Kernel Methods in a Regularization Framework

Yoonkyung Lee
Department of Statistics
The Ohio State University

October 31, 2008
Korean Statistical Society Meeting
Chung-Ang University, Seoul

Overview

- Part I:
  Introduction to Kernel Methods for Statistical Learning and Modeling
- Part II:
  Theory of Reproducing Kernel Hilbert Spaces Methods
- Part III:
  Regularization Approach to Feature Selection

Part I: Introduction to Kernel Methods for Statistical Learning and Modeling

- Statistical learning problems
- Methods of regularization
- Smoothing splines
- Support vector machines
- Statistical issues

Prediction of Annual Household Income

http://www-stat.stanford.edu/∼tibs/ElemStatLearn/

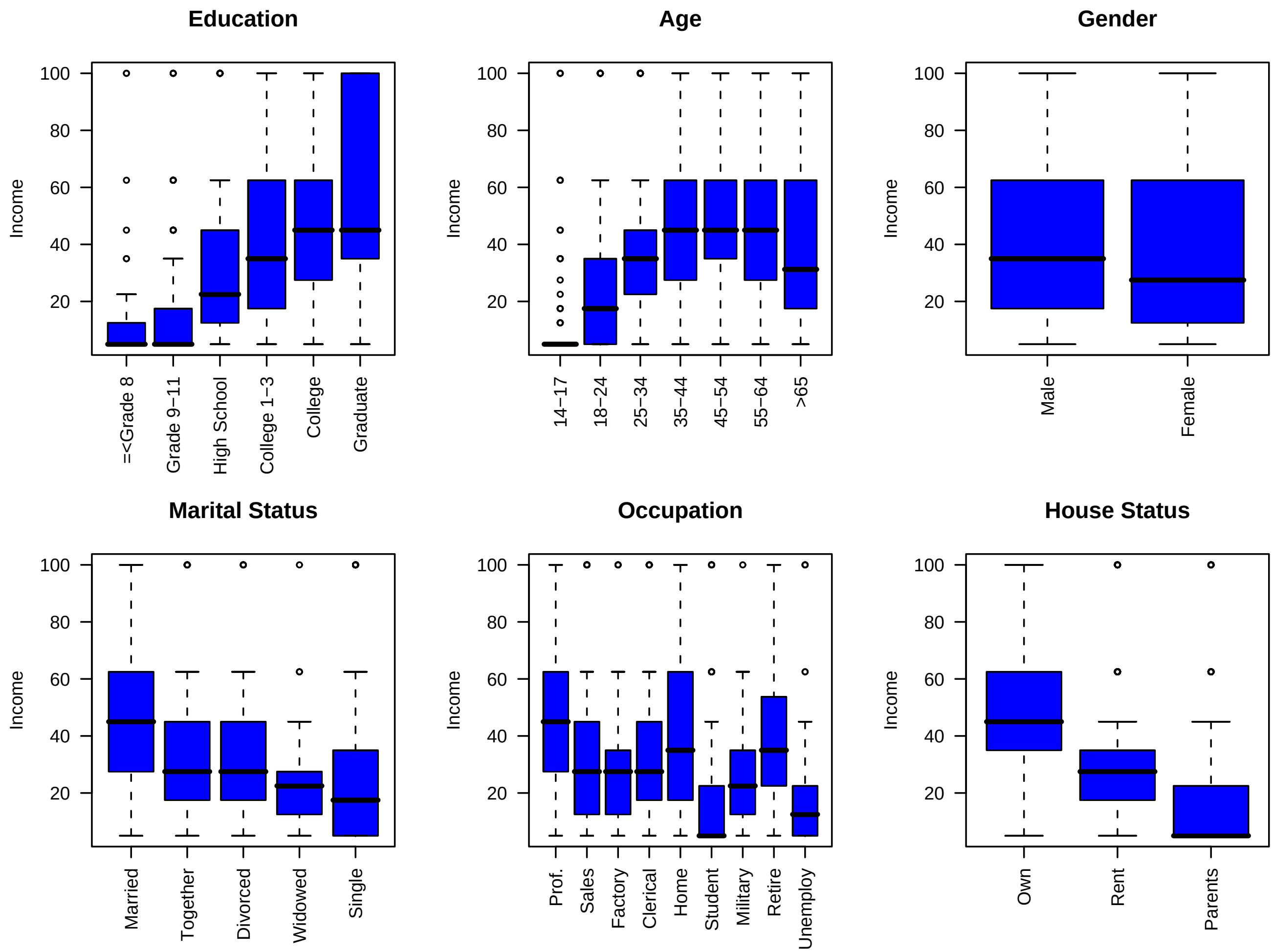


Figure: Boxplots of the annual household income with education, age, gender, marital status, occupation, and householder status out of 13 demographic attributes in the data

Cancer Diagnosis with Microarray Data

- Microarrays measure relative amount of mRNAs of (tens of) thousands of genes.
- Golub et al.(*Science*, 1999): Acute leukemia data
  ALL (acute lymphoblastic leukemia) with subtype B-cell and T-cell lineage, and AML (acute myeloid leukemia)

Acute Leukemia Gene Expression Profiles

Patient	$gene_1$	$gene_2$	$\cdots$	$gene_{7129}$	*class*
1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,7129}$	ALL − B
2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,7129}$	ALL − B
$\vdots$	$\vdots$			$\vdots$	$\vdots$
20	$x_{20,1}$	$x_{20,2}$	$\cdots$	$x_{20,7129}$	ALL − T
$\vdots$	$\vdots$			$\vdots$	$\vdots$
28	$x_{28,1}$	$x_{28,2}$	$\cdots$	$x_{28,7129}$	AML
$\vdots$	$\vdots$			$\vdots$	$\vdots$
38	$x_{38,1}$	$x_{38,2}$	$\cdots$	$x_{38,7129}$	AML

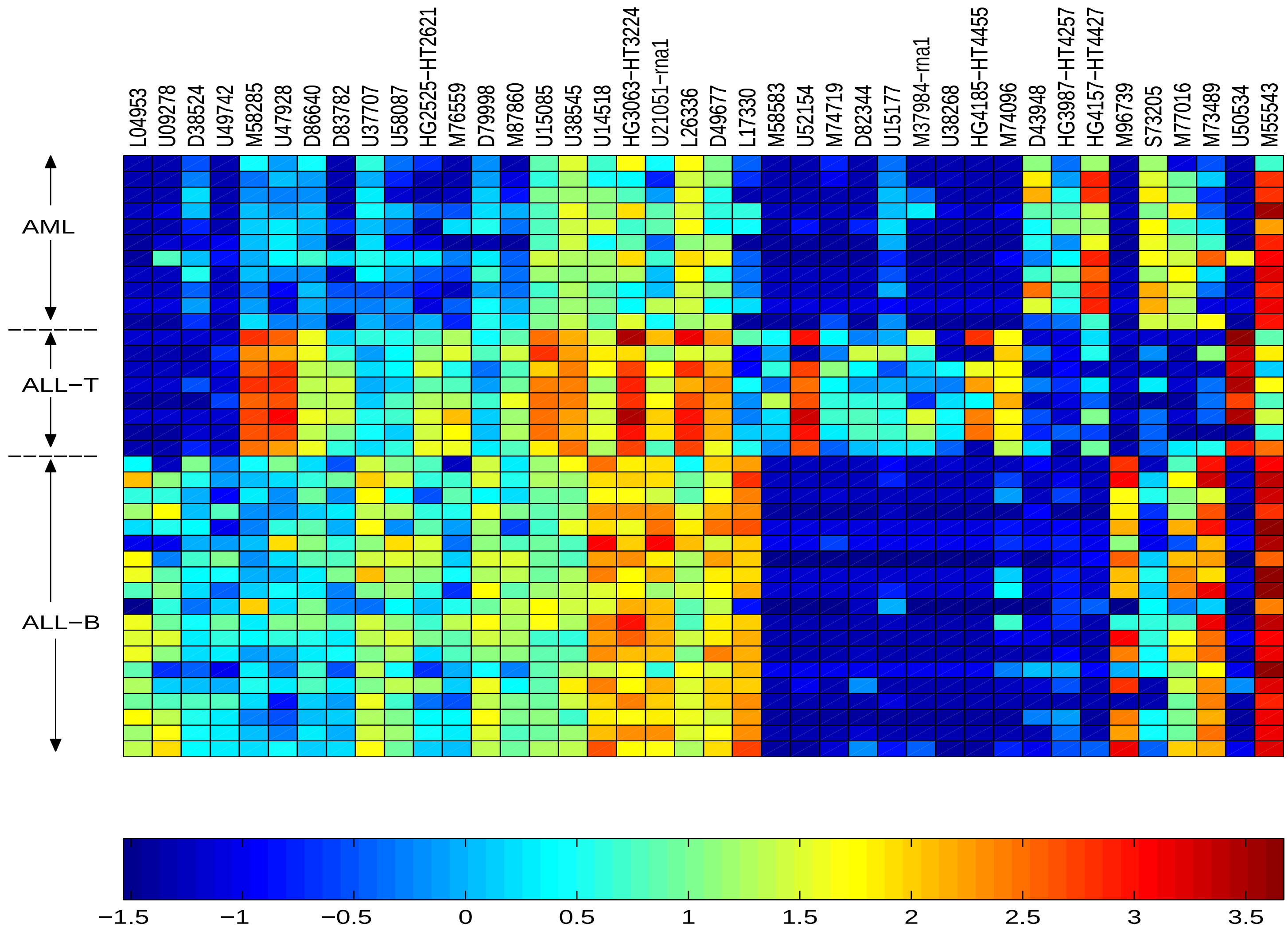


Figure: The heat map shows the expression levels of 40 most important genes for the training samples when they are appropriately standardized. Each row corresponds to a sample, which is grouped into the three classes, and the columns represent genes. The 40 genes are clustered in a way the similarity within each class and the dissimilarity between classes are easily recognized.

Handwritten Digit Recognition

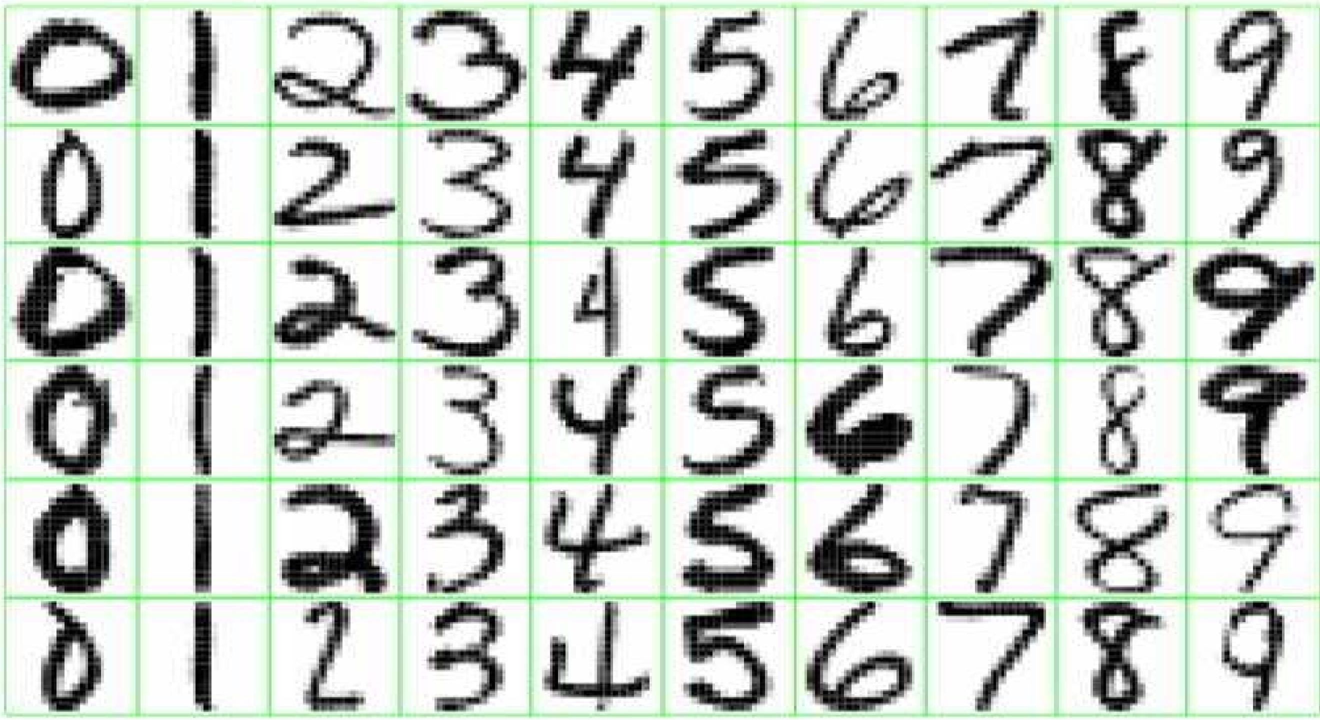


Figure 1.2: *Examples of handwritten digits from U.S. postal envelopes.*

Figure: 16 $\times$ 16 gray scale images

Statistical Learning

- Multivariate function estimation
- A training data set $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$
- Learn functional relationship f between $\boldsymbol{x} = (x_1, \ldots, x_p)$ and y from the training data, which can be generalized to novel cases.
  e.g. $f(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$
- Examples include
  Regression: continuous $y \in \mathbb{R}$, and
  Classification: categorical $y \in \{1, \ldots, k\}$.

Goodness of a Statistical Procedure for Learning

- Accurate prediction with respect to a given loss $\mathcal{L}(y, f(\boldsymbol{x}))$
  e.g. $\mathcal{L}(y, f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2$ for regression
- Flexible (nonparametric) and data-adaptive
- Interpretability
  e.g. subset (variable/feature) selection
- Computational ease for large p (high dimensional input) and n (large sample)

Large Model Space

- Large number of variables for high dimensional data
- Large number of basis functions for nonparametric modeling
- Need to deal with a large parameter (or model) space.
- Classical maximum likelihood estimation (MLE) or empirical risk minimization (ERM) no longer works as the solution may not be well-defined or there may be infinitely many solutions that overfit data.
- How to explore the large model space for stable model fitting and prediction?

