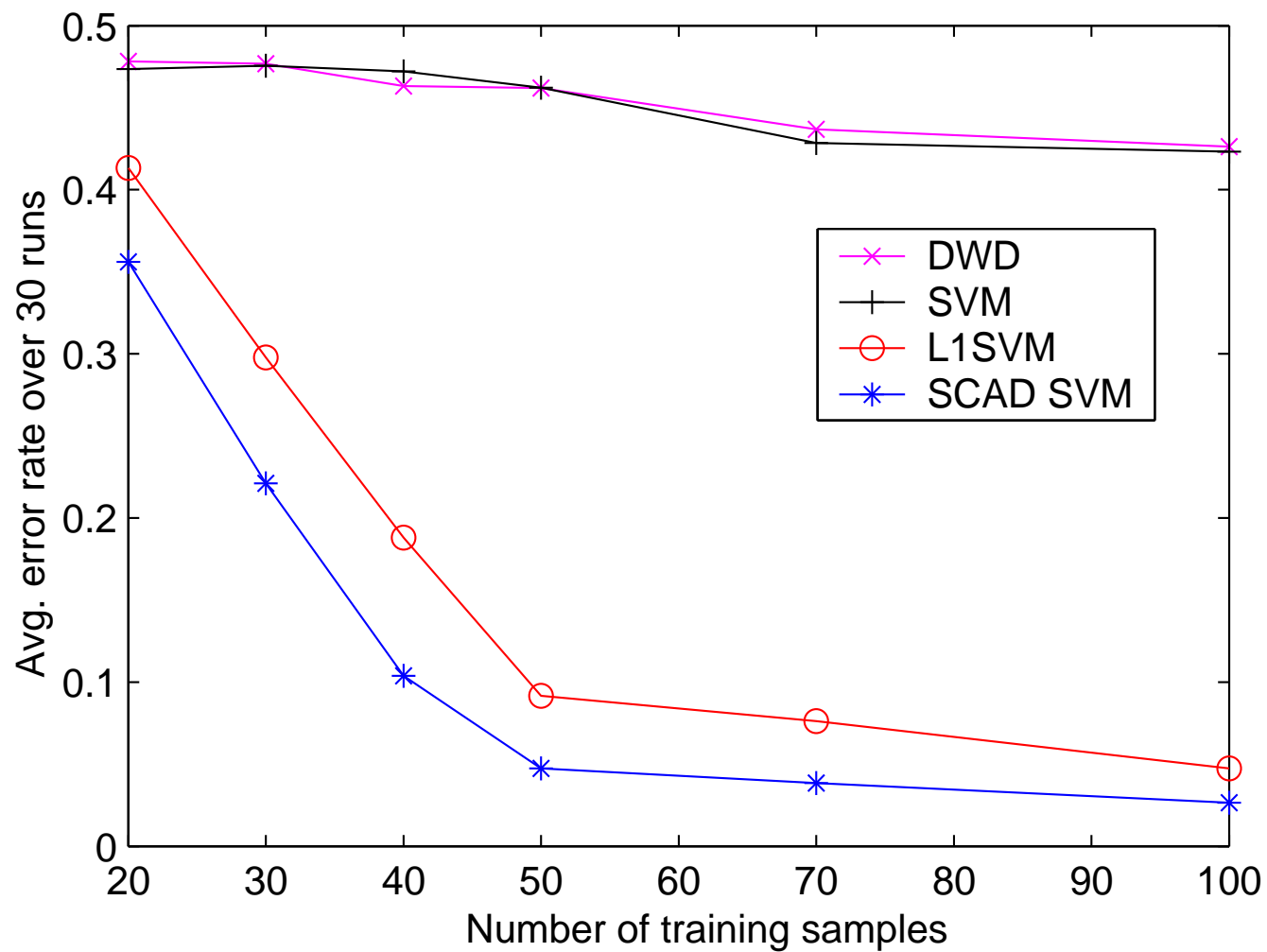
Note: Since the SCAD penalty is unconvex, not guaranteed on global convergence. But it seems to work well in practice.

