

Variable Selection for Support Vector Machines

Hao Zhang

North Carolina State University

Outline

1. Support Vector Machines
2. Variable/Feature Selection in SVM
3. Basis Pursuit Approach
4. Numerical Examples
 - Simulations
 - Applications to Real Data
5. Summary

Two-class Classification

- Response variable $Y \in \{\pm 1\}$; attribute vector $\mathbf{X} = (X^{(1)}, \dots, X^{(d)}) \in R^d$
- Training sets: $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$, with $\mathbf{x}_i = (x_i^{(1)}, \dots, x_i^{(d)})$
- Data i.i.d. from the underlying distribution $P(\mathbf{x}, y)$.
- Need to estimate a function relationship $f(\mathbf{x})$ from the example points. The classification rule is $\text{sign}[f(\mathbf{x})] : R^d \rightarrow \{\pm 1\}$
- The misclassification rate of $f(\cdot)$ at a point (\mathbf{x}, y) is

$$I(f(\mathbf{x}), y) = \begin{cases} 1 & \text{if } yf(\mathbf{x}) < 0, \\ 0 & \text{otherwise.} \end{cases}$$

Bayes Rule

Want to minimize risk

$$R[f] = E_{\mathbf{X}, Y}[I(f(\mathbf{X}, Y))] = \int I(f(\mathbf{x}), y) dP(\mathbf{x}, y)$$

- If $P(\mathbf{X}, Y)$ were known, the optimal rule is Bayes rule $\text{sign}\left[p(\mathbf{X}) - \frac{1}{2}\right]$, where $p(\mathbf{x}) = P\{Y = +1 | \mathbf{X} = \mathbf{x}\}$
- Problem: $P(\mathbf{X}, Y)$ unknown!

Linear Support Vector Machines

- Separating hyperplane with the norm vector \mathbf{w} and intercept b

$$\{\mathbf{x} | f(\mathbf{x}) \equiv \mathbf{w} \cdot \mathbf{x} + b = 0\}, \quad \mathbf{w} \in R^d, \quad b \in R$$

- Objective : For some $C > 0$,

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \left(\sum_{i=1}^n \xi_i \right) \\ \text{subject to} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \quad \text{for all } i = 1, \dots, n \\ & \xi_i \geq 0, \quad i = 1, \dots, n \end{aligned}$$

- Equivalent to

$$\min_{\mathbf{w}, b, \xi} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n [1 - y_i (\mathbf{w} \cdot \mathbf{x}_i + b)]_+$$

Dual Programming for SVM

- Primal Lagrangian: introducing Lagrange multipliers $\alpha_i > 0$,

$$L_P(\mathbf{w}, b, \alpha) \equiv \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i \{y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1\}$$

- Dual problem:

$$\min_{\alpha} L_D(\alpha) = \sum_{j=1}^n \alpha_j - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j$$

subject to $\sum_{i=1}^n \alpha_i y_i = 0, 0 \leq \alpha \leq 1$.

- Solution:

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b$$

Nonlinear Support Vector Machines

- Nonlinearly map data into some inner product space, called “feature space” (of higher or infinite dimensions),

$$\Phi : R^d \rightarrow \mathcal{F}$$

- Apply the linear SVM in the feature space, and the decision rule

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b$$

- The solution involves Φ only through the inner product, we only require the kernel trick: $K(\mathbf{x}_i, \mathbf{x}) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})$.

Kernel

- Mercer Theorem: (Courant & Hilber, 1959)

A kernel $K(\mathbf{x}, \mathbf{z})$ with $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$ is a dot product in some feature space, or $K(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{z})$, if and only if

$$K(\mathbf{x}, \mathbf{z}) = K(\mathbf{z} \cdot \mathbf{x})$$
$$\int \int K(\mathbf{x}, \mathbf{z}) f(\mathbf{x}) f(\mathbf{z}) d\mathbf{x} d\mathbf{z} \geq 0 \quad \forall f \in L^2$$

- Examples:

- d -th degree polynomial: $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x} \cdot \mathbf{z})^d$
- Radial basis: $K(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2)/c$

Regularization Framework Formulation

- The SVM with kernel K is equivalent to a regularization problem in a reproducing kernel Hilbert space (RKHS)
- Find $f(\mathbf{x})$ of the form $h(\mathbf{x}) + b$, where $h \in \mathcal{H}_K$, the RKHS associated with kernel $K(s, t)$, to minimize

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

where $[\tau]_+ = \tau$ if $\tau > 0$; $= 0$ otherwise.

- λ is the tuning parameter
- The loss function $[1 - yf(\mathbf{x})]_+$ is called “hinge-loss”.

Representer Theorem

- Representer Theorem (Kimeldorf and Wahba 1971)

Any solution to the problem above has a representation of the form

$$f(\mathbf{x}) = b + \sum_{i=1}^n c_i K(\mathbf{x}_i, \mathbf{x}).$$

- Let K be the $n \times n$ matrix with (i, j) entry $K(\mathbf{x}_i, \mathbf{x}_j)$, we have

$$\|f\|_{\mathcal{H}_K}^2 = c' K c$$

- Equivalently, the SVM minimizes

$$\frac{1}{n} \sum_{i=1}^n [1 - y_i f(\mathbf{x}_i)]_+ + \lambda c' K c$$

- Dual Problem is typically solved in the SVM literature.

Variable/Feature Selection in Learning

- Purposes:
 - to eliminate irrelevant variables
 - to improve the accuracy of the machine learning process
 - to reduce the collection/storage/computation load
- Filter method - a preprocessing step to induction that can remove irrelevant attributes before induction occurs
 - Pearson correlation coefficients
 - Fisher criterion score
- Wrapper method - a search through the subset during an induction algorithm

Variable Selection in SVM

- Linear SVM
 - Bradley and Mangasarian (1998) – concave minimization
 - Guyon et al. (2000), Rakotomamonjy (2003) – recursive feature elimination (RFE)
 - Jebara and Jaakkola (2000) – in probabilistic setting, feature selection in maximum entropy discrimination (MED)
- Nonlinear SVM
 - Weston et al. (2002), Grandvalet (2003) – adaptive scaling

Shrinkage Methods in Regression

- Frank and Friedman (1993) – Bridge regression

$$\min \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^d \beta_j x_i^{(j)})^2 + \lambda \sum_{j=1}^d |\beta_j|^q, \quad q \geq 1$$

– Tibshirani (1996) – LASSO $q = 1$

- Breiman (1995) – Nonnegative Garrote

$$\sum_{i=1}^n (y_i - \sum_{j=1}^d c_j \hat{\beta}_j x_i^{(j)})^2 \quad \text{subject to} \quad c_j \geq 0, \sum_{j=1}^d |c_j| \leq s.$$

- Smoothly clipped absolute deviation (SCAD) by Fan and Li (2001)

$$\min \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^d \beta_j x_i^{(j)})^2 + \lambda \sum_{j=1}^d p_j(|\beta_j|)$$

Penalty in Variable Selection for Linear SVM

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \sum_{i=1}^n \xi_i + \lambda J_q(\mathbf{w}) \\ \text{subject to} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \quad \text{for all } i = 1, \dots, n \\ & \xi_i \geq 0, \quad i = 1, \dots, n \end{aligned}$$

- $q > 0$,

$$J_1(\mathbf{w}) = \sum_{j=1}^d |w_j|, \quad J_2(\mathbf{w}) = \sum_{j=1}^d w_j^2, \quad J_\infty(\mathbf{w}) = \max_j |w_j|.$$

- Feature selection occurs only when $q = 1$, by Bradley and Mangasarian (1998).