Regularization

- Tikhonov regularization (1943):
  solving ill-posed integral equation numerically
- Process of modifying ill-posed problems by introducing additional information about the solution
- Modification of the maximum likelihood principle or empirical risk minimization principle
  (Bickel & Li 2006)
- Smoothness, sparsity, small norm, large margin, ...
- Bayesian connection

Methods of Regularization (Penalization)

Find $f(\boldsymbol{x}) \in \mathcal{F}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda J(f).$$

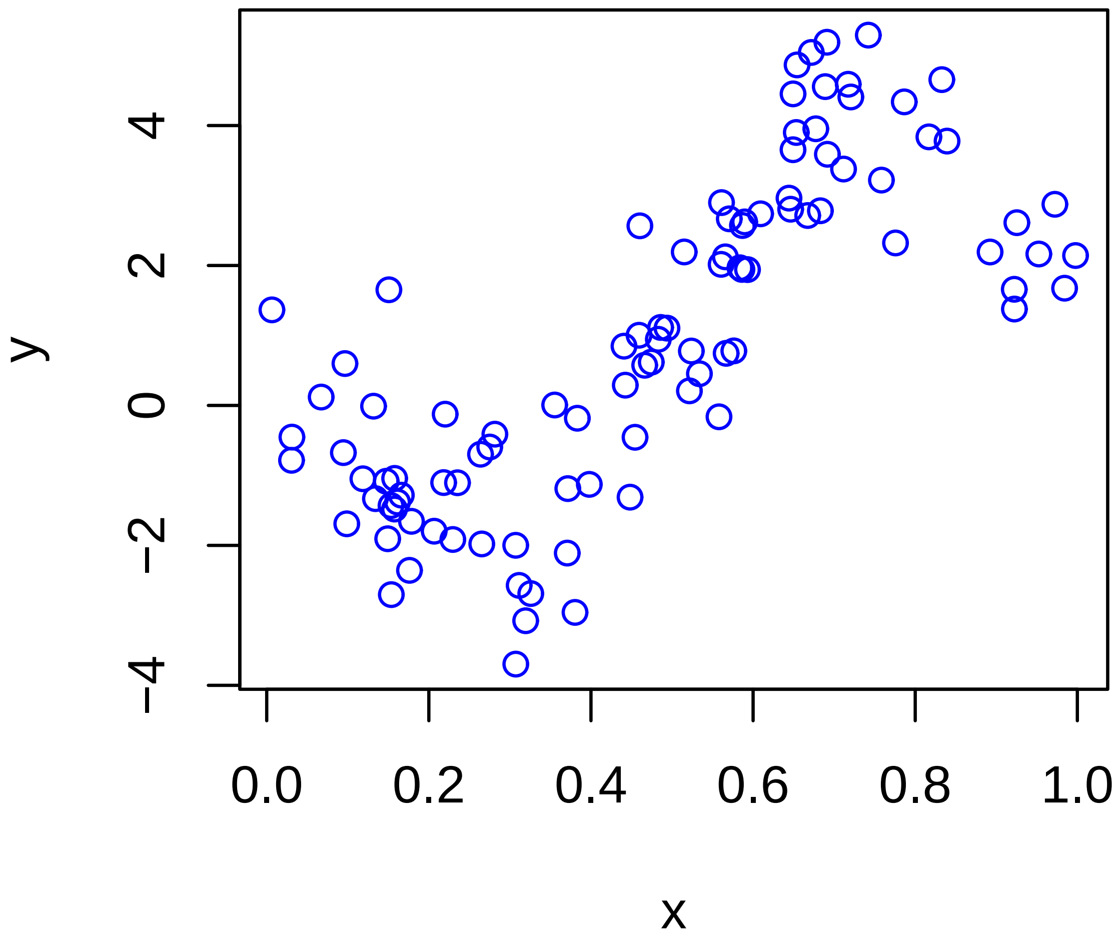
- Empirical risk + penalty
- $\mathcal{F}$: a class of candidate functions
- $J(f)$: complexity of the model f
- $\lambda > 0$: a regularization parameter
- Without the penalty $J(f)$, ill-posed problem

Examples of Regularization Methods

- Ridge regression (Hoerl and Kennard 1970)
- LASSO (Tibshirani 1996)
- Smoothing splines (Wahba 1990)
- Support vector machines (Vapnik 1998)
- Regularized neural network, boosting, logistic regression, ...

Nonparametric Regression

$$y_i = f(x_i) + \epsilon_i \text{ for } i = 1, \ldots, n \text{ where } \epsilon_i \sim N(0, \sigma^2)$$



Smoothing Splines

Wahba (1990), Spline Models for Observational Data.

Find $f(x) \in W_2[0,1]$
$= \{f : \ f, f' \text{ absolutely continuous, and } f'' \in L_2\}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx.$$

- $J(f) = \int_0^1 (f''(x))^2 dx = \|P_1 f\|^2$: curvature of f
- $\lambda \to 0$: interpolation
- $\lambda \to \infty$: linear fit
- $0 < \lambda < \infty$: piecewise cubic polynomial with two continuous derivatives