## Simulation 1

Modified example in Weston (2000).

- $d = 200$  variables, and only the first two are relevant.
- The probability of  $Y = +1$  or  $-1$  is equal.
- First two variables are drawn from a mixture of Normal distributions:
  - with probability 0.7,  $X_1 = Y N(3, 1)$  and  $X_2 = N(0, 1)$ ;
  - with probability 0.3,  $X_1 = N(0, 1)$  and  $X_2 = Y N(3, 1)$ .
- Remaining noise variables are independently generated by  $X_j = N(0, 20)$  for  $j = 3, \dots, 200$ .
- Extra validation set used to tune  $\lambda$ . Choose  $a = 3.7$ .
- Thirty runs for each setting; test error reported.

## Prediction Performance



## Average number of selected variables

	n=20	n=30	n=40	n=50	n=70	n=100
DWD	99.53	99.83	102.13	100.23	98.53	98.97
	(0.83)	(0.73)	(0.76)	(0.67)	(0.70)	(0.61)
SVM	64.07	72.67	81.67	78.93	82.80	87.10
	(1.47)	(1.48)	(1.45)	(0.96)	(1.10)	(0.99)
$L_1$ SVM	15.37	9.10	12.37	11.17	12.47	14.03
	(2.39)	(0.82)	(1.37)	(0.57)	(0.62)	(0.85)
<b>SCAD SVM</b>	8.00	7.90	7.53	6.47	4.73	5.27
	(0.62)	(0.58)	(0.66)	(0.70)	(0.28)	(0.52)

