Variable Selection in Spline ANOVA Models

Hao Helen Zhang

North Carolina State University

Regression Problems

- Continuous Responses

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \cdots, n$$

$x_i = (x_i^1, \cdots, x_i^d) \in R^d$,
$f(x)$ is the unknown regression function
ϵ_i i.i.d. noise with mean 0 and variance σ^2

- Discrete Responses – Binary case $y_i \in \{0, 1\}$
$p(x) = Prob(Y = 1|X = x)$, the logit function

$$f(x) = \log\left[\frac{p(x)}{1 - p(x)}\right]$$

- – To estimate $f(x)$ on the product domain $\mathcal{X} = \prod_{\alpha=1}^{\alpha} \mathcal{X}_\alpha$,
  – To select the important x's.

Variable Selection in Linear Models

Based on the ordinary least squares (OLS) estimates

- Best subset regression (exhaustive search)
  - expensive computation
  - the leaps and bounds by Furnival(1971) efficient for $d < 30$.
- Sequential search methods
  - forward selection; backward elimination
  - stepwise regression by Efroymson (1960)
- Criteria for inclusion/deletion
  - adjusted R^2, Mean Squared Error, Mallow's C_p
  - F-statistic, AIC, BIC

Shrinkage Methods – Penalized Least Squares Estimates

- Bridge regression by Frank & Friedman (1993)

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

  - LASSO by Tibshirani (1996) $q = 1$; Ridge regression $q = 2$.

- Nonnegative garrote by Breiman (1995)

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

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

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

Nonparametric Variable Selection

- Linear models
  - simple, easy to implement, easy to interprete, ...
  - lack of flexibility
- Nonlinear models
  - Classification and Regression Tree (CART)
  - Multivariate Adaptive Regression Spline (MARS)
  - $\cdots$

Smoothing Spline ANOVA Models

In a reproducing kernel Hilbert space (RKHS),

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} \mathcal{C}(y_i, f(x_i)) + \lambda J(f)$$

- provides a rigorous nonparametric framework for multivariate functional estimate
- how to add variable selection features?
  - Likelihood basis pursuit (LBP) for regression
  - SVM basis pursuit for classification

Various Penalties in Regularization Framework

- $\mathcal{C}$ is the fit to the data (e.g. least squares, likelihood, other loss functions)
- $J(f)$ is the penalty of the estimator
  - Ordinary smoothing spline model uses the **squared RKHS norm**
  - Likelihood basis pursuit uses the l_1 **norm of the basis coefficients**
  - COSSO by Lin & Zhang (2002) uses the **sum of component RKHS norms**
- λ is the tuning parameter

Basis Pursuit

1. decompose a signal into an overcomplete set of basis functions
2. choose the optimal decomposition: the smallest l_1 norm of the coefficients

Wavelet smoothing

- Donoho & Johnstone (1994) – SURE shrinkage (Stein Unbiased Risk Estimation)
- Chen & Donoho (1998), Sardy (1997), etc.

Algorithm in Solving the LBP

- Original problem
  - Objective: likelihood plus absolute values of the coefficients
  - The second part is non-differentiable at the origin
- Transformed into a constrained nonlinear optimization
  - Objective: convex and differentiable, nonlinear
  - Polyhedral constraints
- MATLAB, GAMS, MINOS, $\cdots$

Incorporating Categorical Variables

- Examples: sex, race, drinking/smoking history, marital status
- Categorical predictors: $Z = (Z^1, \ldots, Z^r) \in R^r$
- Assume each Z has C categories.
- The simplest case $C = 2$. Binary responses {T, F}
  - Define a mapping $\Phi : \{T, F\} \longrightarrow \{\frac{1}{2}, -\frac{1}{2}\}$ by

$$\begin{aligned} \Phi(z) &= \frac{1}{2} && \text{if} \quad z = T \\ &= -\frac{1}{2} && \text{if} \quad z = F \end{aligned}$$

- For $C > 2$, we need $C - 1$ mappings.