Small λ: Overfit

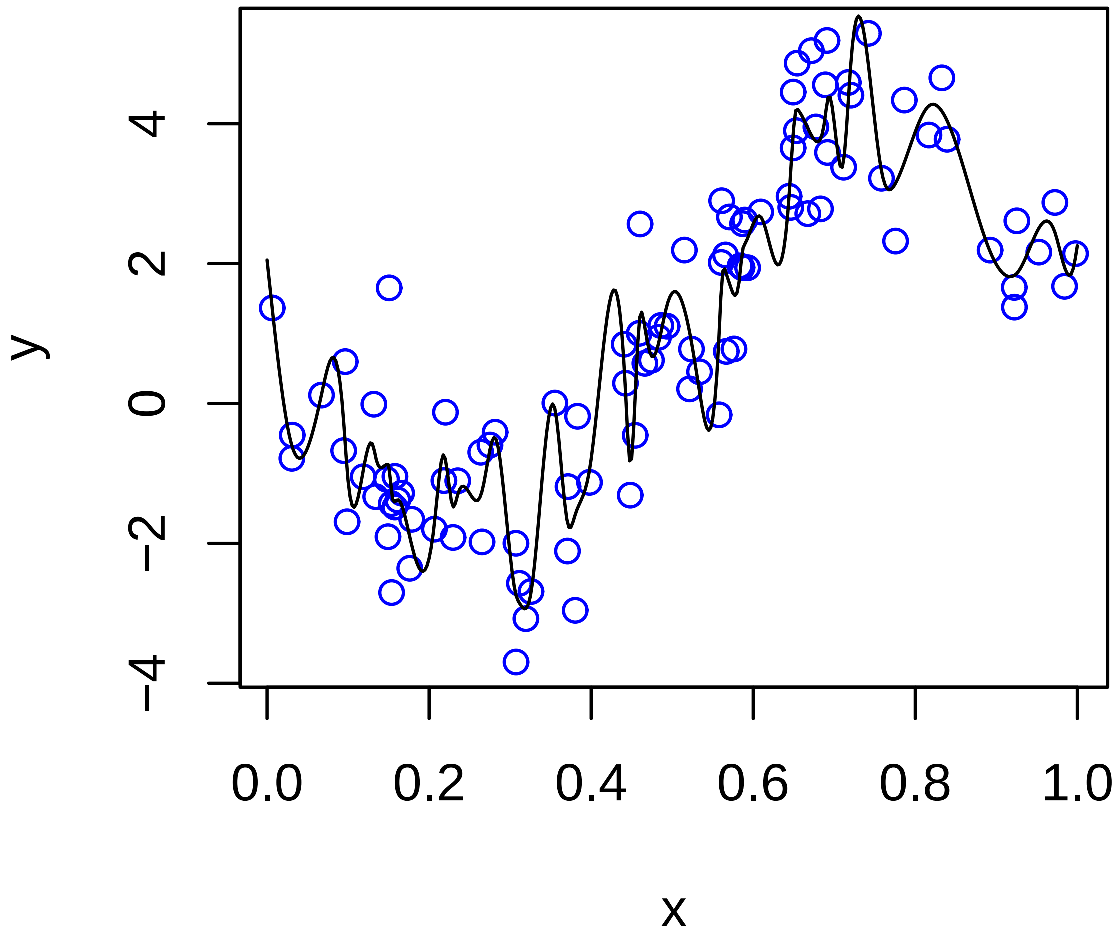


Figure: $\lambda \longrightarrow 0$: interpolation

Large λ: Underfit

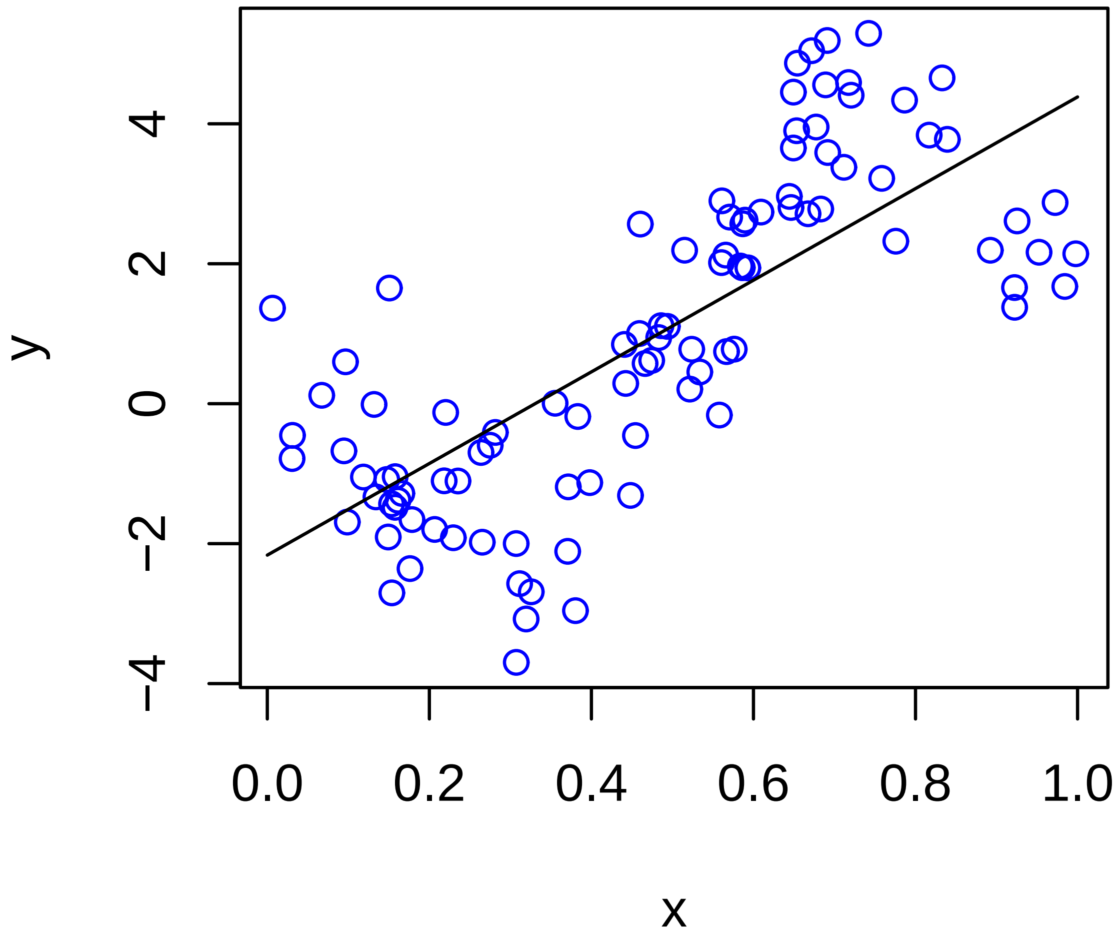


Figure: $\lambda \to \infty$: linear fit

Moderate λ

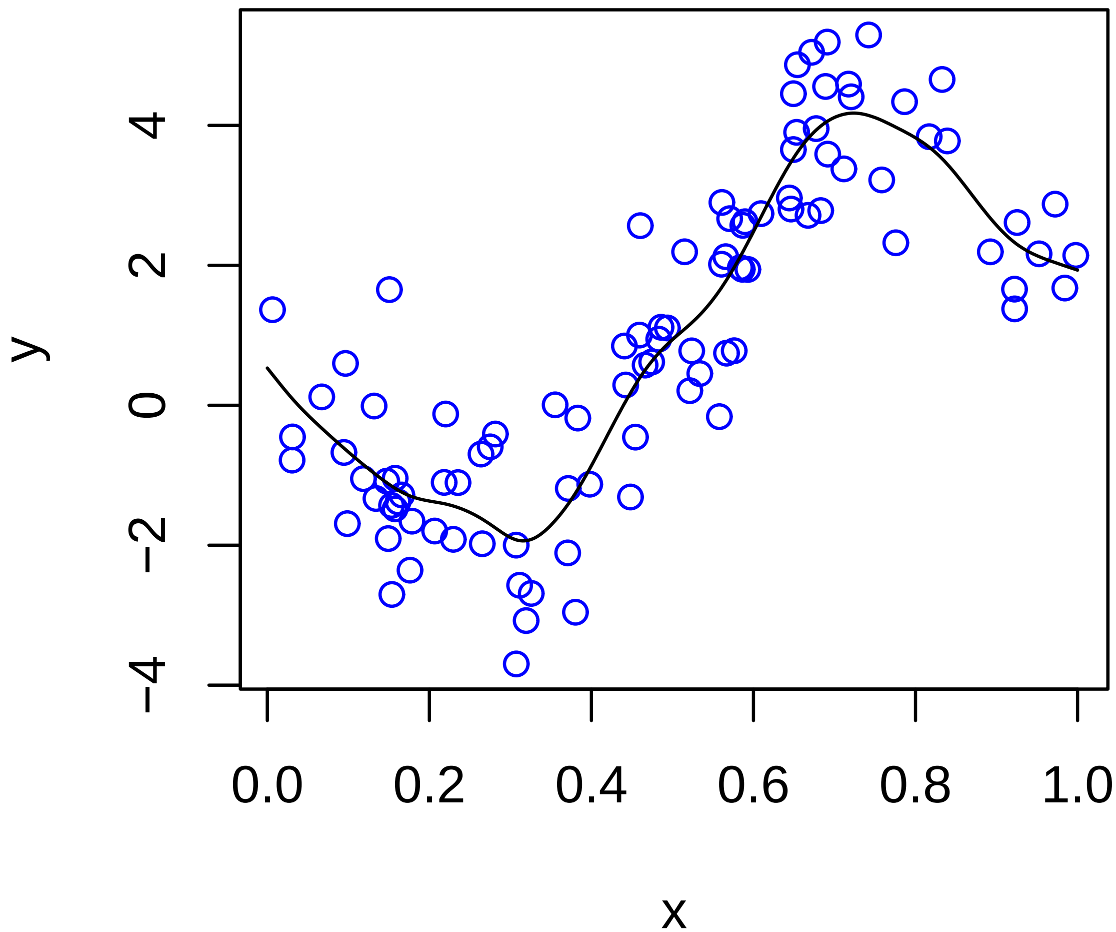


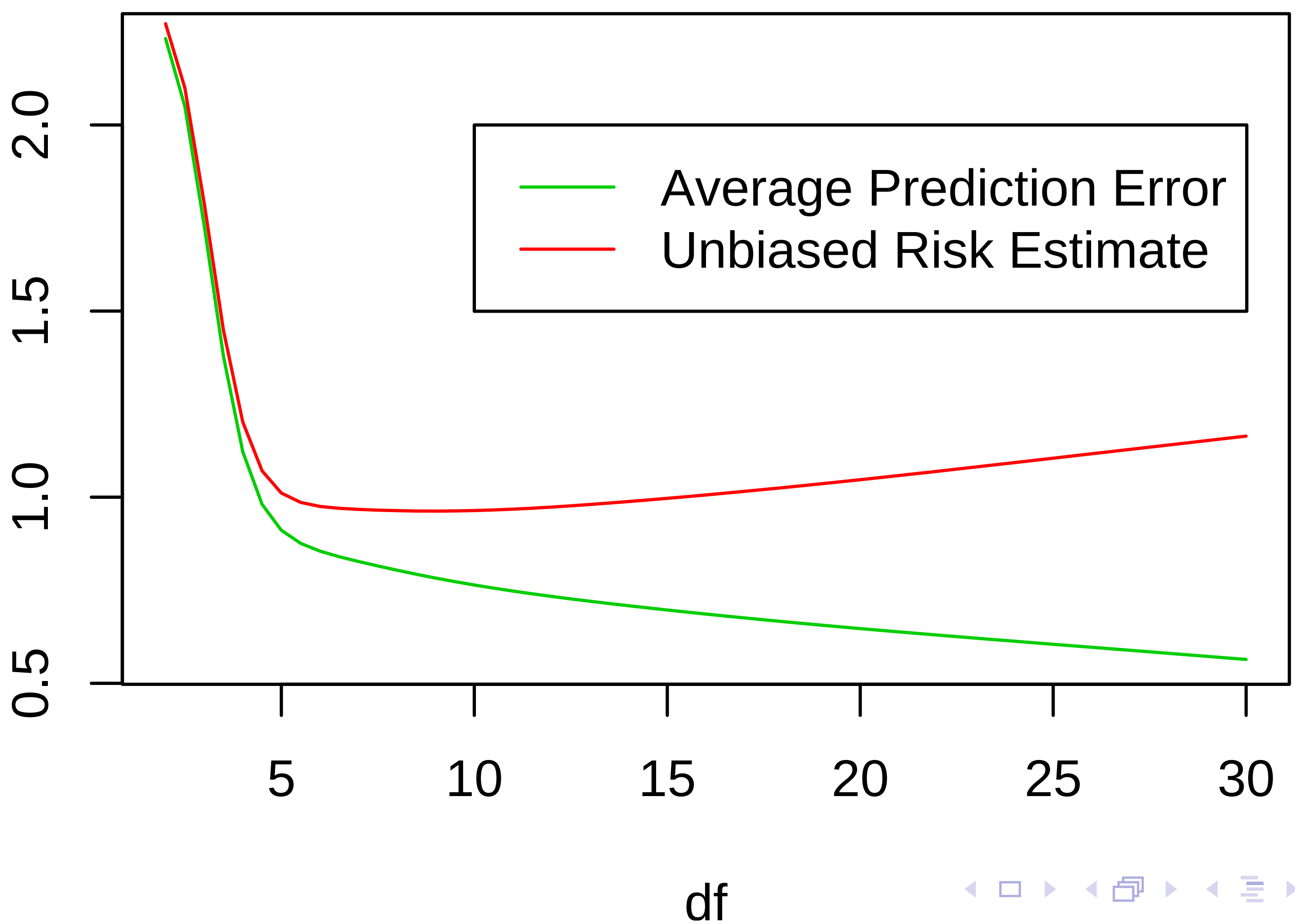
Figure: $0 < \lambda < \infty$: piecewise cubic polynomial with two continuous derivatives

Risk Estimate

- Risk: $E[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_\lambda(x_i)-f(x_i)\right)^2]$
- The Mallows-type criterion:

$$U(\lambda)=\frac{1}{n}\|(I-A(\lambda))y\|^2+2\frac{\sigma^2}{n}tr[A(\lambda)].$$

- $d.f.(\hat{\lambda}_U)=9.$



Classification

- $\boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$
- $y \in \mathcal{Y} = \{1, \ldots, k\}$
- Learn a rule $\phi : \mathbb{R}^p \to \mathcal{Y}$ from the training data $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$.
- The 0-1 loss function:

$$\mathcal{L}(y, \phi(\boldsymbol{x})) = I(y \neq \phi(\boldsymbol{x}))$$

Separating Hyperplane

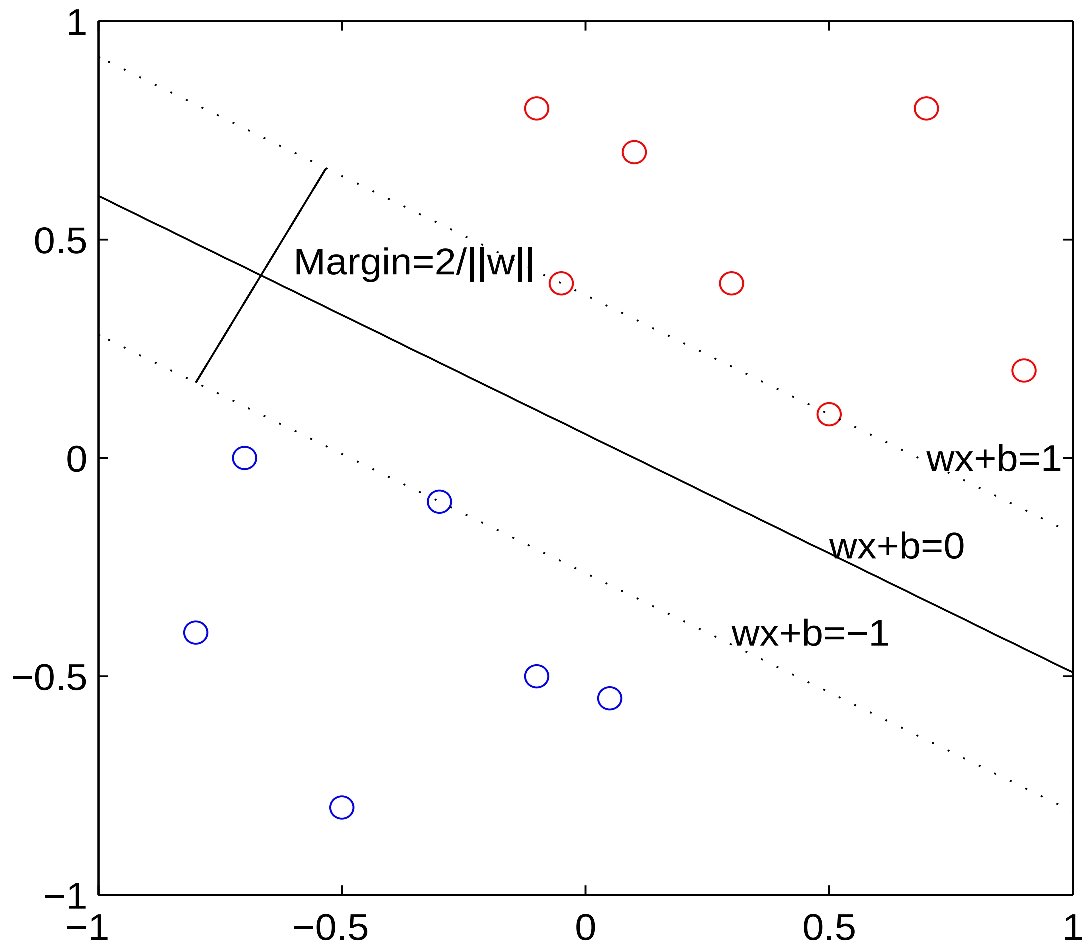


Figure: $y = +1$ in red and $y = -1$ in blue

Support Vector Machines

Boser, Guyon, & Vapnik (1992)
Vapnik (1995), *The Nature of Statistical Learning Theory.*

- $y_i \in \{-1, 1\}$, class labels in binary case
- $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ (real-valued discriminant function)
- Separating hyperplane with the maximum margin:

  f minimizing $\|\boldsymbol{w}\|^2$
  subject to $y_i f(\boldsymbol{x}_i) \geq 1$ for all $i = 1, \ldots, n$
- Classification rule: $\phi(\boldsymbol{x}) = sign(f(\boldsymbol{x}))$

Linear SVM in Non-Separable Case

- Relax the separability condition $y_i f(\boldsymbol{x}_i) \geq 1$.
- Hinge loss: $\mathcal{L}(y, f(\boldsymbol{x})) = (1 - yf(\boldsymbol{x}))_+$ where $(t)_+ = \max(t, 0)$.
- If $yf(\boldsymbol{x}) < 1$, the loss is proportional to the distance of $\boldsymbol{x}$ from the soft margin $yf(\boldsymbol{x}) = 1$
- Find $f \in \mathcal{F} = \{f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b \mid \boldsymbol{w} \in \mathbb{R}^p \text{ and } b \in R\}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(\boldsymbol{x}_i))_+ + \lambda \|\boldsymbol{w}\|^2,$$

where $J(f) = J(\boldsymbol{w}^\top \boldsymbol{x} + b) = \|\boldsymbol{w}\|^2$.

Hinge Loss

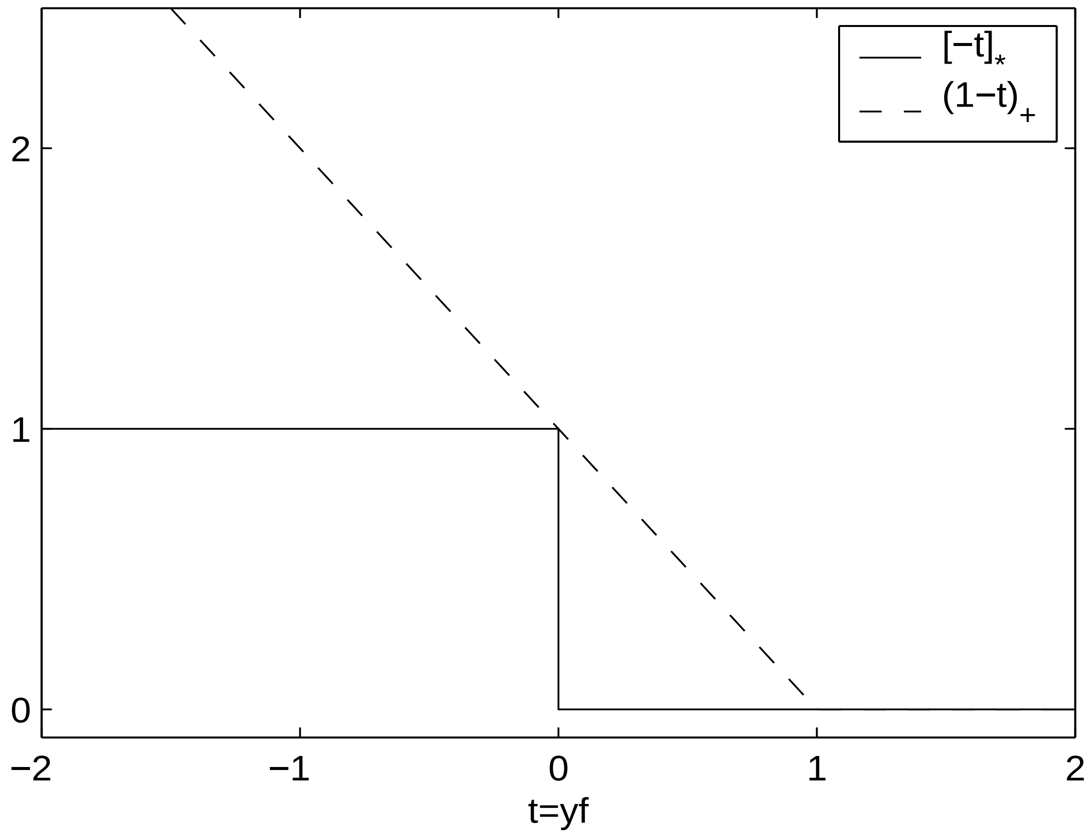


Figure: $(1 - yf(\boldsymbol{x}))_+$ is an upper bound of the misclassification loss function $I(y \neq \phi(\boldsymbol{x})) = [-yf(\boldsymbol{x})]_* \leq (1 - yf(\boldsymbol{x}))_+$ where $[t]_* = I(t \geq 0)$ and $(t)_+ = \max\{t, 0\}$.

Nonlinear SVM

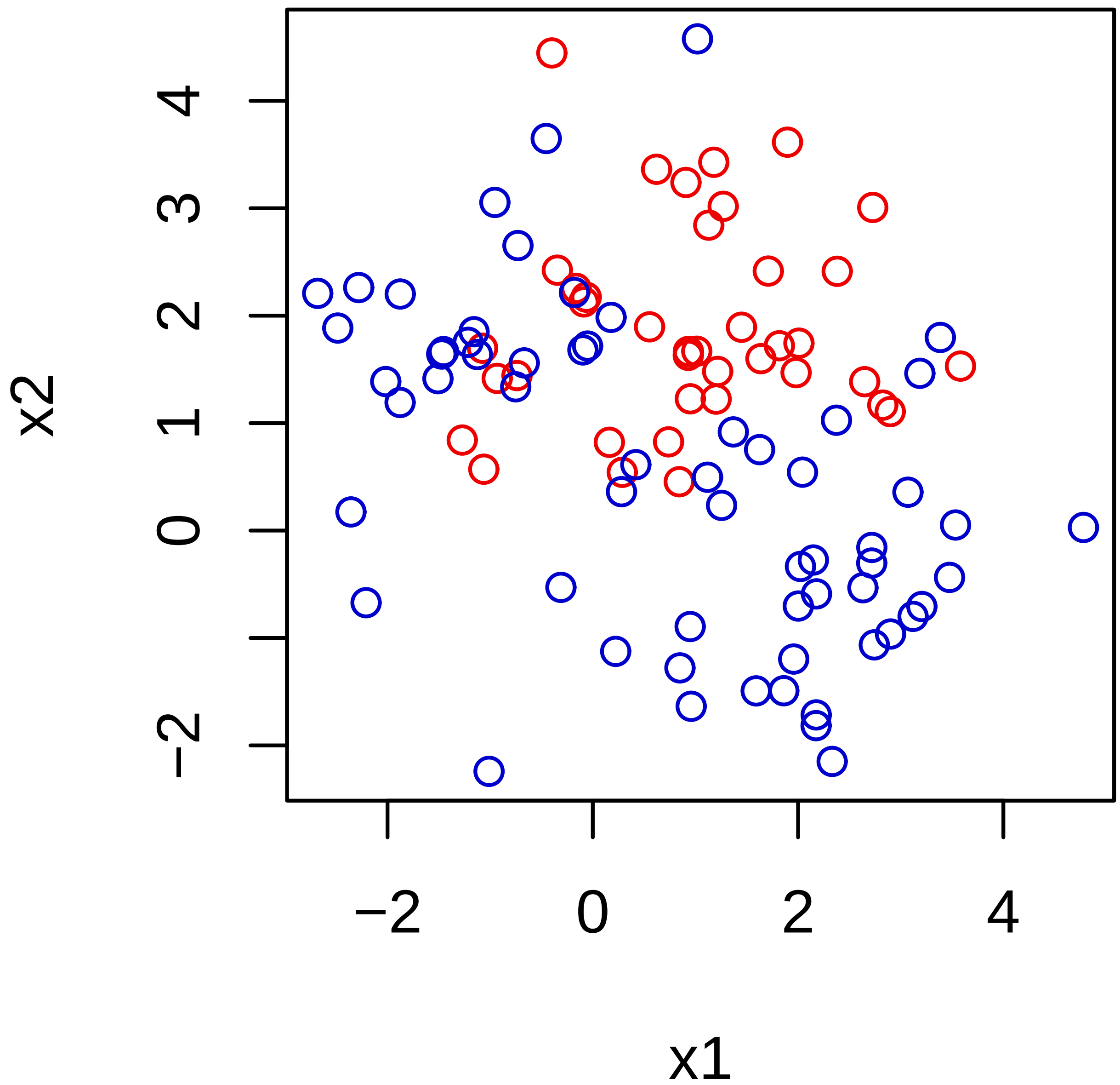
- Linear SVM solution:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} c_i \boldsymbol{x}_i^\top \boldsymbol{x} + b$$