Generalization to nonlinear SVM

$$\frac{1}{n} \sum_{i=1}^n [1 - y_i f(\mathbf{x}_i)]_+ + \lambda J(f)$$

- Want shrinkage on estimate \hat{f} .
- How to choose $J(f)$?

Functional ANOVA Decomposition

Consider a function $f(\underline{x})$ on a product domain $\mathcal{X} = \prod_{j=1}^d \mathcal{X}_j$.

$$f(x^{(1)}, \dots, x^{(d)}) = b_0 + \sum_{j=1}^d f_j(x^{(j)}) + \sum_{j < k} f_{jk}(x^{(j)}, x^{(k)})$$

+ all higher-order interactions

- Let A_j be an averaging operator on \mathcal{X}_j satisfying $A_j^2 = A_j$. An ANOVA decomposition of f is $f = \left\{ \prod_{j=1}^d (I - A_j + A_j) \right\} f = \sum_{\mathcal{S}} \left\{ \prod_{j \in \mathcal{S}} (I - A_j) \prod_{j \notin \mathcal{S}} A_j \right\} f = \sum_{\mathcal{S}} f_{\mathcal{S}}$.
- $b_0 = \prod_{j=1}^d A_j f$, $f_j = (I - A_j) \prod_{k \neq j} A_k f$, and $f_{jk} = (I - A_j)(I - A_k) \prod_{l \neq j, k} A_l f$, and so forth.
- Side conditions on f_j 's, f_{jk} 's satisfied to guarantee uniqueness.

Reproducing Kernel Hilbert Space

Consider a Hilbert space \mathcal{H} of functions on domain \mathcal{X} .

- If the evaluation functional $L_x f = f(x)$ is a continuous in \mathcal{H} , $\forall x \in \mathcal{X}$, then \mathcal{H} is called an RKHS.
- For any RKHS, there exists a reproducing kernel $K(\cdot, \cdot)$ such that $(K(x, \cdot), f) = f(x), \forall f \in \mathcal{H}, \forall x \in \mathcal{X}$.
- Aronszajn (1950) showed one-to-one correspondence between RKHS and a non-negative definite kernel K .
- \mathcal{H}_K : the RKHS generated by the kernel K .

Sobolev Hilbert Space on $[0, 1]$

$$W_2[0, 1] = \{g : g, g' \text{ absolutely continuous, } g'' \in \mathcal{L}_2[0, 1]\}$$

- It is an RKHS when endowed with the norm

$$\|g\|^2 = \left\{ \int_0^1 g(t) dt \right\}^2 + \left\{ \int_0^1 g'(t) dt \right\}^2 + \int_0^1 \{g''(t)\}^2 dt,$$

- The reproducing kernel

$$K(s, t) = 1 + k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s - t|), \text{ with}$$

$$k_1(t) = t - \frac{1}{2}, k_2(t) = \frac{1}{2} \left\{ k_1^2(t) - \frac{1}{12} \right\} \text{ and}$$

$$k_4(t) = \frac{1}{24} \left\{ k_1^4(t) - \frac{1}{2} k_1^2(t) + \frac{7}{240} \right\}. \text{ Wahba (1990) and Gu(2002).}$$

- Orthogonal decomposition: $W_2[0, 1] = [1] \oplus \mathcal{H}_0 \oplus \mathcal{H}_1$.
 - $\mathcal{H}_0 = \text{span}\{k_1(t)\}$, the space of linear functions
 - $\mathcal{H}_1 = \mathcal{H}_{K_1}$ with $K_1(s, t) = K(s, t) - 1 - k_1(s)k_1(t)$, the space of smoothing functions.

Tensor Product RKHS

- Let H^j be an RKHS on domain $\mathcal{X}^{(j)}$ with $\mathcal{H}^j = \{1\} \oplus \bar{\mathcal{H}}^j$.
- Construct the tensor product RKHS on domain $\mathcal{X} = \prod_{j=1}^d \mathcal{X}^{(j)}$ by

$$\bigotimes_{j=1}^d \mathcal{H}^j = [1] \oplus \sum_{j=1}^d \bar{\mathcal{H}}^j \oplus \sum_{j < k} [\bar{\mathcal{H}}^j \otimes \bar{\mathcal{H}}^k] \oplus \dots \quad (1)$$

- Each functional component in the functional decomposition lies in a subspace in (1).
- The reproducing kernel of a tensor product space is the product of the reproducing kernels of the individual spaces.
- Typically only low order interactions are considered for interpretability and visualization.

Basis Approach

- Need to solve $f_j \in \bar{\mathcal{H}}^{(j)}$, $f_{jk} \in \bar{\mathcal{H}}^j \otimes \bar{\mathcal{H}}^k$, etc.
- In sobolev space setting, define $\bar{H}^{(j)} = \mathcal{H}_0 + \mathcal{H}_1$. Then we approximate f_j in terms of the basis functions

$$f_j(x^{(j)}) = b_j k_1(x^{(j)}) + \sum_{i=1}^n c_{ij} K_1(x^{(j)}, x_i^{(j)})$$

- “good”: finite dimensional solution
- “bad”: expensive computation when n is large

Parsimonious Basis Functions

- In the usual penalized likelihood setting, the estimate with $N < n$ basis is a good approximation to the estimate with the full set of n representers. See Xiang and Wahba (1997), Gao et al. (2001), and Lin et al. (2000).
- Choose a subset of size N and relabel them by $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$.
- For each j , we consider the optimization problem in

$$\mathcal{H}_*^{(j)} = \text{span}\{k_1(\cdot), K_{x_1^{(j)}}(\cdot), \dots, K_{x_N^{(j)}}(\cdot)\}$$

- Sampling schemes for subset
 - Simple random scheme
 - Clustering scheme

Additive Model – Choosing Main Effects

- The additive model, is a sum of d functions of one variable, by retaining only main effect components in the ANOVA decomposition.