Main Effects Model (Modified)

The overcomplete set of $1 + d + r + dN$ basis functions:

$$\{1, b^{\alpha}(x), \Phi^{\gamma}(z) \equiv \Phi(z^{\gamma}), B^{\alpha}_{j*}(x)\},$$

for $\alpha = 1, \cdots, d, j* = 1, \cdots, N, \gamma = 1, \cdots, r$.

The likelihood basis pursuit model minimizes

$$\frac{1}{n}\sum_{i=1}^{n}[-l(y_i, f_i)] + \lambda_{\pi}(\sum_{\alpha=1}^{d}|b_{\alpha}| + \sum_{\gamma=1}^{r}|e_{\gamma}|) + \lambda_s \sum_{\alpha=1}^{d}\sum_{j*=1}^{N}|c_{\alpha j*}|,$$

subject to

$$
\begin{aligned}
f(x,z) &= b_0 + \sum_{\alpha=1}^{d} f_\alpha(x^\alpha) + \sum_{\gamma=1}^{r} e_\gamma \Phi^\gamma(z) \\
&= b_0 + \sum_{\alpha=1}^{d} b_\alpha b^\alpha(x) + \sum_{\gamma=1}^{r} e_\gamma \Phi^\gamma(z) \\
&\quad + \sum_{\alpha=1}^{d} \sum_{j*=1}^{N} c_{\alpha j*} B_{j*}^{\alpha}(x)
\end{aligned}
$$

Two-factor Interaction Models (Modified)

Minimize

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}[-l(y_i, f_i)] + \lambda\pi(\sum_{\alpha=1}^{d} |b_\alpha| + \sum_{\gamma=1}^{r} |e_\gamma|) \\
+ \quad & \lambda_{\pi\pi} (\sum_{\alpha<\beta} |b_{\alpha\beta}| + \sum_{\gamma<\theta} |e_{\gamma\theta}| + \sum_{\alpha=1}^{d} \sum_{\gamma=1}^{r} |P_{\alpha\gamma}|) \\
+ \quad & \lambda_{\pi s} (\sum_{\alpha\neq\beta} \sum_{j*=1}^{N} |c^{\pi s}_{\alpha\beta j*}| + \sum_{\alpha=1}^{d} \sum_{\gamma=1}^{r} \sum_{j*=1}^{N} |c^{\pi s}_{\alpha\gamma j*}|) \\
+ \quad & \lambda_s (\sum_{\alpha=1}^{d} \sum_{j*=1}^{N} |c^{s}_{\alpha j*}|) + \lambda_{ss} (\sum_{\alpha<\beta} \sum_{j*=1}^{N} |c^{ss}_{\alpha\beta j*}|),
\end{aligned}
$$

subject to

$$
\begin{aligned}
f(x,z) \quad &= \quad b_0 + \sum_{\alpha=1}^{d} b_\alpha b^\alpha(x) + \sum_{\gamma=1}^{r} e_\gamma \Phi^\gamma(z) + \sum_{\alpha<\beta} b_{\alpha\beta} b^\alpha(x) b^\beta(x) \\
&+ \quad \sum_{\gamma<\theta} e_{\gamma\theta} \Phi^\gamma(z) \Phi^\theta(z) + \sum_{\alpha=1}^{d} \sum_{\gamma=1}^{\theta} P_{\alpha\gamma} b^\alpha(x) \Phi^\gamma(z) \\
&+ \quad \sum_{\alpha=1}^{d} \sum_{j*=1}^{N} c^s_{\alpha j*} B^\alpha_{j*}(x) + \sum_{\alpha<\beta} \sum_{j*=1}^{N} c^{ss}_{\alpha\beta j*} B^\alpha_{j*}(x) B^\beta_{j*}(x) \\
&+ \quad \sum_{\alpha\neq\beta} \sum_{j*=1}^{N} c^{\pi s}_{\alpha\beta j*} B^\alpha_{j*}(x) b^\beta(x) \\
&+ \quad \sum_{\alpha=1}^{d} \sum_{\gamma=1}^{r} \sum_{j*=1}^{N} c^{\pi s}_{\alpha\gamma j*} B^\alpha_{j*}(x) \Phi^\gamma(z)
\end{aligned}
$$