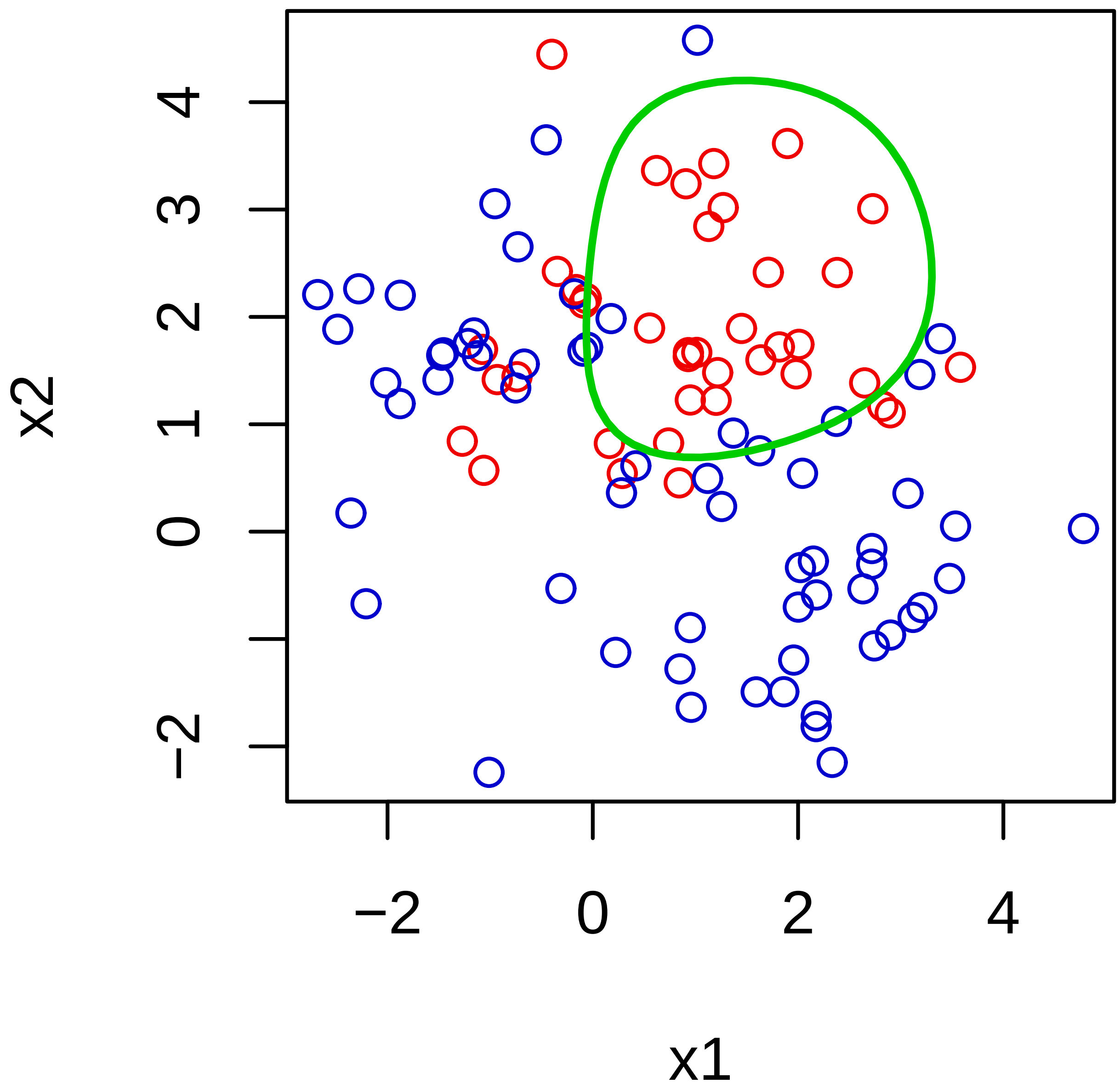
- Replace the Euclidean inner product $\boldsymbol{x}^\top \boldsymbol{x}'$ with $K(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x})^\top \Phi(\boldsymbol{x}')$ for a mapping Φ from $\mathbb{R}^p$ to a higher dimensional 'feature space.'
- Nonlinear kernels:
  $K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^d$, $\exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2 / 2\sigma^2)$, ...
  e.g. For $p = 2$ and $\boldsymbol{x} = (x_1, x_2)$,
  $\Phi(\boldsymbol{x}) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2)$ gives
  $K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^2$.

Classification

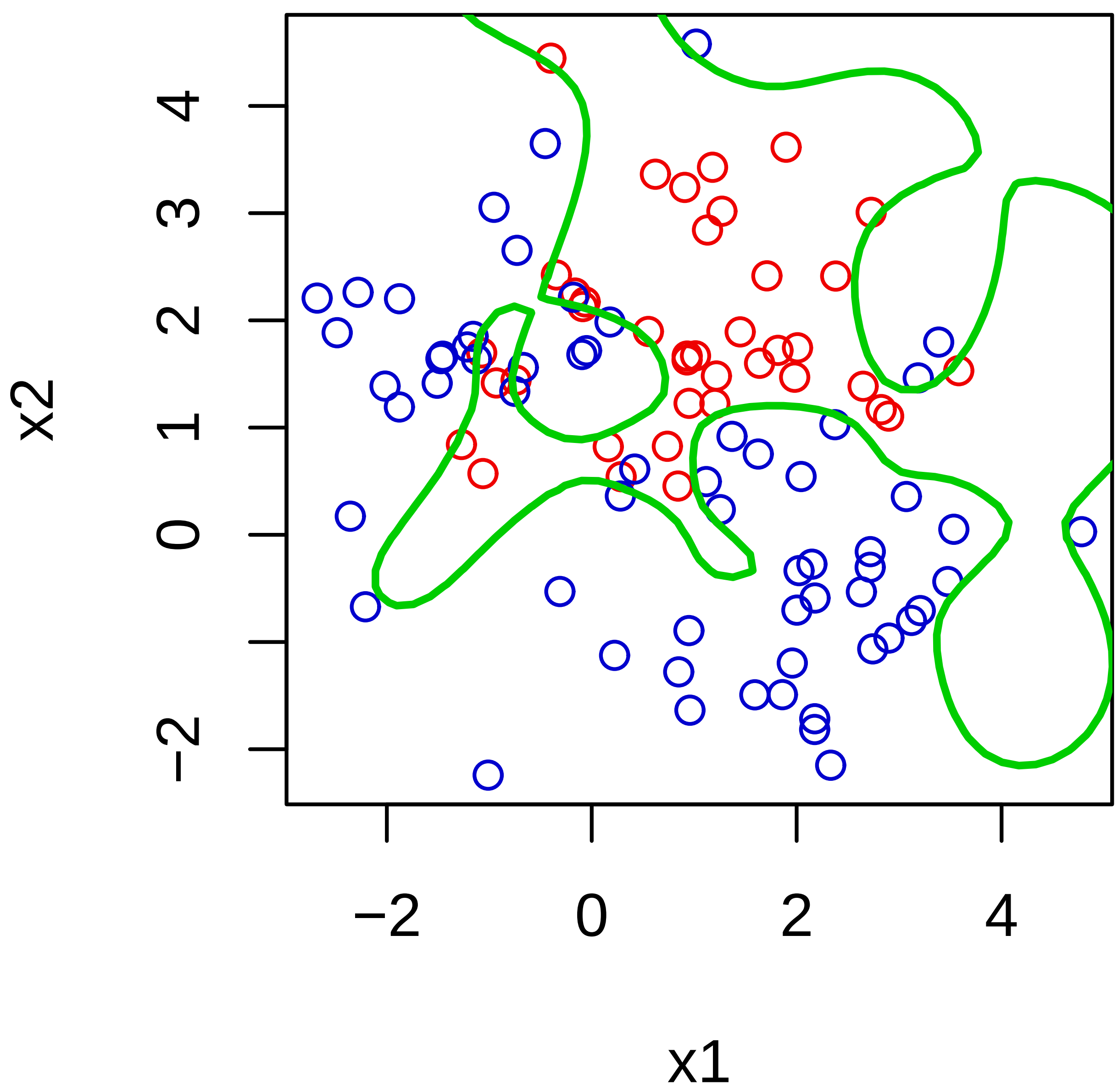
$$y_i \in \{1 : red,\ -1 : blue\}$$



Classification Boundary with a Large λ



Classification Boundary with a Small λ



Test Error Rates

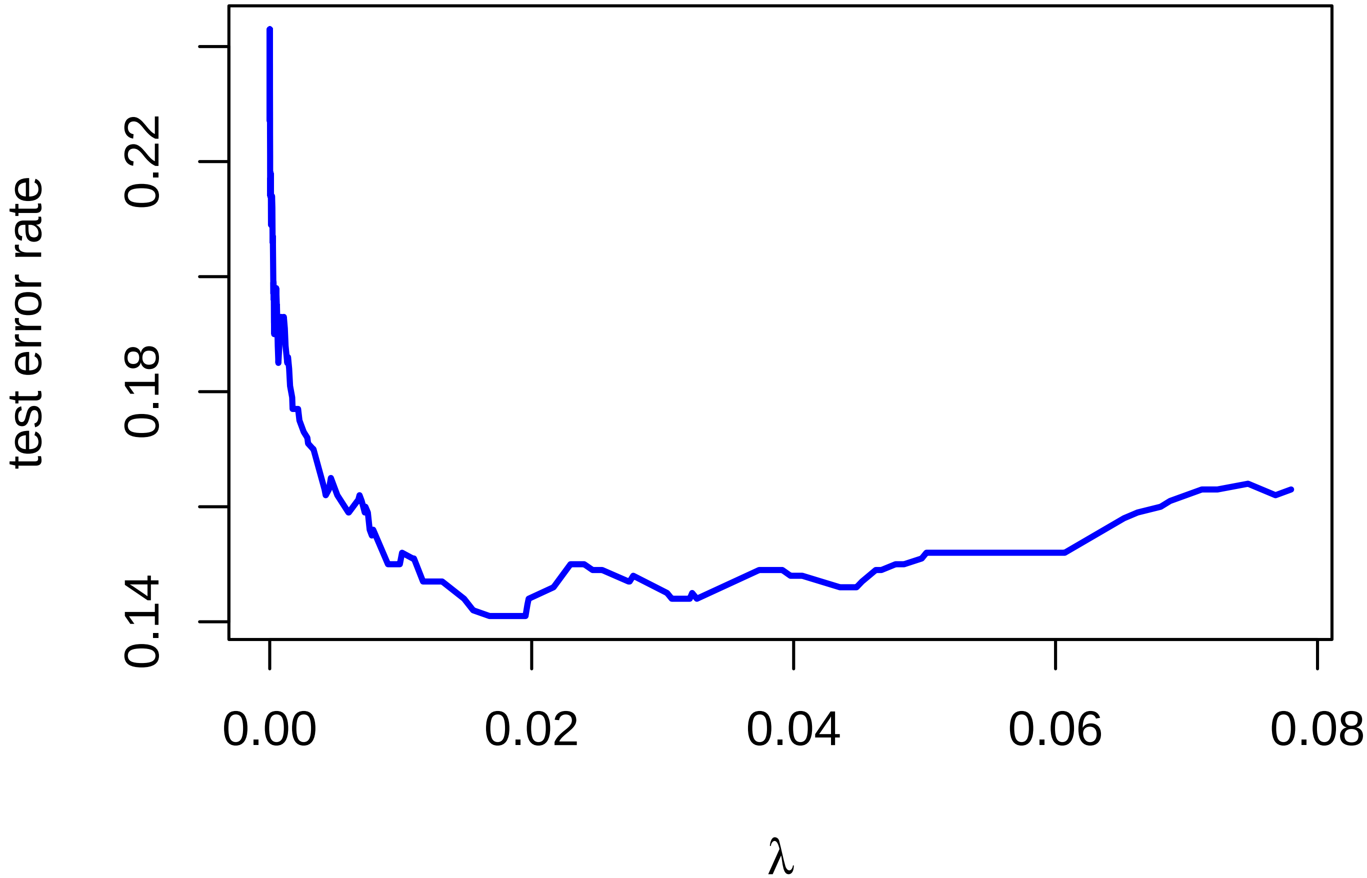


Figure: Error rate over 1,000 test cases as a function of λ.

Statistical Issues

- Risk or generalization error estimation
- Model selection/averaging - choice of tuning parameter(s) (CV, GCV, resampling, risk bounds, ...)
- Variable or feature selection
- Computation
  - e.g. Least squares problem for Smoothing splines and Quadratic Programming for SVM
  - Need to solve a family of optimization problems indexed by λ.
  - Use the characteristics of regularized solutions for efficient algorithms.

Summary

- Many statistical learning methods can be cast in a regularization framework.
- Examples include Smoothing Splines and Support Vector Machines.
- Regularization entails a model selection problem. Tuning parameters need to be chosen to optimize the "bias-variance tradeoff."
- More formal treatment of kernel methods will be given in Part II.

Part II: Theory of Reproducing Kernel Hilbert Spaces Methods

- Regularization in RKHS
- Reproducing kernel Hilbert spaces
- Properties of kernels
- Examples of RKHS methods
- Representer Theorem

Regularization in RKHS

Find $f = \sum_{\nu=1}^{M} d_\nu \phi_\nu + h$ with $h \in \mathcal{H}_K$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda \|h\|^2_{\mathcal{H}_K}.$$

- $\mathcal{H}_K$: a reproducing Kernel Hilbert space of functions defined on a domain which can be arbitrary
- $J(f) = \|h\|^2_{\mathcal{H}_K}$: penalty
- The null space spanned by $\{\phi_\nu\}_{\nu=1}^{M}$

Why consider RKHS?

- Theoretical basis for some of popular regularization methods
- Unified framework for function estimation and modeling various data
- Allow general domains for functions
- Permit geometric understanding
- Can do much more than estimation of function values (e.g. integrals and derivatives)

Reproducing Kernel Hilbert Spaces

- Consider a Hilbert space $\mathcal{H}$ of real valued functions on a domain $\mathcal{X}$.
- A Hilbert space $\mathcal{H}$ is a complete inner product linear space.
- For example, the domain $\mathcal{X}$ could be
  - $\{1, \ldots, k\}$
  - $[0, 1]$
  - $\{A, C, G, T\}$
  - $\mathbb{R}^p$
  - $\mathcal{S}$: sphere.
- A reproducing kernel Hilbert space is a Hilbert space of real valued functions, where the evaluation functional $L_x(f) = f(x)$ is bounded in $\mathcal{H}$ for each $x \in \mathcal{X}$.

Riesz Representation Theorem

- For every bounded linear functional L in a Hilbert space $\mathcal{H}$, there exists a unique $g_L \in \mathcal{H}$ such that $L(f) = (g_L, f)$, $\forall f \in \mathcal{H}$.
- g_L is called the *representer* of L.

Reproducing Kernel

Aronszajn (1950), *Theory of Reproducing kernels*.

- By the Riesz representation theorem, there exists $K_x \in \mathcal{H}$, the representer of $L_x(\cdot)$, such that $(K_x, f) = f(x)$, $\forall f \in \mathcal{H}$.
- $K(x,t) = K_x(t)$ is called the reproducing kernel.
  - $K(x,\cdot) \in \mathcal{H}$ for each x
  - $(K(x,\cdot), f(\cdot)) = f(x)$ for all $f \in \mathcal{H}$
- Note that $K(x,t) = K_x(t) = (K_t(\cdot), K_x(\cdot)) = (K_x(\cdot), K_t(\cdot))$ (the reproducing property)

Example of RKHS

- $\mathcal{F} = \{f \mid f : \{1, \dots, k\} \to \mathbb{R}\} = \mathbb{R}^k$ with the Euclidean inner product $(f, g) = f^t g = \sum_{j=1}^{k} f_j g_j$
- Note that $|f(j)| \le \|f\|$. That is, the evaluation functional $L_j(f) = f(j)$ is bounded in $\mathcal{F}$ for each $j \in \mathcal{X}$.
- $L_j(f) = f(j) = (e_j, f)$ where e_j is the jth column of $\mathbf{I}_k$. Hence $K(i, j) = \delta_{ij} = I[i = j]$ or $[K(i, j)] = \mathbf{I}_k$.