- The function space is

$$\mathcal{H}_* = [1] \oplus \left(\bigoplus_{j=1}^d \text{span}\{k_1(x^{(j)}), K_1(x^{(j)}, x_{l_*}^{(j)})\}, l = 1, \dots, N \right)$$

- The basis pursuit SVM (BPSVM) solves:

$$\min \frac{1}{n} \sum_{i=1}^n [1 - y_i f(\mathbf{x}_i)]_+ + \lambda \left(\sum_{j=1}^d |b_j| + \sum_{j=1}^d \sum_{l=1}^N |c_{lj}| \right), \quad (2)$$

subject to

$$f(\mathbf{x}) = b_0 + \sum_{j=1}^d b_j k_1(x^{(j)}) + \sum_{j=1}^d \sum_{l=1}^N c_{lj} K_1(x^{(j)}, x_{l_*}^{(j)}),$$

- Convex problem with linear constraints.

Two-way Interaction Model

- The function space is

$$\mathcal{H}_* \equiv [1] \oplus \left(\bigoplus_{j=1}^d \mathcal{H}_*^{(j)} \right) \oplus \left(\bigoplus_{j < k} \mathcal{H}_*^{(j)} \otimes \mathcal{H}_*^{(k)} \right).$$

- The optimization problem is:

$$\begin{aligned} \min_{f \in \mathcal{H}_*} & \frac{1}{n} \sum_{i=1}^n [1 - y_i f(\mathbf{x}_i)]_+ + \lambda \left(\sum_{j=1}^d |b_j| + \sum_{j < k} |b_{jk}| \right. \\ & \left. + \sum_{j=1}^d \sum_{l=1}^N |c_{lj}| + \sum_{j < k} \sum_{l=1}^N |c_{jk,l}^{ss}| + \sum_{j \neq k} \sum_{l=1}^N |c_{jk,l}^{\pi s}| \right). \end{aligned} \quad (3)$$

Importance Measure

- Counts of nonzeros c 's: Intuitive; Instable
- L_1 norm: for $f \in [0, 1]$, define $L_1(f) = \int_0^1 |f(t)| dt$. Empirically, generate m points from $[0,1]$ and calculate

$$\begin{aligned}
 L_1(f_j) &= \frac{1}{m} \sum_{i=1}^m |f_j(x_i^{(j)})| \\
 &= \frac{1}{m} \sum_{i=1}^m \left| b_j(x_i^{(j)} - \frac{1}{2}) + \sum_{l=1}^N c_{lj} K(x_i^{(j)}, x_l^{(j)}) \right| \\
 L_1(f_{jk}) &= \frac{1}{m} \sum_{i=1}^m |f_{jk}(x_i^{(j)}, x^{(k)})|.
 \end{aligned}$$

- RKHS norm: for any $f \in \mathcal{H}_K$ with $f(x) = \sum_{i=1}^n c_i K(x_i, x)$, define $\|f\|_{\mathcal{H}_K} = (c' K c)^{\frac{1}{2}}$.

Algorithm – Linear Programming

Define the coefficients by the vector $\mathbf{b} = (b_1, \dots, b_d)$, $\mathbf{c} = (c_{1,1}, \dots, c_{d,N})$, and $\mathbf{f} = (f_1, \dots, f_n) \equiv (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))$, $\mathbf{e} = (1, \dots, 1)$ of length n , and $Y = \text{diag}[y_1, \dots, y_n]$. By introducing some artificial variables $\mathbf{z} = (z_1, \dots, z_n)$, $\xi = (\xi_1, \dots, \xi_d)$, and $\zeta = (\zeta_{1,1}, \dots, \zeta_{d,N})$, the optimization problem in (2) becomes

$$\min_{\mathbf{z}, \mathbf{b}, \mathbf{c}, \xi, \zeta} \sum_{i=1}^n z_i + \lambda_\pi \sum_{j=1}^d \xi_j + \lambda_s \sum_{j=1}^d \sum_{l=1}^N \zeta_{lj}, \quad (4)$$

subject to

$$\mathbf{z} \geq \mathbf{0}, \mathbf{z} + Y\mathbf{f} \geq \mathbf{e}, \xi \geq \mathbf{b}, \xi \geq -\mathbf{b}, \zeta \geq \mathbf{c}, \zeta \geq -\mathbf{c}. \quad (5)$$

This is a linear optimization problem with polyhedral constraints, with the unknowns $(\mathbf{z}, \mathbf{b}, \mathbf{c}, \xi, \zeta)$.

Model Tuning - select tuning parameters

- λ controls the tradeoff between the likelihood \mathcal{L}_t to the training data and the sparsity of coefficients for f
- Let f_λ be the minimizer of the BPSVM, its GCKL is:

$$\begin{aligned} GCKL(\lambda) &= E\left[\frac{1}{n} \sum_{i=1}^n (1 - Y_i f_{\lambda i})_+\right] \\ &= \frac{1}{n} \sum_{i=1}^n [p_i (1 - f_{\lambda i})_+ + (1 - p_i) (1 + f_{\lambda i})_+], \end{aligned}$$

as given in Wahba (1999) and Wahba et al. (2000).

- f_λ is treated fixed and the expectation is taken over the conditional probability $p(Y = +1 | \mathbf{X} = \mathbf{x})$.
- An upper bound on misclassification rate; only computable in simulations.

Model Tuning Criteria

- Generate the extra tuning set, and evaluate the error rate on the tuning set.

$$ETune(\hat{f}_\lambda) = \frac{1}{n_{tune}} \sum_{i=1}^{n_{tune}} I(\text{sign}[\hat{f}_\lambda(\mathbf{x}_{tune,i})] \neq y_{tune,i}).$$

- k-fold cross validation (k-CV).
 - Leave-one-out (LOO) is an extreme case of k -fold cross validation, in which k is equal to the sample size n .
 - In real applications, choose k to be 5 or 10.

Simulation 1 - Linear Model

- There are 6 variables: X_1, \dots, X_6 .

- The true logit function is

$$f(x) = \frac{x_1 + x_2 + x_3}{3}$$

- Important variables: X_1, X_2, X_3
- $n = 200$. $N = 40$.
- Tune λ within $\log_2(\lambda) \in [-12, -1]$.

Parameter Tuning Result

- The parameter λ were tuned by using three criteria: GCKL, ETune, and CV.
- A separate tuning set of $n_{tune} = 200$ was generated with the underlying probability distribution.
- The number of folds $k = 5$ in the cross validation tuning.
- The optimal parameters are $\hat{\lambda}_{GCKL} = 2^{-7}$, $\hat{\lambda}_{ETune} = 2^{-5.5}$, while 5-CV gives multiple minima $\hat{\lambda}_{5-CV} = \{2^{-8}, 2^{-8.5}, 2^{-9}, 2^{-9.5}\}$.