Beaver Dam Eye Study (BDES)

- Funded by National Eye Institute (part of NIH)
- Collect information for the prevalence & incidence of age-related cataract, macular degeneration & diabetic retinopathy.
- Between 1987 and 1988, 5925 eligible people (age 43-84) identified in Beaver Dam, WI. Among them, $4926 (83.1\%)$ participated in the baseline exam.
- Two five-year follow-ups after the baseline examination

Response Variable Y: Five-year mortality for non-diabetic patients. Y was defined to be 1 if a patient participated the baseline examination and died prior to the start of the first 5-year follow-up.

Continuous Variables

- X_1: *pky* - pack years smoked (packs per day/20)*years
- X_2: *sch* - highest year of school/college completed
- X_3: *inc* - total household personal income
- X_4: *bmi* - body mass index, kg/m^2
- X_5: *glu* - glucose (serum), mg/dL
- X_6: *cal* - calcium (serum), mg/dL
- X_7: *chl* - cholesterol (serum), mg
- X_8: *hgb* - hemoglobin (blood), g/dL
- X_9: *sys* - systolic blood pressure, mmHg
- X_{10}: *age* - age at baseline examination, years

Categorical Variables

- Z_1: *cv* - history of cardiovascular disease. (0 =No,1 =yes)
- Z_2: *sex* - sex. (0 =Female, 1 = Male)
- Z_3: *hair* - hair color. (0 =Blond/Red, 1 = Brown/Black)
- Z_4: *hist* - history of heavy drinking. (0 = Never, 1 = Past/Currently)
- Z_5: *nout* - winter leisure time. (0 = Spent mostly indoors, 1 = half/mostly spent outdoors)
- Z_6: *mar* - marital status. (0 = Never/Separated/Divorced/Widowed, 1 = Married)
- Z_7: *sum* - part of day spent outdoors in summer. ($0 =< 1/4$ of the day, $1 >= 1/4$ of the day)
- Z_8: *vtm* - vitamin use. (0 = Never, 1 = Past/Current)