The Mercer-Hilbert-Schmidt Theorem

- If $\int_{\mathcal{X}} \int_{\mathcal{X}} K^2(s,t)dsdt < \infty$ for a continuous symmetric non-negative definite K, then there exists an orthonormal sequence of continuous eigenfunctions $\Phi_1, \Phi_2, \ldots$ in $L_2[\mathcal{X}]$ and eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ with $\sum_{i=1}^{\infty} \lambda_i^2 < \infty$ with $\int_{\mathcal{X}} K(s,t)\Phi_i(s)ds = \lambda_i \Phi_i(t)$ and

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i \Phi_i(s) \Phi_i(t).$$

- The inner product in $\mathcal{H}$ of functions f with $\sum_i (f_i^2/\lambda_i) < \infty$

$$(f,g) = \sum_{i=1}^{\infty} \frac{f_i g_i}{\lambda_i}$$

where $f_i = \int_{\mathcal{X}} f(t)\Phi_i(t)dt$.

- Feature mapping: $\Phi(x) = (\sqrt{\lambda_1}\Phi_1(x), \sqrt{\lambda_2}\Phi_2(x), \ldots)$

Reproducing Kernel is Non-Negative Definite

- A bivariate function $K(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is non-negative definite if for every n, and every $x_1, \ldots, x_n \in \mathcal{X}$, and every $a_1, \ldots, a_n \in \mathbb{R}^n$,

$$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \geq 0.$$

In other words, letting $a = (a_1, \ldots, a_n)^t$,

$$a^t \left[K(x_i, x_j) \right] a \geq 0.$$

- For a reproducing kernel K,

$$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) = \left\| \sum_{i=1}^{n} a_i K(x_i, \cdot) \right\|^2 \geq 0.$$

The Moore-Aronszajn Theorem

- For every RKHS $\mathcal{H}$ of functions on $\mathcal{X}$, there corresponds a unique reproducing kernel (RK) $K(s,t)$, which is n.n.d.

- Conversely, for every n.n.d. function $K(s,t)$ on $\mathcal{X}$, there corresponds a unique RKHS $\mathcal{H}$ that has $K(s,t)$ as its RK.

How to construct RKHS given an n.n.d. $K(s,t)$

- For each fixed $x \in \mathcal{X}$, define $K_x(\cdot) = K(x, \cdot)$.
- Taking all finite linear combinations of the form $\sum_i a_i K_{x_i}$ for all choices of n, $a_1, \ldots, a_n$, and $x_1, \ldots, x_n$, construct a linear space $\mathcal{M}$.
- Define an inner product via $(K_{x_i}, K_{x_j}) = K(x_i, x_j)$ and extend it by linearity.

$$(\sum_i a_i K_{x_i}, \sum_j b_j K_{t_j}) = \sum_{i,j} a_i b_j (K_{x_i}, K_{t_j}) = \sum_{i,j} a_i b_j K(x_i, t_j).$$

- For any f of the form $\sum_i a_i K_{x_i}$, $(K_x, f) = f(x)$.
- Complete $\mathcal{M}$ by adjoining all the limits of Cauchy sequences of functions in $\mathcal{M}$.

Sum of Reproducing Kernels

- The sum of two n.n.d. matrices is n.n.d.
- Hence, the sum of two n.n.d. functions defined on the same domain $\mathcal{X}$ is n.n.d.
- In particular, if $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, and $K_j(s,t)$ is the RK of $\mathcal{H}_j$ for $j = 1, 2$, then $K(s,t) = K_1(s,t) + K_2(s,t)$ is the RK for the tensor sum space of $\mathcal{H}_1 \oplus \mathcal{H}_2$.

Product of Reproducing Kernels

- Suppose that $K_1(x_1, x_2)$ is n.n.d. on $\mathcal{X}_1$ and $K_2(t_1, t_2)$ is n.n.d. on $\mathcal{X}_2$.
- Consider the tensor product of K_1 and K_2

$$K\Big((x_1, t_1), (x_2, t_2)\Big) = K_1(x_1, x_2)K_2(t_1, t_2)$$

on $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$.
$K(\cdot, \cdot)$ is n.n.d. on $\mathcal{X}$.

- The tensor product of two RK's K_1 and K_2 for $\mathcal{H}_1$ and $\mathcal{H}_2$ is an RK for the tensor product space of $\mathcal{H}_1 \otimes \mathcal{H}_2$ on $\mathcal{X}$.

Constructing Kernels

- Use reproducing kernels on a univariate domain as building blocks.
- Tensor sums and products of reproducing kernels.
- Systematic approach to estimating multivariate functions.
- Other tricks to expand and design kernels: *Haussler (1999)*, Convolution Kernels on Discrete Structures.
- Learning kernels (n.n.d. matrices) from data: *Lanckriet et al. (2004)*, Learning the Kernel Matrix with Semidefinite Programming, JMLR.

Cubic Smoothing Splines

Find $f(x) \in W_2[0,1]$
$= \{f : \ f, f' \text{ absolutely continuous, and } f'' \in L_2\}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx.$$

- The null space: $M = 2$, $\phi_1(x) = 1$, and $\phi_2(x) = x$.
- The penalized space: $\mathcal{H}_K = W_2^0[0,1] =$
  $\{f \in W_2[0,1] : \ f(0) = 0, \ f'(0) = 0\}$ is an RKHS with
  i)$(f,g) = \int_0^1 f''(x)g''(x)dx$
  ii) $\|f\|^2 = \int_0^1 (f''(x))^2 dx$
  iii) $K(x,x') = \int_0^1 (x-u)_+(x'-u)_+ du$.

SVM in General

Find $f(\boldsymbol{x}) = b + h(\boldsymbol{x})$ with $h \in \mathcal{H}_K$ minimizing

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

- The null space: $M = 1$ and $\phi_1(\boldsymbol{x}) = 1$
- Linear SVM:
  $\mathcal{H}_K = \{h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} \mid \boldsymbol{w} \in \mathbb{R}^p\}$ with
  i) $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}'$
  ii) $\|h\|^2_{\mathcal{H}_K} = \|\boldsymbol{w}^\top \boldsymbol{x}\|^2_{\mathcal{H}_K} = \|\boldsymbol{w}\|^2$
- Nonlinear SVM: $K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^d$ (polynomial kernel), $\exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2 / 2\sigma^2)$ (Gaussian kernel), ...

Representer Theorem

Kimeldorf and Wahba (1971)

- The minimizer $f = \sum_{\nu=1}^{M} d_\nu \phi_\nu + h$ with $h \in \mathcal{H}_K$ of

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda \|h\|^2_{\mathcal{H}_K}$$

has a representation of the form

$$\hat{f}(\boldsymbol{x}) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\boldsymbol{x}) + \sum_{i=1}^{n} c_i K(\boldsymbol{x}_i, \boldsymbol{x}).$$

- $\|h\|^2_{\mathcal{H}_K} = \sum_{i,j} c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Sketch of the Proof

- Write $h \in \mathcal{H}_K$ as

$$h(\boldsymbol{x}) = \sum_{i=1}^{n} c_i K(\boldsymbol{x}_i, \boldsymbol{x}) + \rho(\boldsymbol{x})$$

  where $\rho(\boldsymbol{x})$ is some element in $\mathcal{H}_K$ perpendicular to $K(\boldsymbol{x}_i, \boldsymbol{x})$ for $i = 1, \ldots, n$.
- Note that $h(\boldsymbol{x}_i) = (K(\boldsymbol{x}_i, \cdot), h(\cdot))_{\mathcal{H}_K}$ does not depend on ρ and $\|h\|^2_{\mathcal{H}_K} = \sum_{i=1}^n \sum_{j=1}^n c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \|\rho\|^2_{\mathcal{H}_K}$
- Then, $\|\rho\|^2_{\mathcal{H}_K}$ needs to be zero.
- Hence, the minimizer $\hat{f}$ is of the form

$$\hat{f}(\boldsymbol{x}) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\boldsymbol{x}) + \sum_{i=1}^{n} c_i K(\boldsymbol{x}_i, \boldsymbol{x}).$$

Remarks on the Representer Theorem

- It holds for an arbitrary loss function $\mathcal{L}$.
- Minimizer of RKHS method resides in a finite dimensional space.
- So the solution is computable even if the RKHS had infinite dimension.
- The resulting optimization problems with the representation $\hat{f}$ depend on $\mathcal{L}$ and the penalty $J(f)$.

Loss Function and its Risk Minimizer f

$\mathcal{L}(y, f(x))$	$\text{argmin } E[\mathcal{L}(Y, f(X))\|X = x]$
Regression	
$(y - f(x))^2$	$E(Y\|X = x)$
$\|y - f(x)\|$	$median(Y\|X = x)$
Classification with $y = \pm 1$	
SVM: $(1 - yf(x))_+$	$sign\{p(x) - 1/2\}$
Logistic regression:	
$\log\{1 + \exp(-yf(x))\}$	$\log p(x)/(1 - p(x))$
Boosting: $\exp(-yf(x))$	$(1/2) \log p(x)/(1 - p(x))$
	where $p(x) = P(Y = 1\|X = x)$

Hinge vs -log likelihood

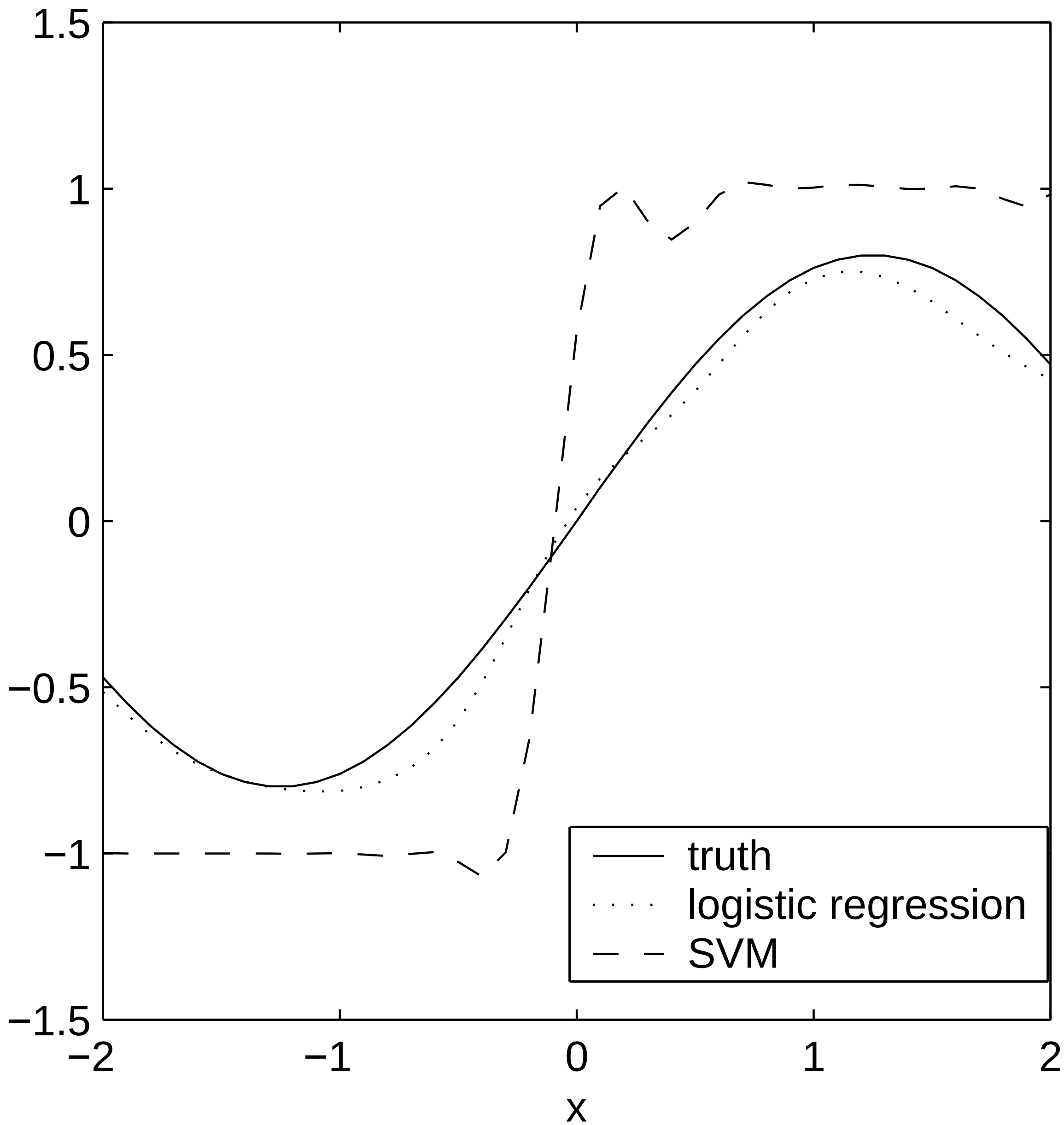


Figure: Solid: $2p(x) - 1$, dotted: $2\hat{p}_{LG}(x) - 1$ and dashed: $\hat{f}_{SVM}(x)$

Smoothing Splines for Modeling Non-Gaussian Response

- Generalized linear models for the data with y from exponential families can be extended nonparametrically.
- The main idea of the extension of smoothing splines to non-Gaussian case is to replace the squared error loss by the negative log likelihood function associated with y.
- Thus, the penalized least squares method becomes the method of penalized negative log likelihood.
- Non-Gaussian data can be treated in the framework of RKHS method in a unified way, yet they require some computational modification.

Logistic Regression

- Suppose that we have independent data points of (x_i, y_i), $i = 1, \ldots, n$, and $Y_i | x_i \sim Bernoulli(p(x_i))$, where $p(x) = P(Y = 1 | X = x)$. Note that $y_i \in \{0, 1\}$.
- The likelihood of the data conditional on x is

$$\prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}.$$

- Model the logit function of $p(x)$, $\log\{p(x)/(1 - p(x))\}$ by $f \in \mathcal{H}$.
- In terms of the logit $f(x)$, the likelihood is given by

$$\prod_{i=1}^{n} \exp(y_i f(x_i) - \log(1 + e^{f(x_i)})) = \exp(\sum_{i=1}^{n} y_i f(x_i) - \log(1 + e^{f(x_i)})).$$

Formulation of Penalized Logistic Regression

- Use the negative log likelihood as a loss function.
- Find $\hat{f}_\lambda \in \mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\{-y_i f(x_i) + \log(1 + e^{f(x_i)})\} + \lambda \|P_1 f\|^2.$$

- By the representer theorem,

$$\hat{f}_\lambda(x) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} c_i K(x_i, x).$$

- The optimization problem of finding $d = (d_1, \ldots, d_M)^t$ and $c = (c_1, \ldots, c_n)^t$ entails iterations.
- The solution can be approximated by iteratively reweighted least squares.