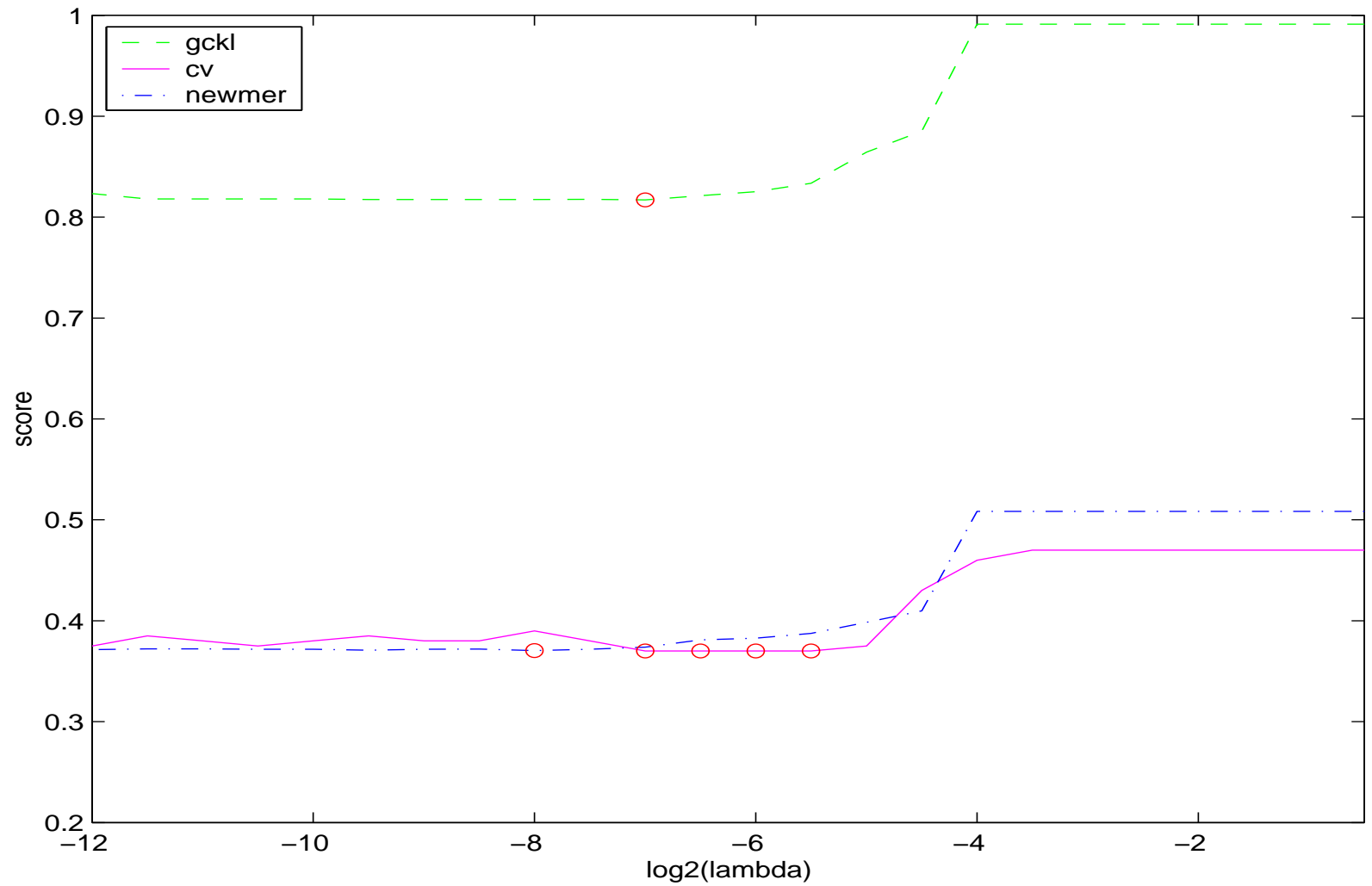


Figure 1: Tuning parameter λ with three different criteria

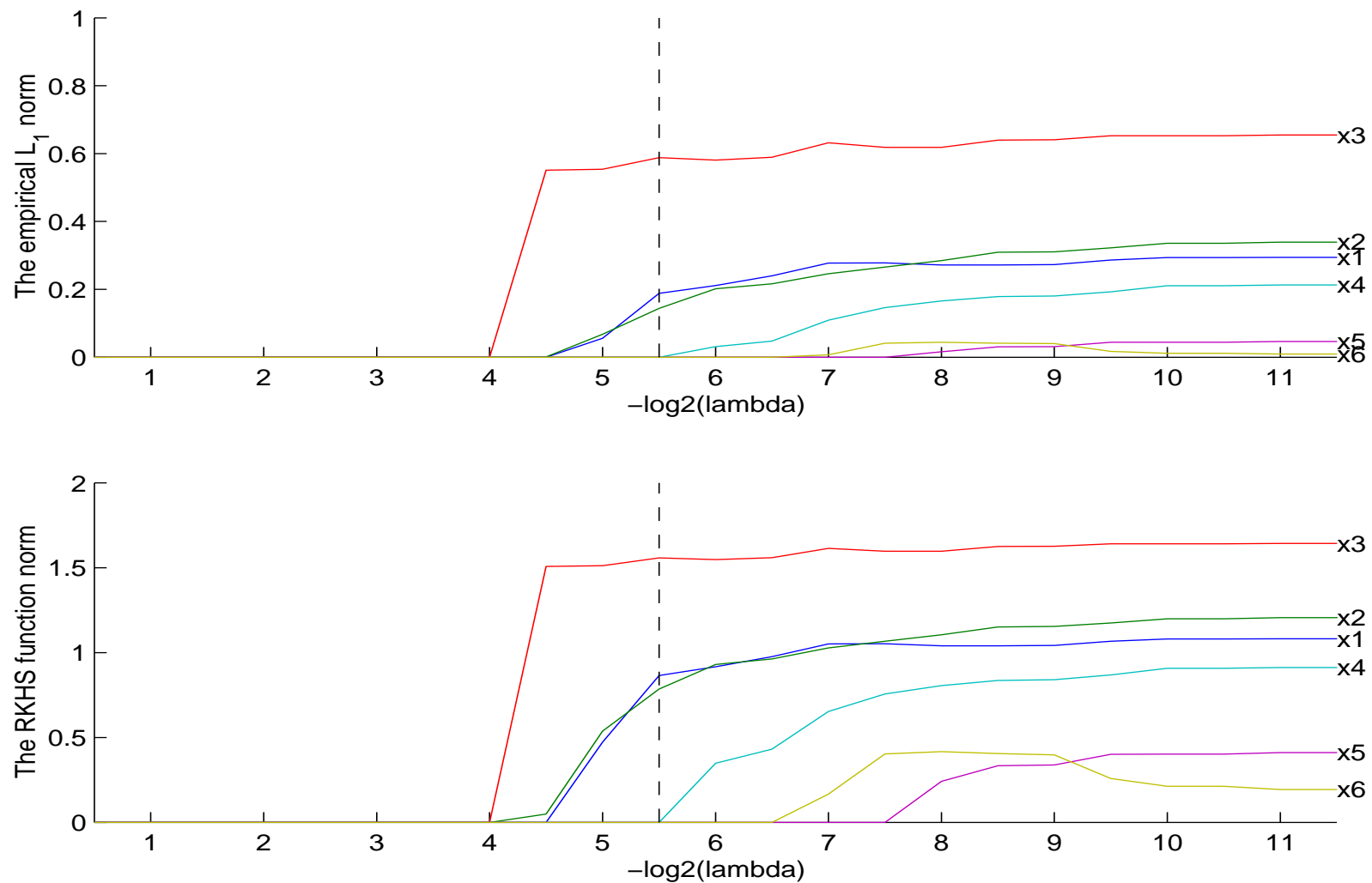


Figure 2: The empirical L_1 norm for the estimated components against λ

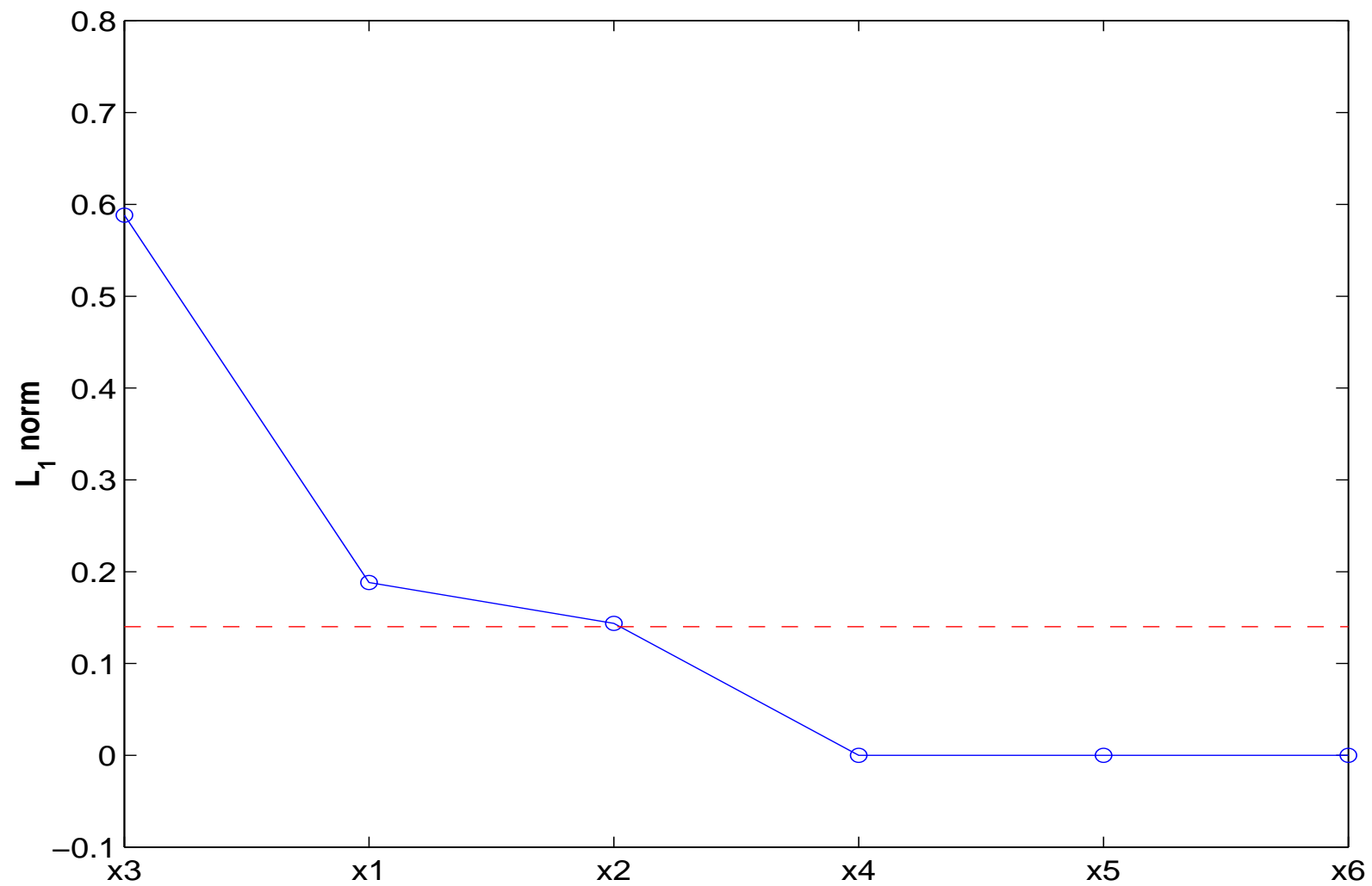


Figure 3: L_1 norm for individual variables (using $\hat{\lambda}_{CV}$)

Sequential Permutation Test

- Order the L_1 norms by $L_{(1)} > L_{(2)} > \dots > L_{(d)}$
- For each $j \geq 1$, conditional on the first $j - 1$ components are relevant (in the model), test: $H_0 : L_{(j)} = 0$ vs $H_1 : L_{(j)} > 0$
- Stop the first insignificant result appears.

Table 3: P-values given by the permutation test

	x_3	x_1	x_2	x_4
p-value	0.01	0.09	0.04	0.4

Estimated Risk

- Since the true probability function $p_0(\mathbf{x})$ is known in this simulation study, we know the Bayes classifier is $\text{sign}[p_0(\mathbf{x}) - \frac{1}{2}]$.
- Risk of f is