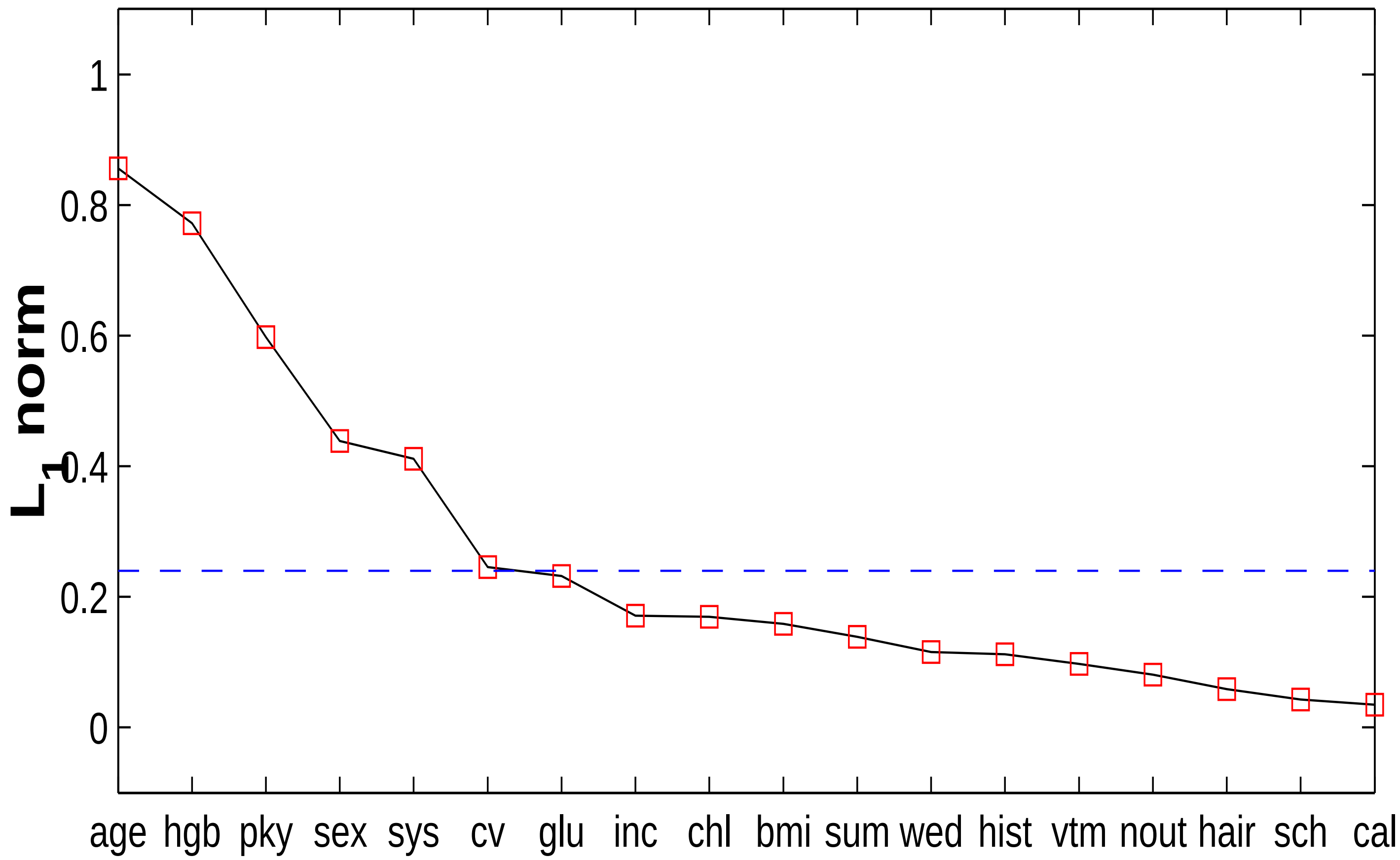

Figure 1: L_1 norm scores for the main effects model (BDES)