Machine Learning View on Kernels

- $K(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x})^\top \Phi(\boldsymbol{x}')$ through feature mapping Φ from $\mathbb{R}^p$ to a feature space.
- Kernel trick:
  Kernelize, that is, replace the Euclidean inner product $x^\top x'$ in your linear method with $K(x, x')$!
- This idea goes beyond supervised learning problems.
  - nonlinear dimension reduction and data visualization: kernel PCA
  - clustering: kernel k-means algorithm
  - ...

Summary

- RKHS methods provide a unified framework for statistical model building.
- Kernels are now used as a versatile tool for flexible modeling and learning in various contexts.
- Feature selection for kernel methods will be discussed in Part III based on the idea of kernel construction by sums and products.

Part III: Regularization Approach to Feature Selection

- Feature selection procedures
- LASSO, modeling with l_1 constraint
- Generalization of LASSO for kernel methods
- Application for finding biomarkers

Motivation for Feature Selection

- Key questions in many scientific investigations.
- Achieve parsimony (Occam's razor)
  "Entities should not be multiplied beyond necessity."
- Enhance interpretation.
- Often reduce variance, hence improve prediction accuracy.

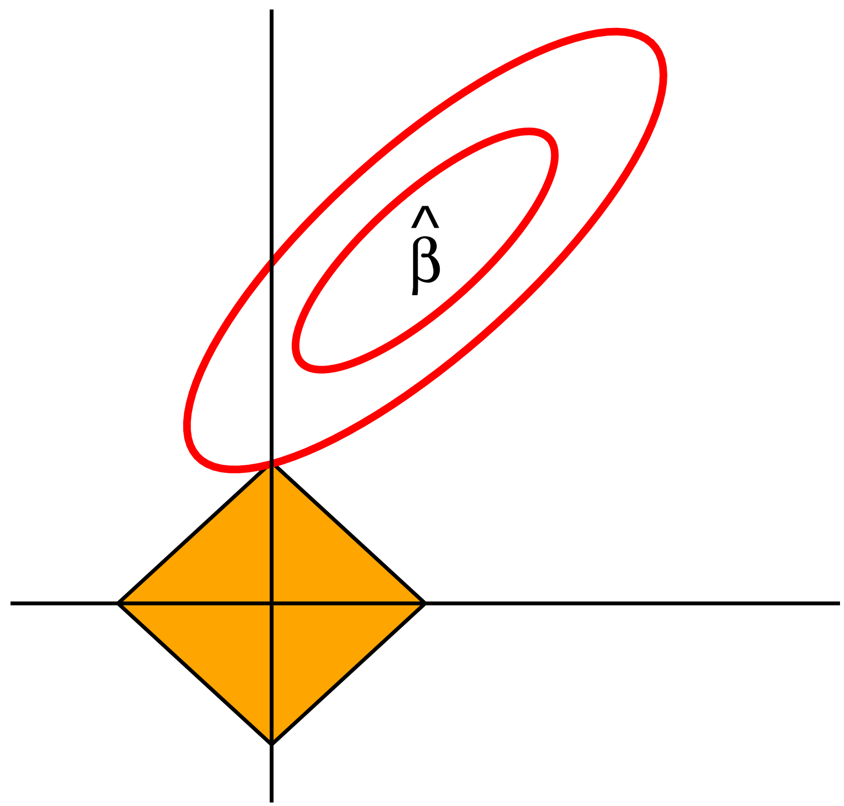
Feature Selection Procedures

- Combinatorial approach:
  Best subset selection, Forward selection, Backward elimination, Stepwise regression
  e.g. *Guyon et al. (2002)*, Recursive feature selection

- l_1 penalty for simultaneous fitting and selection:
  e.g. *Bradley and Mangasarian (1998)*,
  Linear SVM with l_1 penalty
  Tibshirani (1996), LASSO
  (Least Absolute Shrinkage and Selection Operator)
  Chen and Donoho (1994), Basis Pursuit
- Other variants for groupings of variables, model hierarchy, adaptiveness, and efficiency.

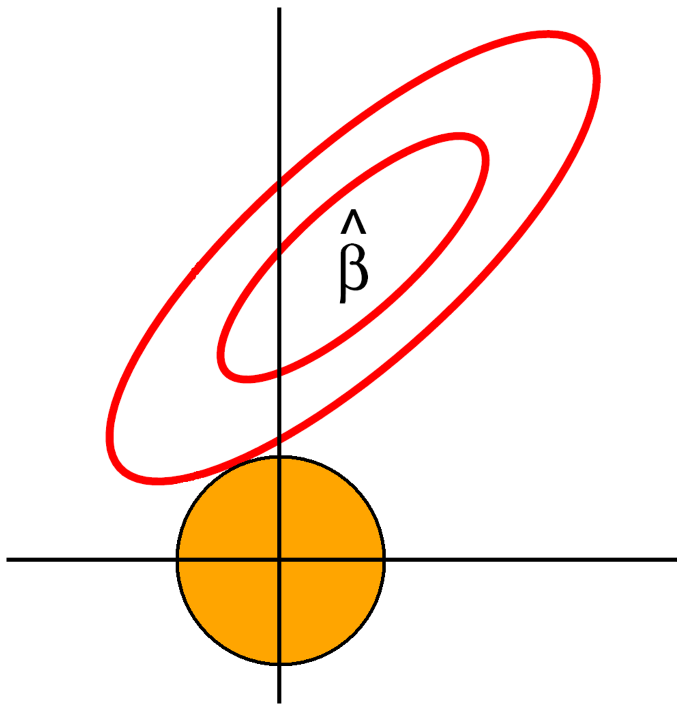
LASSO

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \|\beta\|_1 \Leftrightarrow \min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \text{ s.t. } \|\beta\|_1 \leq s.$$

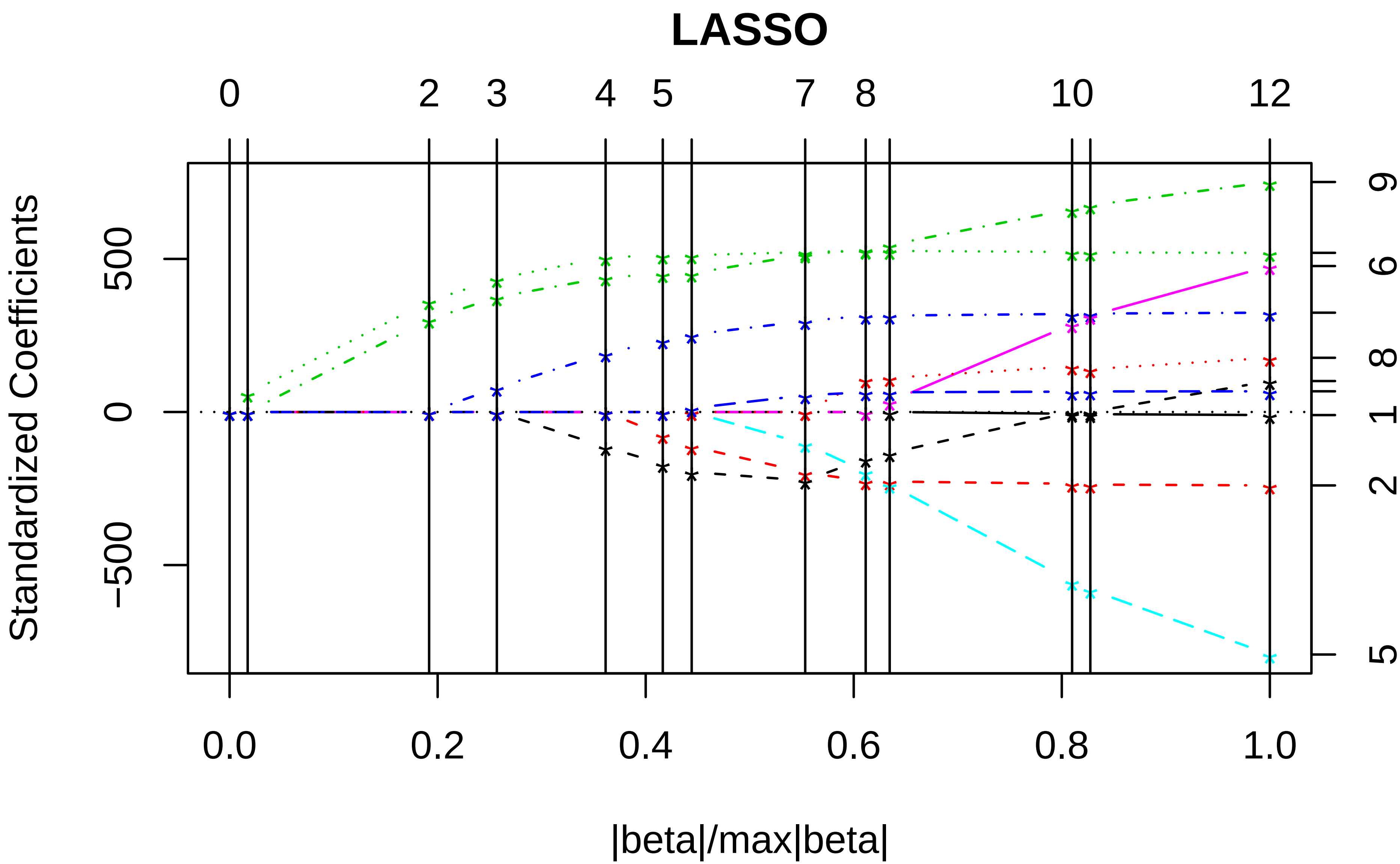


Ridge Regression

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \|\beta\|_2^2.$$



LASSO Coefficient Paths



Median Regression with l_1 Penalty for Income Data

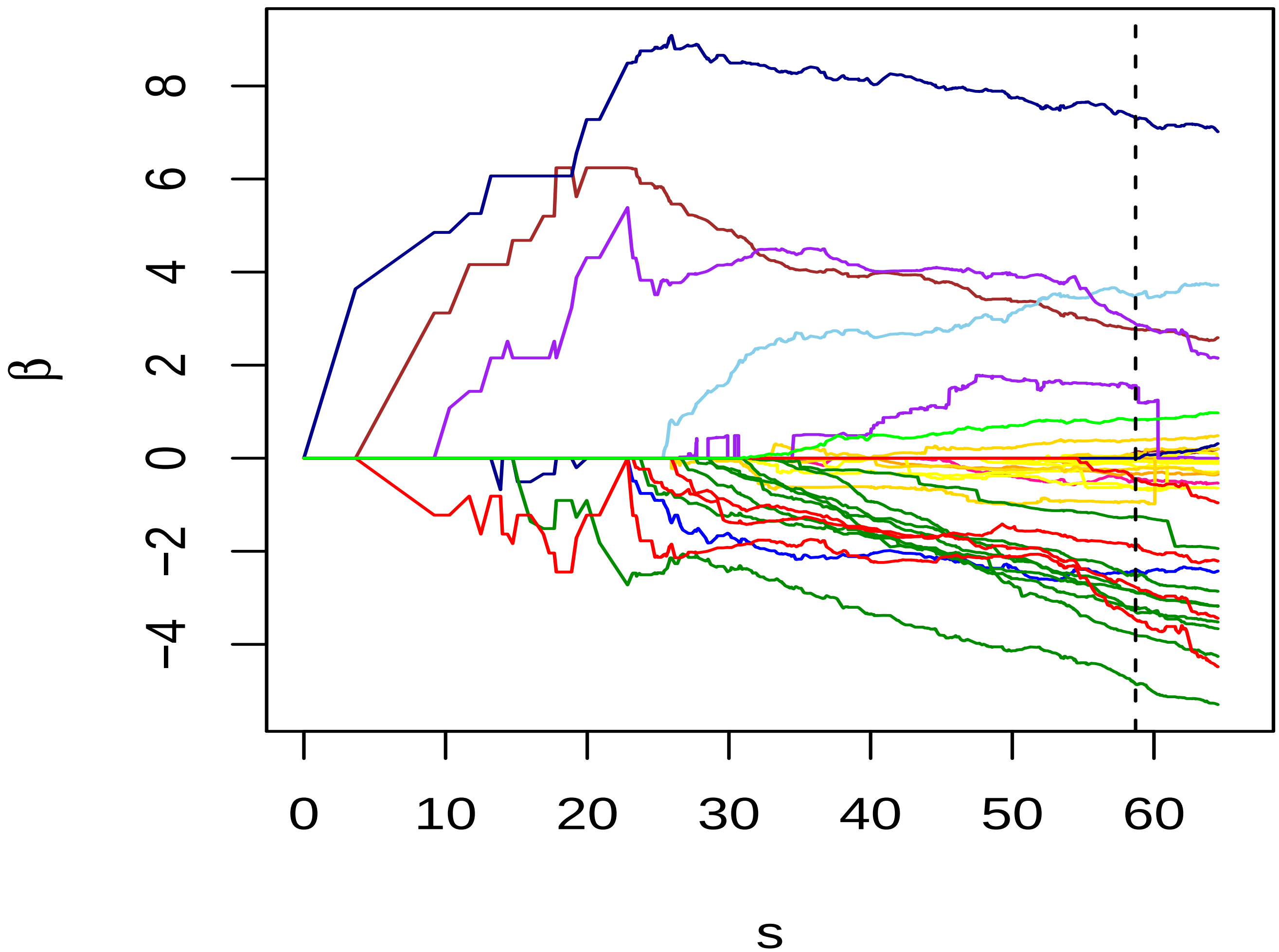


Figure: The coefficient paths of the main effects model. Positive: home ownership (in dark blue relative to renting), education (in brown), dual income due to marriage (in purple relative to ‘not married’), age (in skyblue), and male (in light green). Negative: single or divorced (in red relative to ‘married’) and student, clerical worker, retired or unemployed (in green relative to professionals/managers)