$$R(f) = E_{\mathbf{x},y}[I(f(\mathbf{x}, y))] = \int I(f(\mathbf{x}), y)dP(\mathbf{x}, y).$$

- Empirically, $R[f]$ is estimated based on a test set of size $n_{test} = 10000$.

Table 2: The test set error given by various classification rules

	Bayes Rule	\hat{f}_{5-CV}	<i>SVM</i>
test error	0.3626	0.3820	0.4162.

Table 1: The average empirical L_1 norms for components over 35 runs.

(In the parentheses are the standard errors of the average norms)

	x_1	x_2	x_3	x_4	x_5	x_6
L_1 norm	0.3438	0.3673	0.3600	0.0791	0.0639	0.0885
	(0.0293)	(0.0312)	(0.0344)	(0.0133)	(0.0143)	(0.0172)

Simulation 2 - Nonlinear Model

- There are $d = 8$ covariates, taken uniformly from $[0, 1]$ independently.
- The sample size $n = 800$ and the basis size $N = 50$.
- The true probability function is

$$\log \left(\frac{p_0(\mathbf{x})}{1 - p_0(\mathbf{x})} \right) = \frac{4}{3}x_1 + \pi \sin(\pi x_3) + 8x_6^5 + \frac{2}{e - 1}e^{x_8} - 5$$

- Tune the parameter λ in the range of $\log_2(\lambda) \in [-20, -1]$.