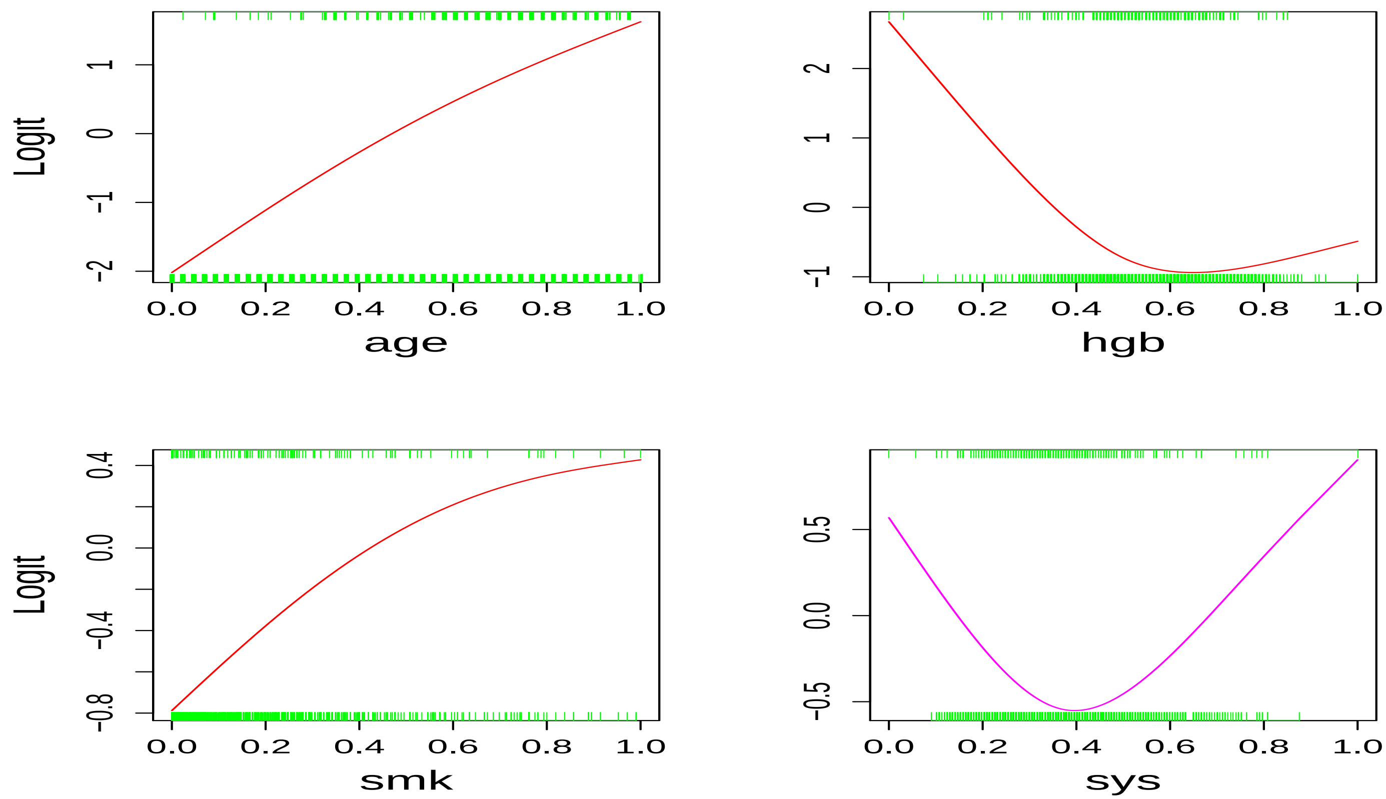

Figure 2: Estimated univariate logit component for important variables (BDES)

Classification Problem

Consider two-category classification first Class label $y_i \in \{-1, 1\}$
$x_i = (x_i^1, \cdots, x_i^d) \in R^d$
Estimate the boundary function $f(x)$

The classification rule is

$$\text{sign}[f(x)] : R^d \to \{\pm 1\}$$

Support Vector Machines

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} [1 - y_i f(x_i)]_+ + \lambda ||f||^2_{\mathcal{H}_K}$$

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

- The loss function $[1 - yf(x)]_+$ is called “hinge-loss”.
- The classification rule is sign$[f(x)]$.

SVM Basis Pursuit

- Main effects model

$$\min \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(x_i)]_+ + \lambda(\sum_{\alpha=1}^{d}|b_\alpha| + \sum_{\alpha=1}^{d}\sum_{j*=1}^{N}|c_{\alpha j*}|), \quad (1)$$

  subject to

$$f(x) = b_0 + \sum_{\alpha=1}^{d} b_\alpha b^\alpha(x) + \sum_{\alpha=1}^{d}\sum_{j*=1}^{N} c_{\alpha j*} B_{j*}^{\alpha}(x),$$

- Two-way interaction model

- Advantage of Computation: Linear programming with linear constraints.