Median Regression with l_1 Penalty for Income Data

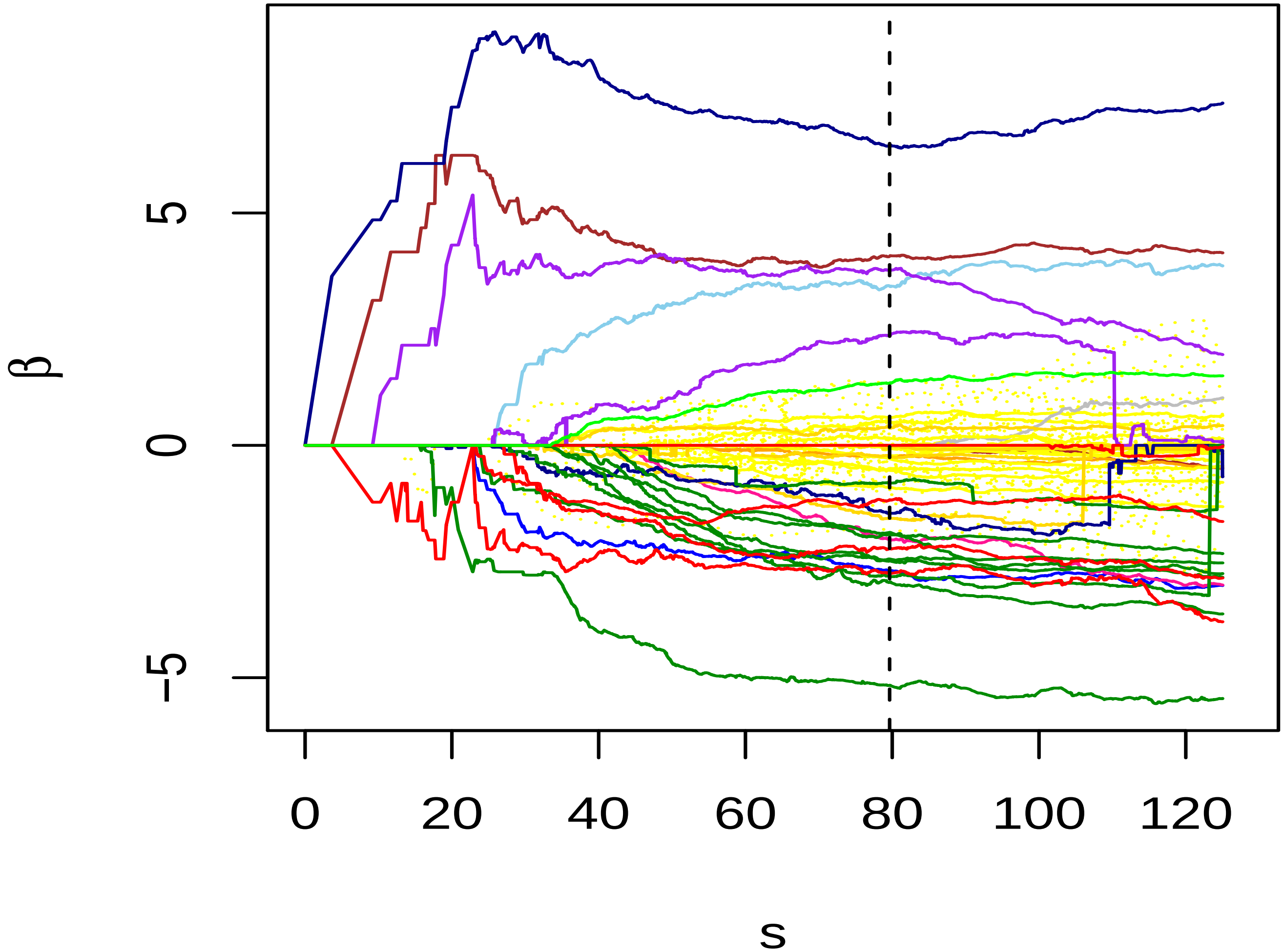


Figure: The coefficient paths of a partial two-way interaction model. Positive: ‘dual income $*$ home ownership’, ‘home ownership $*$ education’, and ‘married but no dual income $*$ education’. Negative: ‘single $*$ education’ and ‘home ownership $*$ age’

Risk Path

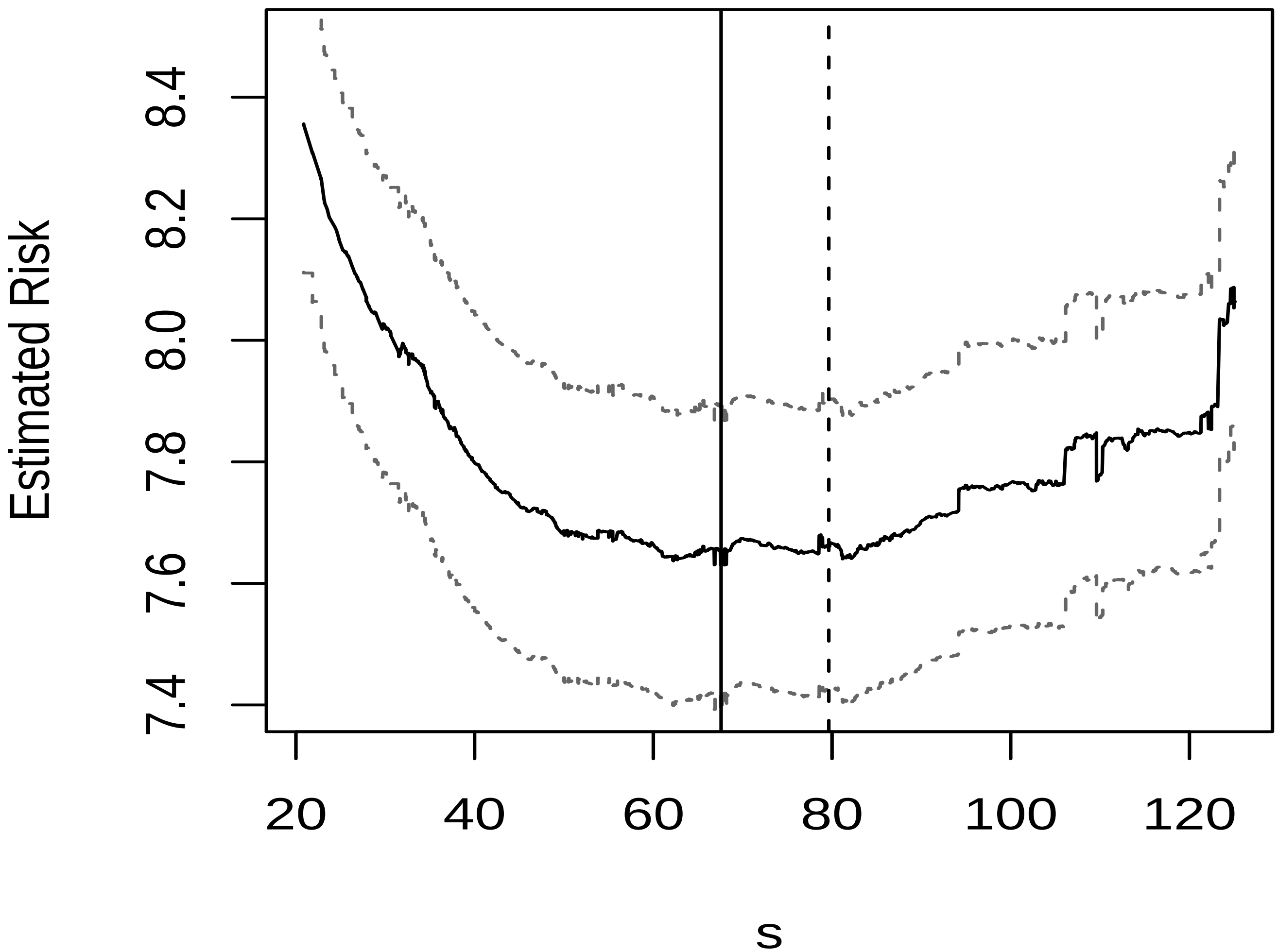


Figure: The risks of the two-way fitted models are estimated by using a test data set with 4,876 observations.

Generalization of LASSO

- Kernel methods may be difficult to interpret when the embedding into feature space is implicit.
- Regression:
  Lin and Zhang (2003), COmponent Selection and Smoothing Operator
  Gunn and Kandola (2002), Structural modelling with sparse kernel
- Classification:
  Zhang (2006) for the binary SVM
  Lee et al. (2006) for the multiclass SVM

Strategy for Feature Selection

- Structured representation of f using functional ANOVA decomposition
- A sparse solution approach with l_1 penalty
- A unified treatment for regression and classification (both linear and nonlinear cases)
- Inexpensive additional computation
- Systematic elaboration of f with features

Functional ANOVA Decomposition

- For f defined on a product domain $\mathcal{X} = \prod_{j=1}^{p} \mathcal{X}_j$,

$$
\begin{aligned}
f &= \prod_j [A_j + (I - A_j)] f \\
&= (\prod_j A_j) f + \sum_i (\prod_{j \neq i} A_j)(I - A_i) f \\
&\quad + \sum_{i<j} (\prod_{r \neq i,j} A_r)(I - A_i)(I - A_j) f + \cdots
\end{aligned}
$$

- Functional "overall mean" + "main effects" + "two-way interactions" + $\cdots$.
- Define A_j appropriately so that the decomposition of A_j and $I - A_j$ corresponds to $\{1\} \oplus \bar{\mathcal{H}}_j$.

ANOVA Spaces and Kernels

Wahba (1990), smoothing spline ANOVA models

- Function: $f(\boldsymbol{x}) = b + \sum_{\alpha=1}^{p} f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x_\alpha, x_\beta) + \cdots$

- Functional space: $f \in \mathcal{H} = \otimes_{\alpha=1}^{p}(\{1\} \oplus \bar{\mathcal{H}}_\alpha)$,
  $\mathcal{H} = \{1\} \oplus \sum_{\alpha=1}^{p} \bar{\mathcal{H}}_\alpha \oplus \sum_{\alpha<\beta}(\bar{\mathcal{H}}_\alpha \otimes \bar{\mathcal{H}}_\beta) \oplus \cdots$

- Reproducing kernel (r.k.):
  $K(\boldsymbol{x}, \boldsymbol{x}') = 1 + \sum_{\alpha=1}^{p} K_\alpha(\boldsymbol{x}, \boldsymbol{x}') + \sum_{\alpha<\beta} K_{\alpha\beta}(\boldsymbol{x}, \boldsymbol{x}') + \cdots$

Two-way ANOVA Decomposition of $f \in W_2[0,1] \otimes W_2[0,1]$

- Two-way ANOVA decomposition of f

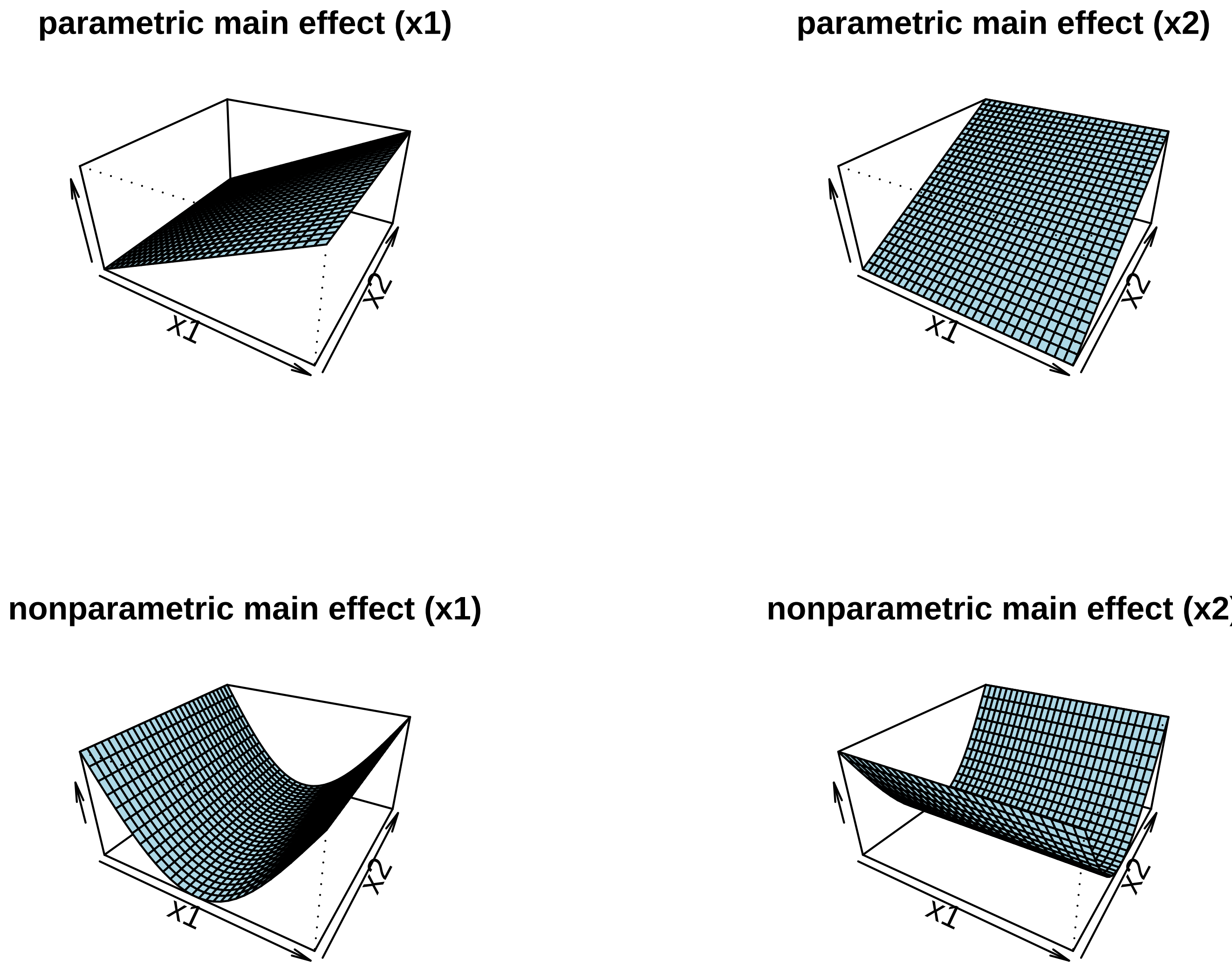
$$f(x_1, x_2) = f_0 + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2)$$

with the side conditions that $\int_0^1 f_j(x_j)dx_j = 0$ and $\int_0^1 f_{12}(x_1, x_2)dx_j = 0$ for $j = 1, 2$.

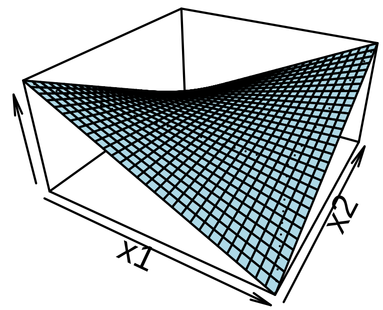
- The corresponding functional components

	$\{1\}$	$\{k_1(x_2)\}$	$\mathcal{H}_1^2$
$\{1\}$	mean	p-main effect (x_2)	n-main effect (x_2)
$\{k_1(x_1)\}$	p-main effect (x_1)	p$\times$p-interaction	p$\times$n-interaction
$\mathcal{H}_1^1$	n-main effect (x_1)	n$\times$p-interaction	n$\times$n-interaction

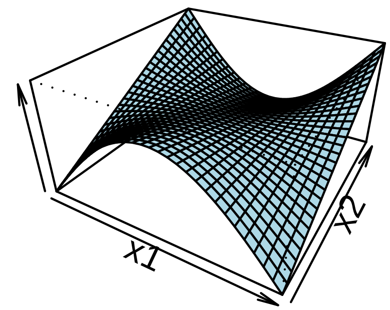
where “p” and “n” mean parametric and nonparametric, respectively.


parametric main effect (x1)
parametric main effect (x2)
nonparametric main effect (x1)
nonparametric main effect (x2)
x1
x2

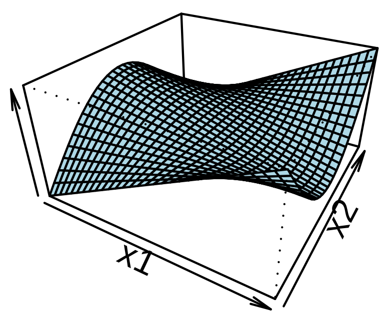
p−p interaction



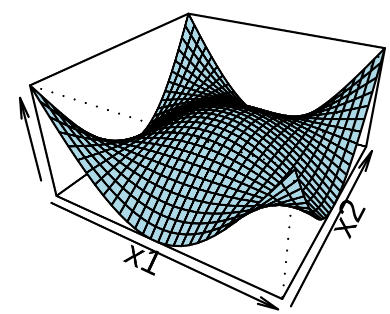
n−p interaction



p−n interaction



n−n interaction



l_1 Penalty on $\boldsymbol{\theta}$

- Modification of r.k. by rescaling parameters $\boldsymbol{\theta} \geq 0$
  $K_\theta(\boldsymbol{x}, \boldsymbol{x}') = 1 + \sum_{\alpha=1}^{p} \theta_\alpha K_\alpha(\boldsymbol{x}, \boldsymbol{x}') + \sum_{\alpha<\beta} \theta_{\alpha\beta} K_{\alpha\beta}(\boldsymbol{x}, \boldsymbol{x}') + \cdots$
- Truncating $\mathcal{H}$ to $\mathcal{F} = \{1\} \oplus_{\nu=1}^{d} \mathcal{F}_\nu$, find $f(\boldsymbol{x}) \in \mathcal{F}$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda \sum_{\nu} \theta_\nu^{-1} \|P^\nu f\|^2.$$

  Then $\hat{f}(\boldsymbol{x}) = \hat{b} + \sum_{i=1}^{n} \hat{c}_i \left[\sum_{\nu=1}^{d} \theta_\nu K_\nu(\boldsymbol{x}_i, \boldsymbol{x}) \right].$
- For sparsity, minimize

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda \sum_{\nu=1}^{d} \theta_\nu^{-1} \|P^\nu f\|^2 + \lambda_\theta \sum_{\nu=1}^{d} \theta_\nu$$

  subject to $\theta_\nu \geq 0, \forall \nu$.