Model Tuning

- The parameter λ tuned by using three criteria: GCKL, ETune, and CV.
- $n_{tune} = 500$; $k = 5$ for cross validation.
- The optimal parameters are $\hat{\lambda}_{GCKL} = 2^{-15}$, $\hat{\lambda}_{ETune} = 2^{-16}$, and $\hat{\lambda}_{5-CV} = 2^{-14}$.

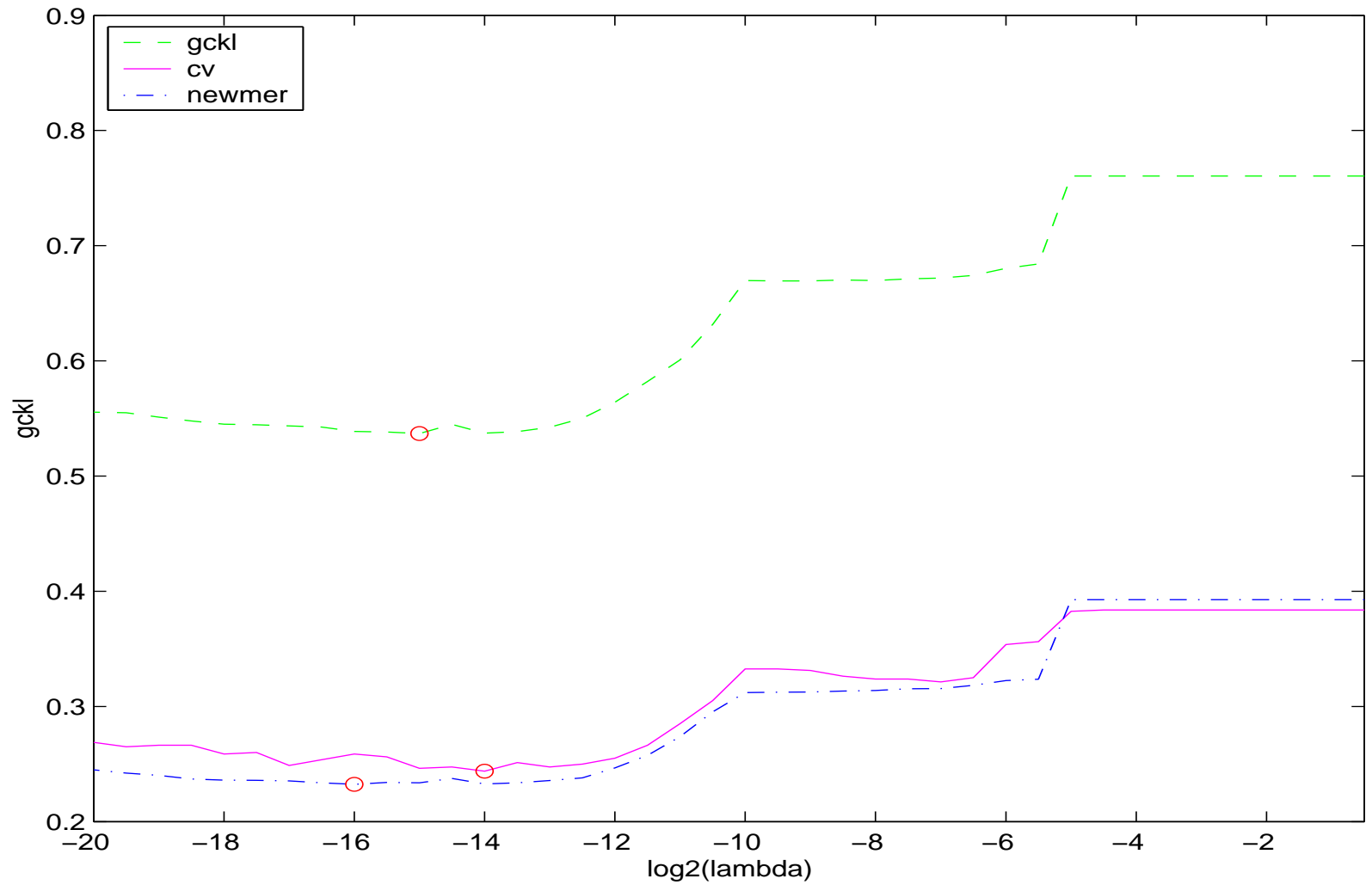


Figure 4: Tuning parameter λ with three different criteria

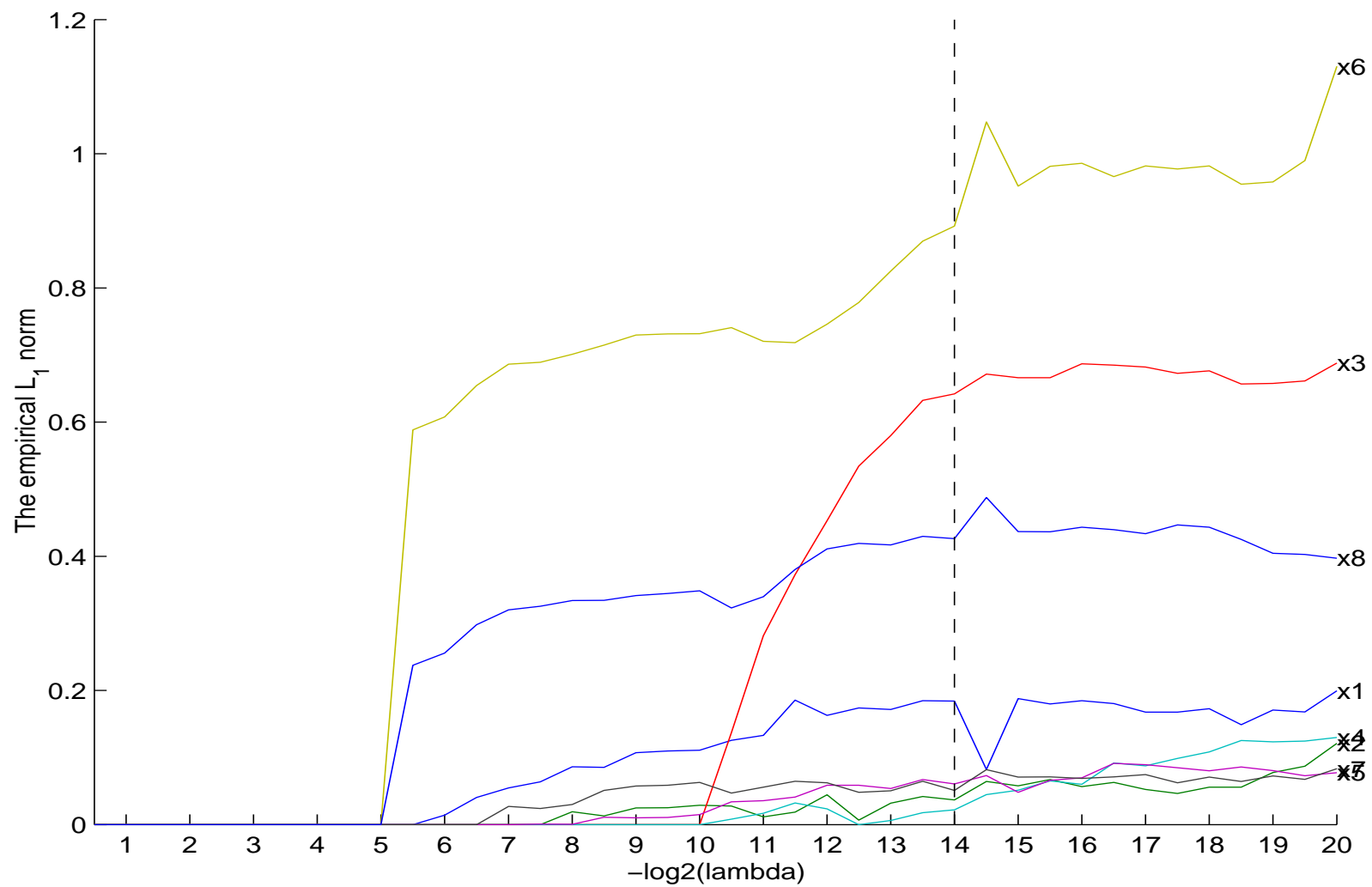


Figure 5: The empirical L_1 norm for the estimated components against λ

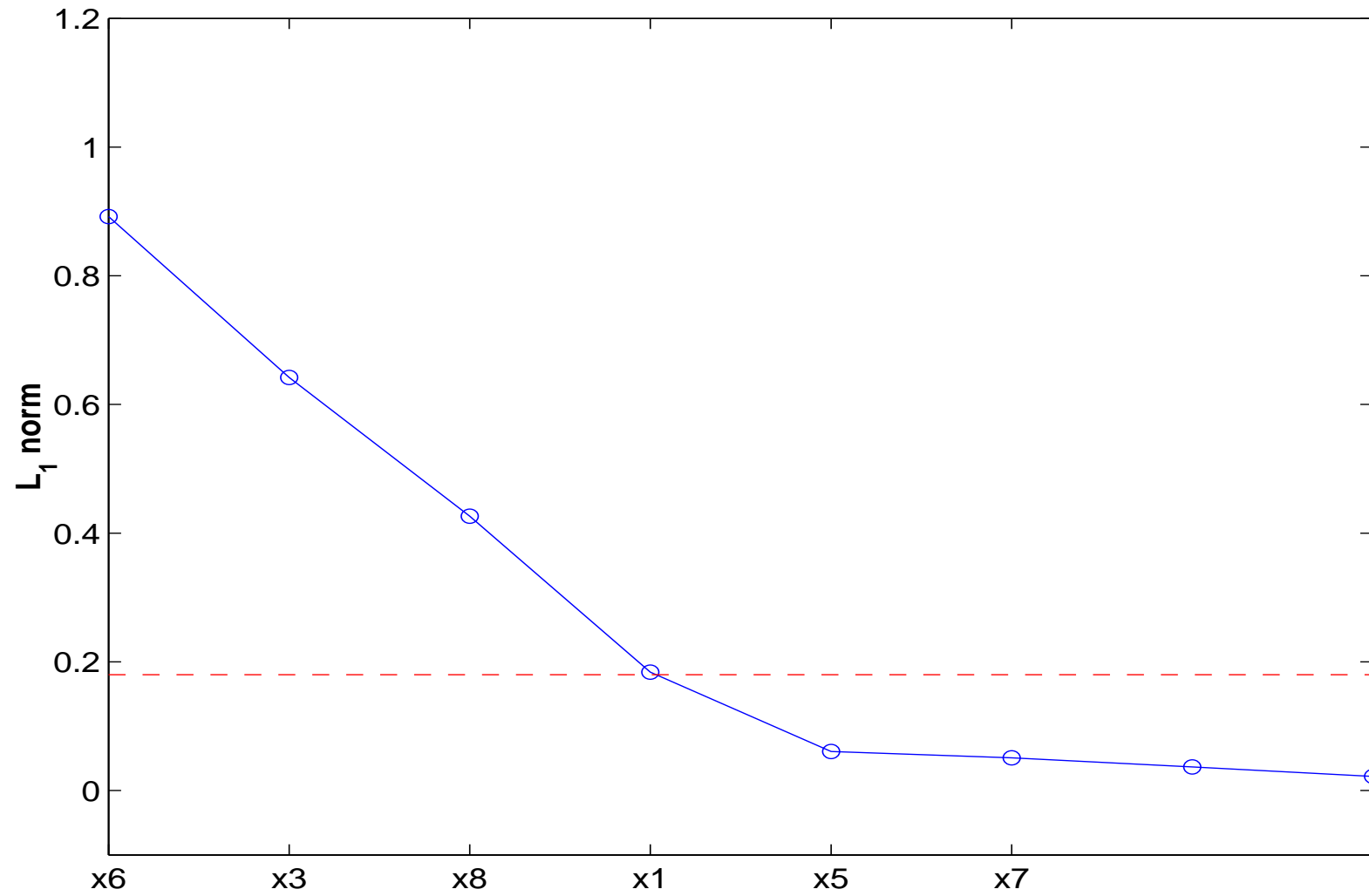


Figure 6: L_1 norm for individual variables (using $\hat{\lambda}_{CV}$)

Table 4: The test error given by various classification rules

	Bayes Rule	\hat{f}_{5-CV}	<i>SVM</i>
test error	0.2256	0.2400	0.3902.

Permutation Test

Table 5: p-values given by the permutation test

	x_6	x_3	x_8	x_1	x_5
p-value	0.02	0.02	0.02	0.02	> 0.52

Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR)

- Ongoing epidemiological study of a cohort of patients receiving medical care in an 11-county area in Wisconsin
- Baseline examination in 1980-1982; the first follow-up in 1984-1986; the second follow-up in 1990-1992
- Retinopathy severity level was graded for each eye.
- Group interested: younger (age < 30) onset group with type 1 diabetes (sample size 668)

Response Variable Y: Four-year progression of diabetic retinopathy

- Y was defined to be 1 if at first follow-up exam, the retinopathy level of a patient progressed two steps from the baseline.