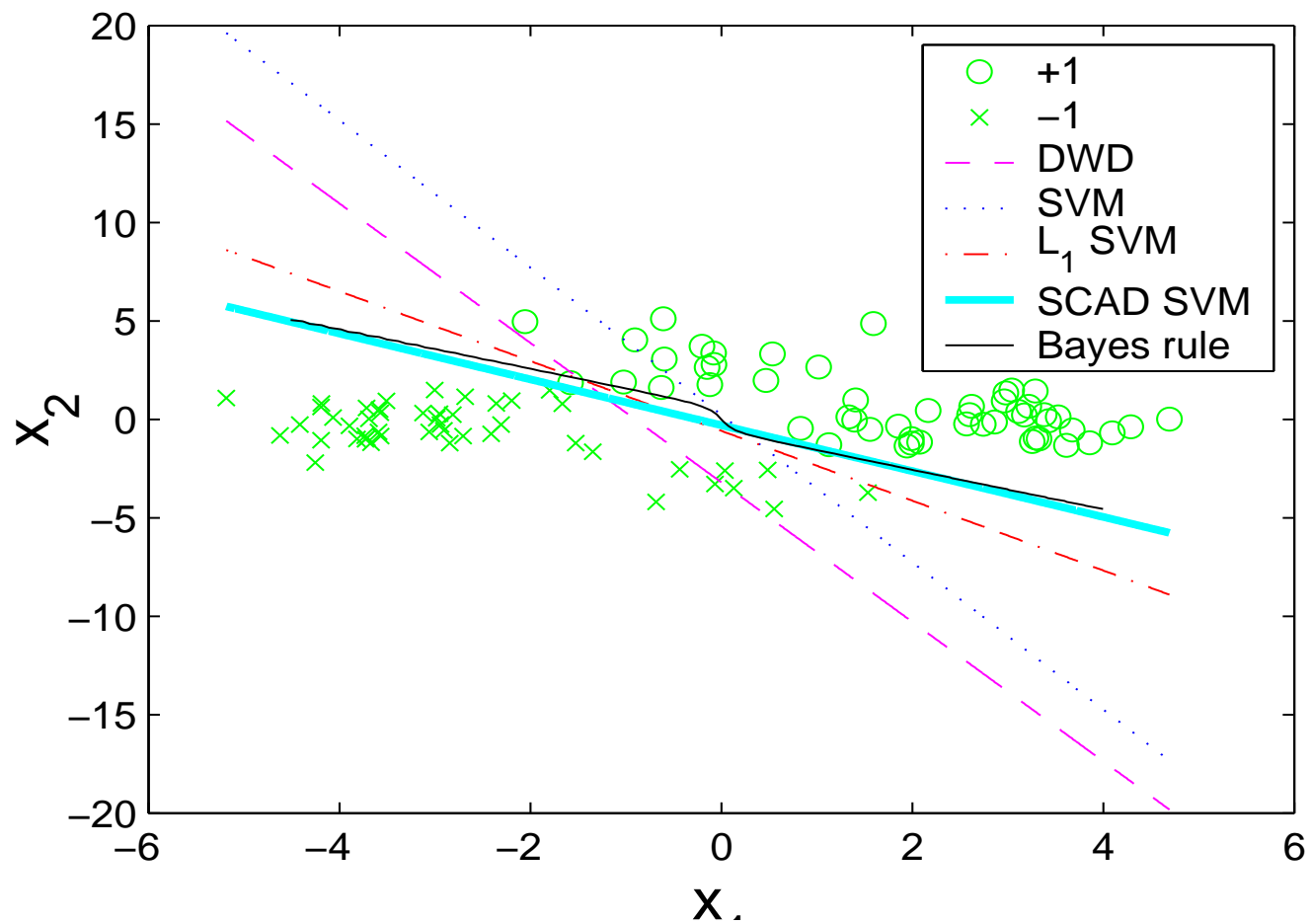


Figure 3: Bayes rule and four classification boundaries (projected onto first two dimensions) in the simulated example.

## UNC Breast Cancer Dataset

- From UNC Microarray Database (Perou 2000, Vantveer 2002, Sotiriou 2003)
- Classify two types of breast cancer: *Luminal vs Not Luminal*
- Three different sources
  - Stanford (5,974 genes and 104 patients)
  - Rosetta (24,187 genes and 97 patients)
  - Singapore (7,650 genes and 99 patients)
- The data was combined and corrected using DWD batch correction (Hu 2005, Benito 2004)
- The combined data set has 2,924 genes and totally 300 patients.

## Gene Ranking Methods

Firstly, rank genes according to some criteria, then use top-ranked genes to classify cancers. Commonly-used ranking criteria:

- correlation coefficients (Golub et al. 1999, Furey et al. 2000, Pavlidis et al. 2001)
- hypothesis testing statistics: two-sample t-test (Devore et al. 1997, Thomas et al. 2001, Pan et al. 2002, He 2004).
- BW ratio: the ratio of between-class to within-class sum of squares (Dudoit et al. 2002).

Easy to use in practice. However

- Has to specify the number of selected  $q$  genes in advance, often subjectively;
- Use individual gene information; correlation among genes ignored.

## Approximate t-statistic for Ranking

For each gene  $x_j$ , calculate the means  $\mu_j^+$ ,  $\mu_j^-$ , the standard deviations  $\sigma_j^+$ ,  $\sigma_j^-$ . Define

$$w_j = (\mu_j^+ - \mu_j^-) / (\sigma_j^+ + \sigma_j^-).$$

- Golub et al. (1999) selected  $q$  genes with the largest positive  $w_j$ 's and  $q$  genes with the largest negative  $w_j$ 's.
- Furey et al. (2000) used  $|w_j|$  to select top  $q$  genes.
- Pavlidis et al. (2001) used a Fisher-discriminant type correlation coefficient.

## Prediction Accuracy

	Stanford	Rosetta	Singapore	Average
t-test (q=50)	.202	.217	.192	.203
t-test (q=100)	.192	.206	.111	.170
SVM	.154	.175	.051	.127
$L_1$ SVM	.125	.216	.081	.141
<b>SCAD SVM</b>	.115	.175	.061	.117

Gene selection result:

	Stanford	Rosetta	Singapore	Average
$L_1$ SVM	59	63	72	65
<b>SCAD SVM</b>	15	19	31	22