Example 1

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

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

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

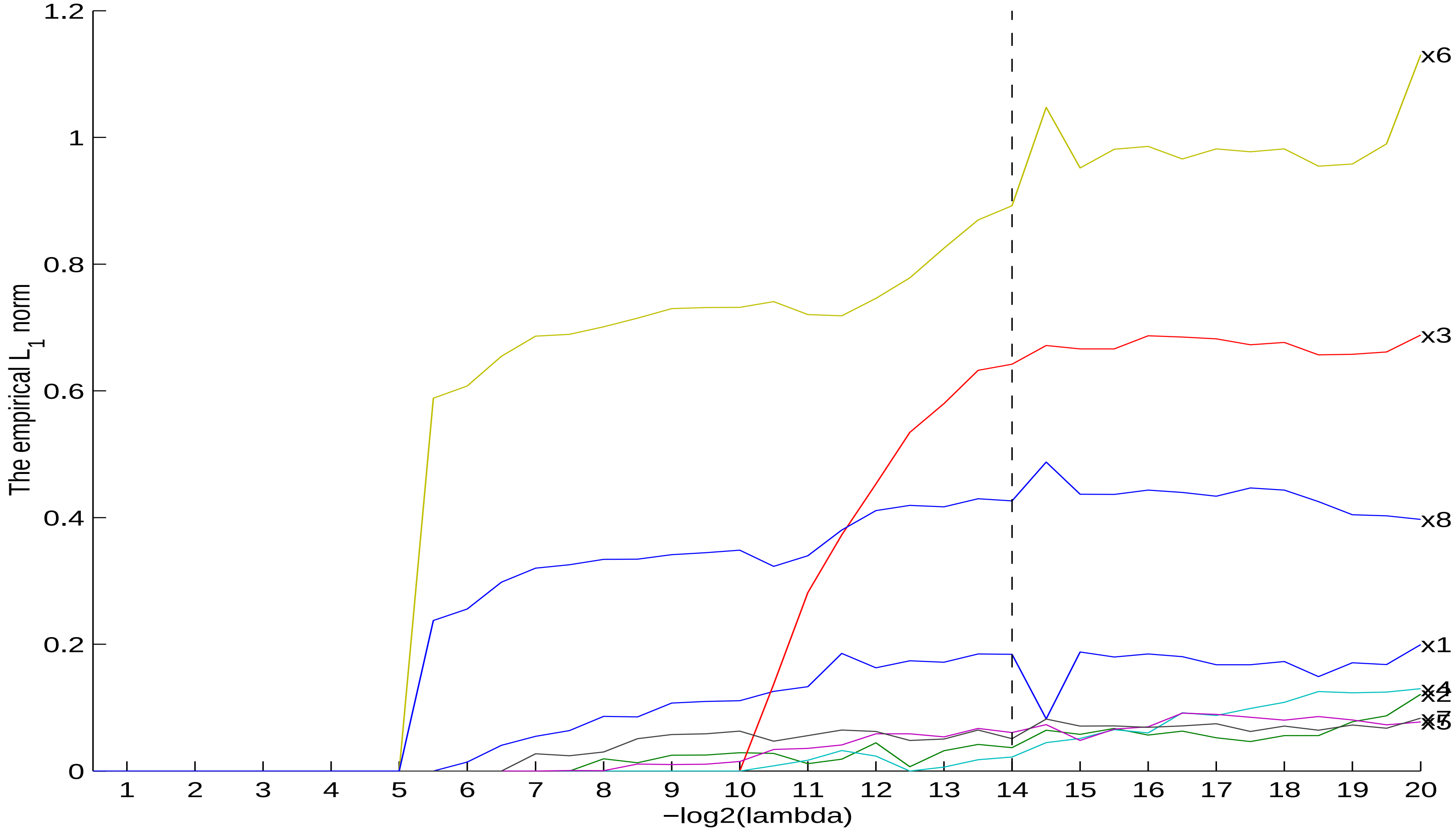

Figure 3: The empirical L_1 norm for the estimated components against λ

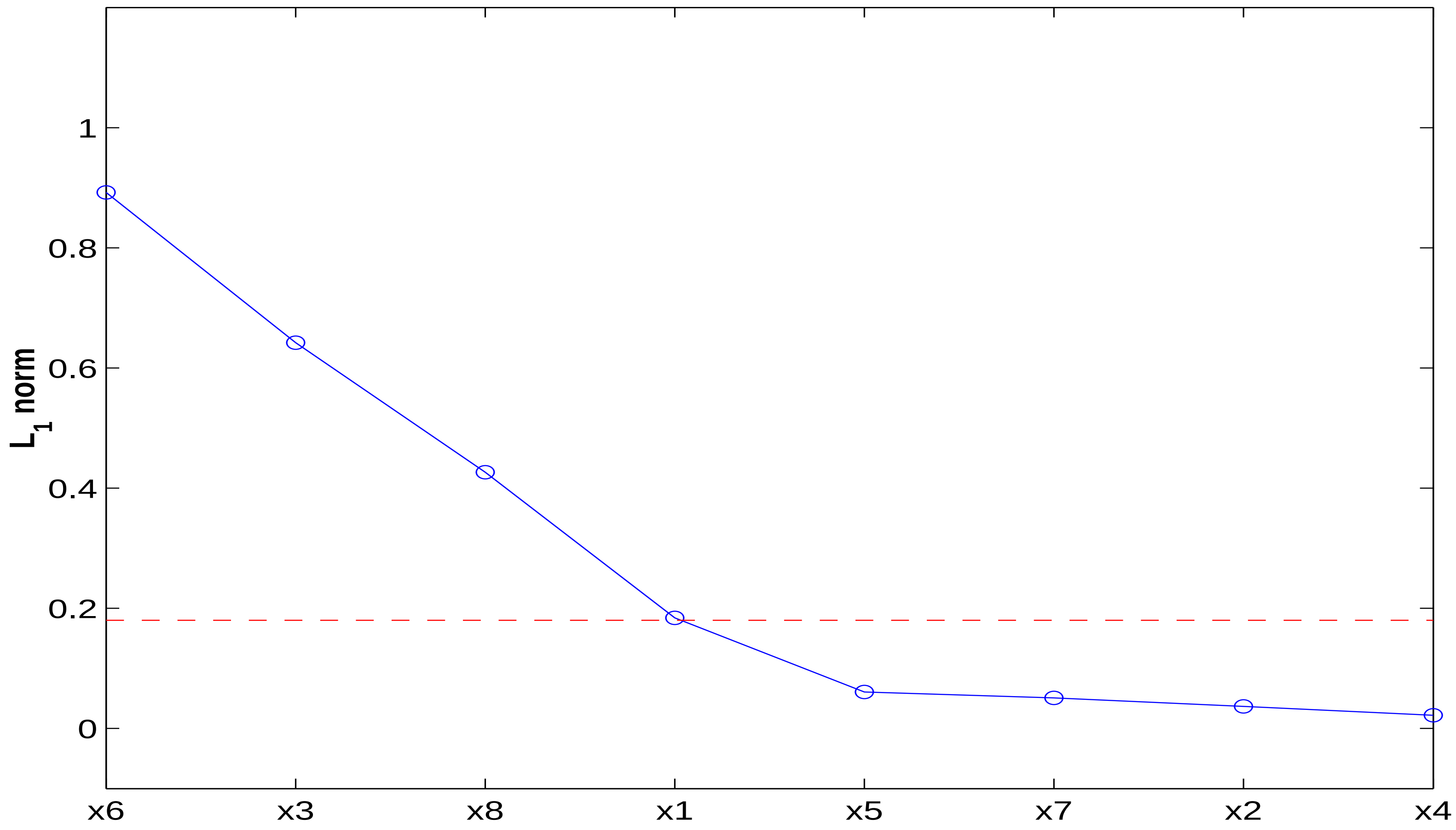

Figure 4: L_1 norm for individual variables (using $\hat{\lambda}_{CV}$)

Table 4: The test error given by various classification rules

	Bayes Rule	SVM-BP	SVM
test error	0.2256	0.2400	0.3902

Summary

- Likelihood basis pursuit
  - Spline ANOVA framework
  - Simultaneous estimation and variable/component selection
- SVM basis pursuit
  - Two-category classification
  - Multiple-category classification