Potential Risk Factors

- DUR: duration of diabetes at the time of baseline examination (years)
- GLY: glycosylated hemoglobin, a measure of hyperglycemia (%)
- BMI: body mass index (kg/m^2)
- SYS: systolic blood pressure (mmHg)
- RET: retinopathy level
- PULSE: pulse rate
- INS: insulin dose (kg/day)
- SCH: years of school completed (years)
- IOP: intraocular pressure (mmHg)

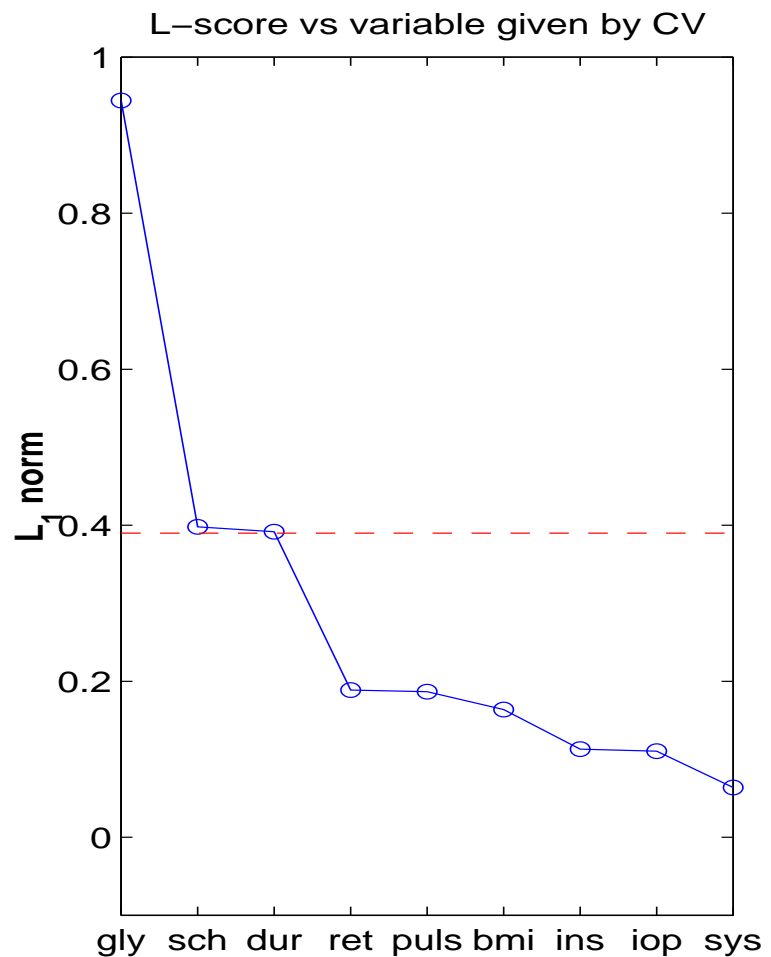
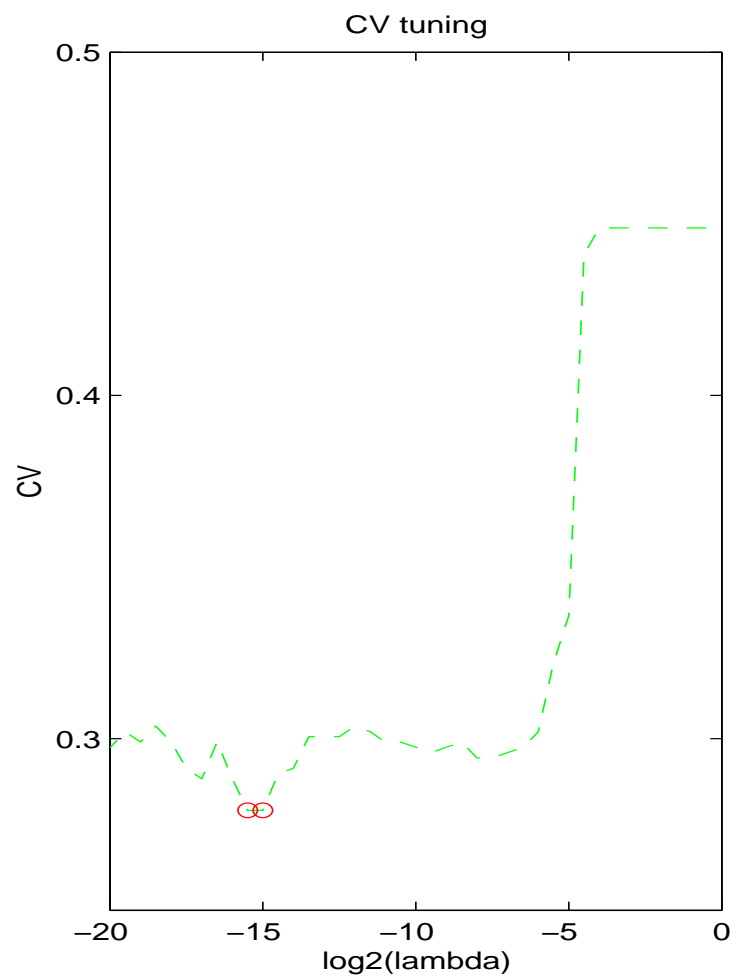


Figure 7: L_1 norm scores for the WESDR main effects model

- The approach picks out three most important variables *gly*, *sch*, and *dur*, that also appeared in Wahba et al. (1995) and Zhang et al. (2003).
- Empirical error: 0.2701

Summary

- Variable selection approach for nonlinear SVM, simultaneously fitting model and selecting components.
- Smoothing-Spline ANOVA model allows the flexibility of the model
- Able to choose high-order interactions between variables
- Simple Linear Programming.
- Extend to non-standard case and multi-class case
- More comparisons needed with existing approaches

References

FAN, J. and LI, R. Z. (2001). Variable selection via penalized likelihood. *J. Am. Statist. Assoc.* **96**
1348–1360.

App: Model Tuning - select tuning parameters

- λ controls the tradeoff between the likelihood \mathcal{L} to the training data and the sparsity of coefficients for f
- Let p and f be the true function; p_λ and f_λ be the estimates when using λ . Then **Kullback-Liebler (KL)** distance of p and p_λ is:

$$KL(p, p_\lambda) \equiv [p(f - f_\lambda) - (b(f) - b(f_\lambda))], \quad (6)$$

where $b(f) = \log(1 + e^f)$.

- **Comparative KL (CKL)** distance is defined by only keeping the terms related to λ ,

$$CKL(p, p_\lambda) = [- p f_\lambda + b(f_\lambda)] \quad (7)$$

App: Generalized Approximate Cross Validation (GACV)

- The sampling version of CKL is given by

$$CKL(p, p_\lambda) = \frac{1}{n} \sum_{i=1}^n [- p_i f_{\lambda i} + b(f_{\lambda i})] \quad (8)$$

- Leaving-out-one **Cross Validation (CV)** function for CKL is defined as

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n [- y_i f_{\lambda i}^{[-i]} + b(f_{\lambda i})]$$

- **GACV** is derived by using Leaving-out-one lemma (See Appendix for details). It is a calculable proxy for CKL.