## Variable Selection for SVM

UGid	SCAD	L1	Total	t-test	Int.	Name
Hs.169946	3	3	6	3	Y	GATA binding protein 3
Hs.79136	3	2	5	3	Y	solute carrier family 39 (metal ion transporter), member 6
Hs.80420	3	2	5	3	Y	chemokine (C-X3-C motif) ligand 1
Hs.1657	2	3	5	3	Y	estrogen receptor 1
Hs.26770	2	3	5	3	Y	fatty acid binding protein 7, brain
Hs.1041	2	2	4	0	N	v-ros UR2 sarcoma virus oncogene homolog 1 (avian)
Hs.137476	2	2	4	0	N	paternally expressed 10
Hs.252938	2	2	4	0	N	low density lipoprotein-related protein 2
Hs.298654	2	2	4	0	N	dual specificity phosphatase 6
Hs.369508	2	2	4	0	N	phosphoserine phosphatase-like
Hs.412999	2	2	4	1	N	cystatin A (stefin A)
Hs.9795	2	2	4	0	Y	acyl-Coenzyme A oxidase 2, branched chain
Hs.98998	2	2	4	0	Y	tenascin C (hexabrachion)
Hs.2962	2	1	3	0	Y	S100 calcium binding protein P
Hs.442844	2	1	3	0	Y	fibromodulin
Hs.75256	2	1	3	0	N	regulator of G-protein signalling 1
Hs.111676	1	2	3	0	Y	protein kinase H11
Hs.2178	1	2	3	0	Y	histone 2, H2be
Hs.420563	1	2	3	0	N	NADH dehydrogenase (ubiquinone) Fe-S protein 1, 75kDa (NADH-coenzyme Q reductase)
Hs.437638	1	2	3	3	Y	X-box binding protein 1
Hs.458430	1	2	3	2	N	N-acetyltransferase 1 (arylamine N-acetyltransferase)
Hs.89603	1	2	3	2	Y	mucin 1, transmembrane
Hs.91448	1	2	3	0	N	dual specificity phosphatase 14
Hs.191842	0	3	3	2	Y	cadherin 3, type 1, P-cadherin (placental)
Hs.437457	0	3	3	0	Y	lactotransferrin
Hs.75736	0	3	3	0	Y	apolipoprotein D
Hs.79187	0	3	3	0	N	coxsackie virus and adenovirus receptor

Figure 4: Gene selection frequency for breast cancer data.

## Selected Important Genes

- Genes selected at least three times in total by SCAD and  $L_1$  SVM
- “intrinsic” genes, selected by individual-based technique (Perou 2000)
- Last column is the corresponding descriptive name of UniGene identifiers.

### Results:

- Top gene Hs.169946 is selected by all the methods in each learning.
- Hs.79136 and Hs.80420 are selected three times by the SCAD SVM but only twice by the  $L_1$  SVM.
- All the top five genes are also selected by Perou (2000) and the t-test, suggesting that these genes may be highly relevant to breast cancer pathway.

## Metabolism Dataset

Contain the quantitative measurements of all small molecule metabolites in biological samples.

- Most metabolites are not informative in predicting disease or non-disease outcomes (Stitt 2003).
- Need to incorporate variable selection with classification techniques.
- Our metabolism data set is provided by Metabolon Inc.
  - contains metabolic profiles of 63 samples: 32 healthy subjects and 31 subjects diagnosed with a certain disease.
  - Within the patient group, 9 subjects are taking medication and 22 are not.
  - For each sample, its metabolic profile contains the intensity levels of 317 compounds (metabolites).

## LOOCV Error and Number of Selected Metabolites

	test error	metabolite selected
t-test (q=50)	0.370 (0.018)	50
t-test (q=100)	0.235 (0.016)	100
Furey (q=50)	0.375 (0.011)	50
Furey (q=100)	0.230 (0.011)	100
DWD	0.159 (0.012)	315
SVM	0.190 (0.013)	307
$L_1$ SVM	0.174 (0.012)	32
<b>SCAD SVM</b>	0.143 (0.020)	18

## Multiclass Classification

Code  $y$  as  $\{1, \dots, K\}$ , and decision vector  $\mathbf{f} = (f_1, \dots, f_K)$ .

- $f_j$  is a map from input domain  $R^p$  to  $R$ , representing the class  $j$ .
- The classifier assigns a new input  $\mathbf{x}$  to the class with the largest  $f_k(\mathbf{x})$

$$\phi(\mathbf{x}) = \arg \max_{k=1, \dots, K} f_k(\mathbf{x})$$

- Assume  $(\mathbf{x}_i, y_i)$ 's iid from  $p_k(\mathbf{x}) = \Pr(Y = k | \mathbf{X} = \mathbf{x})$ 
  - Generalization error for  $\mathbf{f}$ ,  $\text{GE}(\mathbf{f}) = P(Y \neq \arg \max_k f_k(\mathbf{x}))$ .
  - Bayes rule which minimizes the GE is

$$\phi_B(\mathbf{x}) = \arg \min_{k=1, \dots, K} [1 - p_k(\mathbf{x})] = \arg \max_{k=1, \dots, K} p_k(\mathbf{x}).$$

## Multicategory SVM and $L_1$ MSVM

For linear classifiers,  $f_k(\mathbf{x}) = b_k + \sum_{j=1}^p w_{kj}x_j$ ,  $k = 1, \dots, K$ .

- Standard MSVM (Lee et al. 2004)

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K I(y_i \neq k) [f_k(\mathbf{x}_i) + 1]_+ + \lambda \sum_{k=1}^K \sum_{j=1}^p w_{kj}^2.$$

under the sum-to-zero constraints.

- Wang and Shen (2005) proposed  $L_1$  MSVM by imposing the  $L_1$  penalty on all the coefficients

$$\min_{\mathbf{b}, \mathbf{w}} \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K I(y_i \neq k) [b_k + \mathbf{w}_k^T \mathbf{x}_i + 1]_+ + \lambda \sum_{k=1}^K \sum_{j=1}^p |\mathbf{w}_{kj}|,$$

under the sum-to-zero constraint.

## Simultaneous Selection

Define the weight matrix  $W$  such that its  $(k, j)$  entry is  $w_{kj}$ .

	$x_1$	$\cdots$	$x_j$	$\cdots$	$x_d$
Class 1	$w_{11}$	$\cdots$	$w_{1j}$	$\cdots$	$w_{1d}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Class k	$w_{k1}$	$\cdots$	$w_{kj}$	$\cdots$	$w_{kd}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Class K	$w_{K1}$	$\cdots$	$w_{Kj}$	$\cdots$	$w_{Kd}$

- $L_1$  MSVM treats all  $w_{kj}$ 's equally without distinction.
- We take into account: some coefficients are associated with the same covariate. More natural to treat them as a group rather than separately.

## Variable Selection for MSVM

- Define  $\mathbf{w}_k = (w_{k1}, \dots, w_{kp})^T$  for the  $k$ th row vector of  $W$
- Define  $\mathbf{w}_{(j)} = (w_{1j}, \dots, w_{Kj})^T$  for the  $j$ th column vector of  $W$ .

Each variable  $x_j$  is associated with  $K$  weights  $\mathbf{w}_{(j)}$ .

$$\min_{\mathbf{b}, \mathbf{w}} \quad \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K I(y_i \neq k) [b_k + \mathbf{w}_k^T \mathbf{x}_i + 1]_+ + \lambda \sum_{j=1}^p J(|\mathbf{w}_{(j)}|),$$

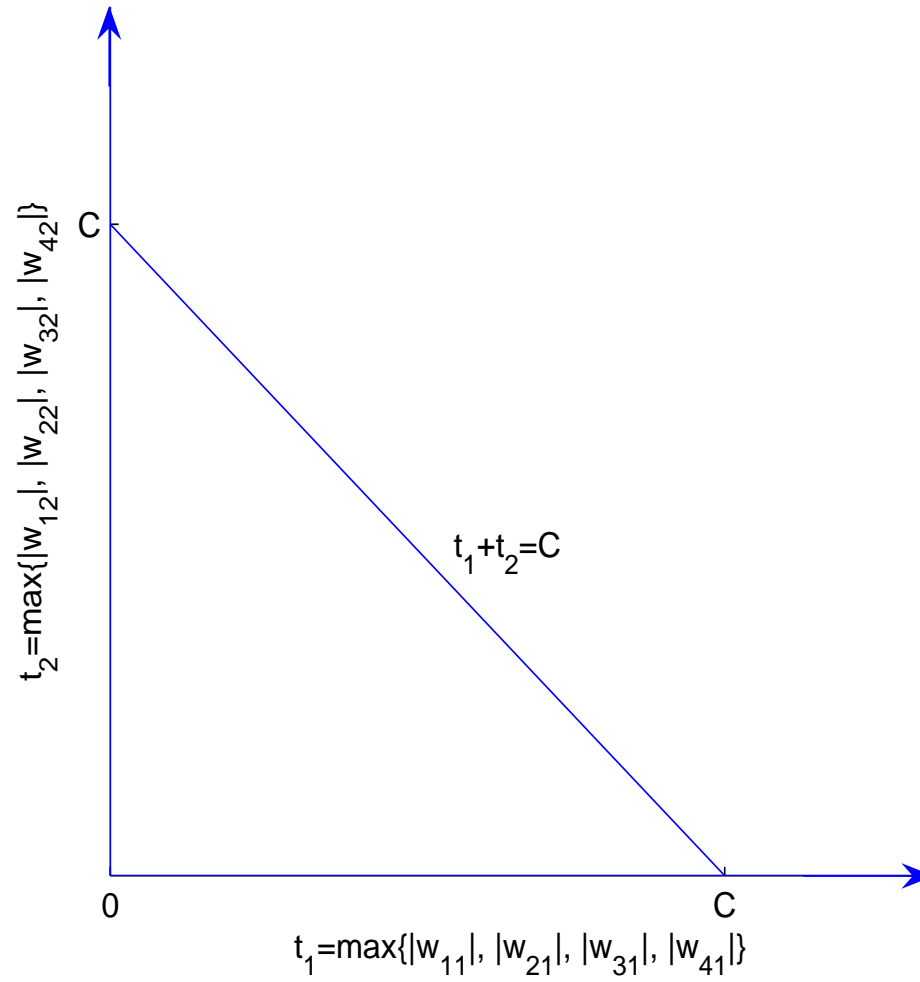
subject to  $\mathbf{1}^T \mathbf{b} = 0, \quad \mathbf{1}^T \mathbf{w}_{(j)} = 0, \quad \text{for } j = 1, \dots, p.$

## Choices of Norm

- $L_q$  norm:  $J_q(|\mathbf{w}_{(j)}|) = \left[ \sum_{k=1}^K w_{kj}^q \right]^{1/q} \quad (q \geq 1)$
- $L_2$  norm:  $J_2(|\mathbf{w}_{(j)}|) = \left[ \sum_{k=1}^K w_{kj}^2 \right]^{1/2}$
- $L_\infty$  norm:  $\|\mathbf{w}_{(j)}\|_\infty = \max_{k=1, \dots, K} |w_{kj}|$
- SCAD related penalty (on-going work)

Related to group variable selection

- Yuan and Lin (2004), Yuan and Zou (2006)
- Different grouping: their group corresponds multiple variables, our group corresponds to one variable



## Supnorm MSVM

Define the matrix  $A$  by  $a_{ik} = I(y_i \neq k)$ ,  $i = 1, \dots, n$ ;  $k = 1, \dots, K$ .

Introduce slack variables  $\xi_{ik}$  such that

$$\xi_{ik} = [b_k + \mathbf{w}_k^T \mathbf{x}_i + 1]_+ \quad \text{for } i = 1, \dots, n; \quad k = 1, \dots, K.$$

Then the problem becomes

$$\min_{\mathbf{b}, \mathbf{w}, \xi} \quad \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K a_{ik} \xi_{ik} + \lambda \sum_{j=1}^p \|\mathbf{w}_{(j)}\|_\infty,$$

subject to  $\mathbf{1}^T \mathbf{b} = 0, \quad \mathbf{1}^T \mathbf{w}_{(j)} = 0, \quad j = 1, \dots, p,$

$$\xi_{ik} \geq b_k + \mathbf{w}_k^T \mathbf{x}_i + 1, \quad \xi_{ik} \geq 0,$$

$$i = 1, \dots, n; \quad k = 1, \dots, K.$$

## Supnorm MSVM (continued)

Further, we introduce a second set of slack variables

$$\eta_j = \|\mathbf{w}_{(j)}\|_\infty = \max_{k=1,\dots,K} |w_{kj}|,$$

which also bring a set of new constraints

$$|w_{kj}| \leq \eta_j, \quad \text{for } k = 1, \dots, K; \quad j = 1, \dots, p.$$

Write  $w_{kj} = w_{kj}^+ - w_{kj}^-$ , and define  $\boldsymbol{\beta} = (\eta_1, \dots, \eta_p)^T$ .

$$\begin{aligned} \min_{\mathbf{b}, \mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\eta}} \quad & \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K a_{ik} \xi_{ik} + \lambda \sum_{j=1}^p \eta_j, \\ \text{subject to} \quad & \mathbf{1}^T \mathbf{b} = 0, \quad \mathbf{1}^T [\mathbf{w}_{(j)}^+ - \mathbf{w}_{(j)}^-] = 0, \quad j = 1, \dots, p, \\ & \xi_{ik} \geq b_k + [\mathbf{w}_k^+ - \mathbf{w}_k^-]^T \mathbf{x}_i + 1, \quad \xi_{ik} \geq 0, \quad i = 1, \dots, n; \quad k = 1, \dots, K, \\ & \mathbf{w}_{(j)}^+ + \mathbf{w}_{(j)}^- \leq \boldsymbol{\eta}, \quad \mathbf{w}_{(j)}^+ \geq \mathbf{0}, \quad \mathbf{w}_{(j)}^- \geq \mathbf{0}, \quad j = 1, \dots, p. \end{aligned}$$

## Four-Class Linear Example

- the input vector  $\mathbf{x}$  in 20-dimensional space.
- First two components of the input vector are generated from the mixture Gaussian in the following way: for class  $k = 1, 2, 3, 4$ , generate  $(x_1, x_2)$  independently from  $N(\boldsymbol{\mu}_k, \sigma_1^2 I_2)$ , with  $\boldsymbol{\mu}_1 = (\sqrt{2}, \sqrt{2})$ ,  $\boldsymbol{\mu}_2 = (-\sqrt{2}, \sqrt{2})$ ,  $\boldsymbol{\mu}_3 = (-\sqrt{2}, -\sqrt{2})$ ,  $\boldsymbol{\mu}_4 = (\sqrt{2}, -\sqrt{2})$ ,
- Remaining eighteen components are i.i.d. generated from  $N(0, \sigma_2^2)$ .
- We generate the same number of observations in each class.
- Here  $n = 200$  and  $n' = 40,000$ . Bayes error is 0.292.

## Classification and Variable Selection Results

Method	Test Error	No. of Zeros	Model Size	Corr. Models
L2 MSVM	0.346 (0.029)	0.00	20.00	0/100
L1 MSVM	0.418 (0.036)	18.31	17.92	4/100
Supnorm	0.296 (0.006)	70.00	2.50	80/100

Supnorm MSVM has the smallest testing error 0.296.

## Variable Selection for SVM

Method	$x_1, x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	Selection Frequency					$z_5$
									$x_0$	$z_1$	$z_2$	$z_3$	$z_4$	
L2	100	100	100	100	100	100	100	100	100	100	100	100	100	100
L1	100	90	86	90	85	90	90	85	86	91	92	89	93	90
Sup	100	4	3	3	5	0	2	1	3	4	5	2	3	2

## Summary

- SCAD penalty for variable selection of binary SVM
- Supnorm penalty for MSVM
- Ongoing work:
  - characterize the solution path of Supnorm MSVM
  - Adaptive penalties
  - explore other penalties

Joint work with Y. Liu, Y. Wu, and J. Zhu.