Related to Kernel Learning

- *Micchelli and Pontil (2005)*, Learning the kernel function via regularization
- $\mathcal{K} = \{K_\nu, \nu \in \mathcal{N}\}$: a compact and convex set of kernels
- A variational problem for optimal kernel configuration

$$\min_{K \in \mathcal{K}} \left(\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda J(f) \right)$$

One-Step Update for Structured Regression

- Given $\hat{b}$ and $\{\hat{c}_j\}$, recalibrate $\boldsymbol{\theta}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{b} - \sum_{\nu=1}^{d}\theta_\nu\left[\sum_{j=1}^{n}\hat{c}_j K_\nu(\boldsymbol{x}_j, \boldsymbol{x}_i)\right]\right)^2$$

$$+ \quad \lambda\sum_{\nu}\theta_\nu \sum_{i,j=1}^{n}\hat{c}_i\hat{c}_j K_\nu(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

subject to $\theta_\nu \geq 0, \forall\nu,$ and $\sum_{\nu}\theta_\nu \leq s$

Nonnegative Garrote

Breiman, L. (1995),
Better Subset Regression Using the Nonnegative Garrote

- Stepwise variable selection can be unstable with respect to small perturbations in the data.
- Starting with the full LSE, it both shrinks and zeroes coefficients.
- Given $\hat{\beta}^{LS}$, take $(c_1, \ldots, c_p)$ to minimize

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} c_j \hat{\beta}_j^{LS} x_{ij})^2$$

  subject to $c_j \geq 0$ and $\sum_{j=1}^{p} c_j \leq s$.
- Generally lower prediction error than best subset selection

Structured SVM with ANOVA decomposition

- The binary case (Zhang, 2006):
  Find $f(\boldsymbol{x}) = b + h(\boldsymbol{x})$ minimizing
  $$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(\boldsymbol{x}_i))_+ + \lambda \sum_{\nu=1}^{d} \theta_\nu^{-1} \|P^\nu f\|^2 + \lambda_\theta \sum_{\nu=1}^{d} \theta_\nu$$
  subject to $\theta_\nu \geq 0, \forall \nu$.
- The multiclass case (Lee et al., 2006) :
  Find $\boldsymbol{f} = (f^1, \ldots, f^k) = (b^1 + h^1(\boldsymbol{x}), \ldots, b^k + h^k(\boldsymbol{x}))$ with the sum-to-zero constraint minimizing
  $$\frac{1}{n}\sum_{i=1}^{n}\sum_{j \neq y_i}(f^j(\boldsymbol{x}_i) - y_i^j)_+ + \frac{\lambda}{2}\sum_{j=1}^{k}\left(\sum_{\nu=1}^{d}\theta_\nu^{-1}\|P^\nu h^j\|^2\right)$$
  $$+\lambda_\theta \sum_{\nu=1}^{d}\theta_\nu \text{ subject to } \theta_\nu \geq 0, \forall \nu,$$
  where $(y^1, \ldots, y^k)$ is a class code with $y^j = 1$ and $-1/(k-1)$ elsewhere, if $y = j$, and $\phi(\boldsymbol{x}) = \arg\max_j [f^j(\boldsymbol{x})]$.

Updating Algorithm

Letting $\boldsymbol{C} = (b, \{c_j\})$ and denoting the objective function by $\Phi(\boldsymbol{\theta}, \boldsymbol{C})$,

- Initialize $\boldsymbol{\theta}^{(0)} = (1, \ldots, 1)^t$ and $\boldsymbol{C}^{(0)} = \operatorname{argmin} \Phi(\boldsymbol{\theta}^{(0)}, \boldsymbol{C})$.
- At the m-th iteration ($m = 1, 2, \ldots$)

  (θ-step) find $\boldsymbol{\theta}^{(m)}$ minimizing $\Phi(\boldsymbol{\theta}, \boldsymbol{C}^{(m-1)})$ with $\boldsymbol{C}$ fixed.

  (c-step) find $\boldsymbol{C}^{(m)}$ minimizing $\Phi(\boldsymbol{\theta}^{(m)}, \boldsymbol{C})$ with $\boldsymbol{\theta}$ fixed.

- One-step update can be used in practice.

Two-Way Regularization

- c-step solutions range from the simplest model (or majority rule) to the complete overfit to the data as λ decreases. (Standard regularization procedure)
- θ-step solutions range from the constant model to the full model with all the variables as λ_θ decreases. (Functional component pursuit)

Breast Cancer Data

- Sharma et al. (2005), Early detection of breast cancer based on gene-expression patterns in peripheral blood cells, *Breast Cancer Research*.
- Develop accurate and convenient methods for detection of breast cancer using blood samples.
- 60 unique blood samples from 24 women with breast cancer and 32 women with no signs of the disease
- Mean normalized and cube-root transformed expression levels of 1,368 cDNAs
- The nearest shrunken centroid method (Tibshirani et al. 2002) was used in the original paper.

Searching for Gene Signatures

- The Nearest Shrunken Centroid method
- The ℓ_1-norm SVM with
  - 1,368 linear terms
  - 1,368 linear and quadratic terms
- The structured kernel SVM with 1,368 nonparametric main effects terms
- 'External' 6-fold cross-validation
  (Ambroise and McLachlan, *PNAS* 2002)
  - Split 60 samples into a training set of 50 and a test set of 10.
  - Internal 5-fold CV for selection of an optimal tuning parameter (a threshold for NSC, and penalty parameters for SVM).

Computation

- Both ℓ_1-norm SVM and the θ-step of the structured kernel SVM require parametric linear programming.
- Solutions are piecewise constant in the tuning parameter (λ or λ_θ).
- Simplex algorithm can be used to generate the entire regularized solution path *(Yao and Lee, 2007)*.

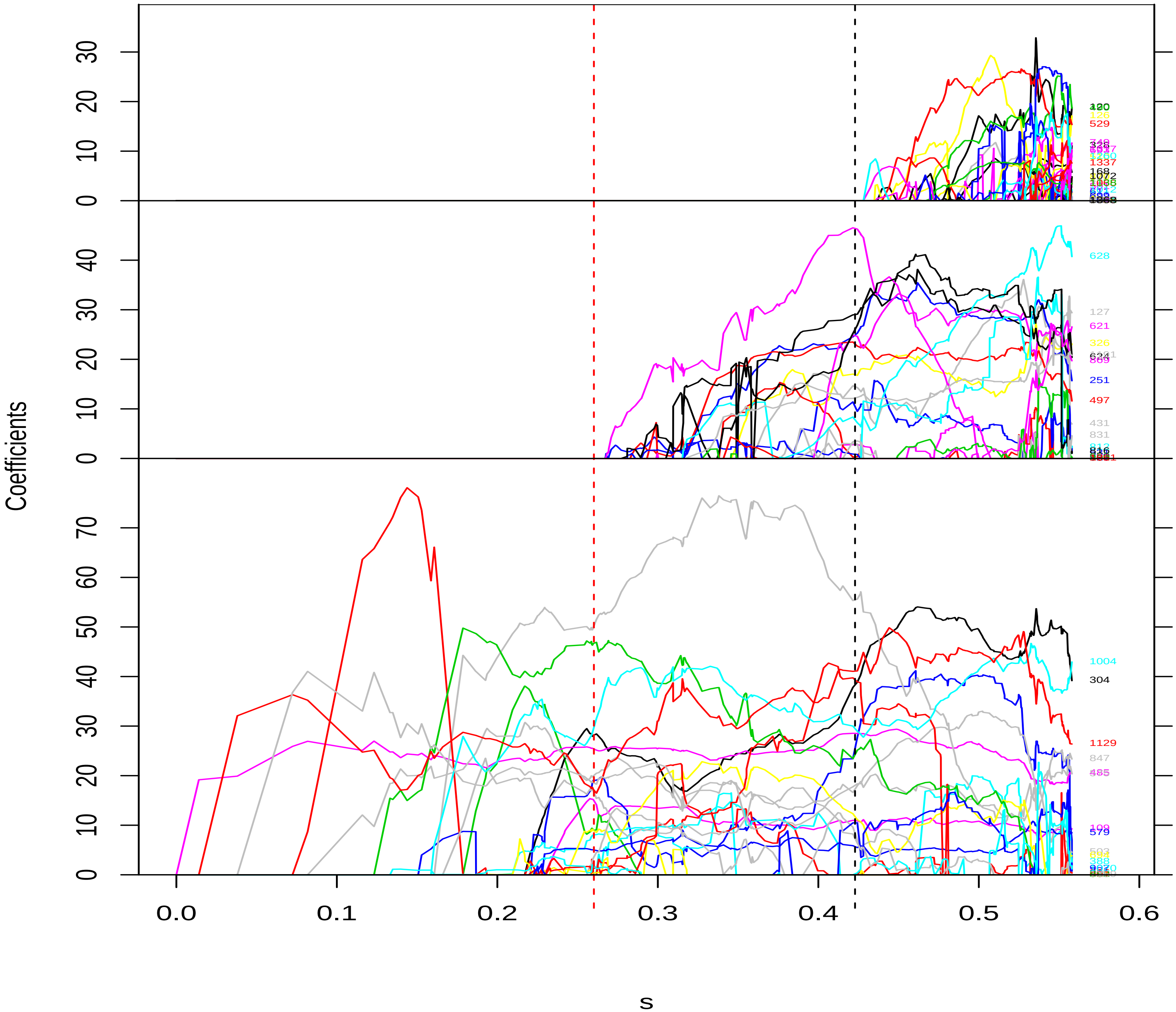


Figure: Recalibration parameter path for structured SVM.

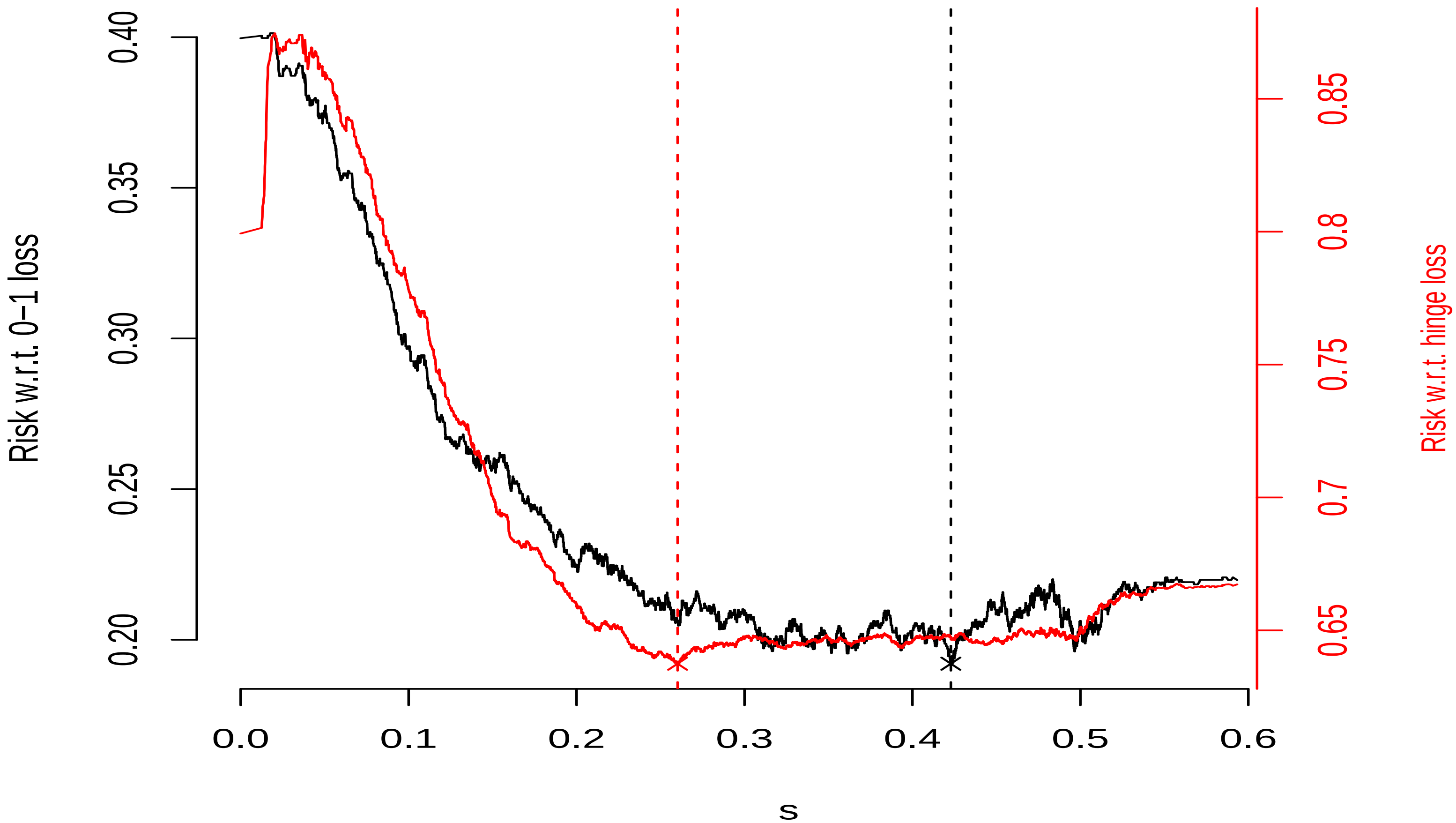


Figure: Error rate path for the θ-step of structured SVM.

Error Rates

- External 6-fold CV
- Comparison

Method	NSC	L.SVM	LQ.SVM	StructSVM
Mean	0.186	0.197	0.279	0.170
SE	0.050	0.051	0.058	0.048

Summary

- Integrate feature selection with kernel methods using l_1 type penalty.
- Enhance interpretation without compromising prediction accuracy.
- General approach for structured and sparse representation with kernels.
- RKHS methods can solve a wide range of statistical learning problems in a principled way.

Software

- Smoothing spline ANOVA models:
  `ssanova` for Gaussian response and `gssanova` for non-Gaussian response in `gss` R library
- Support vector machines:
  `kernlab` for SVM, spectral clustering and kernel PCA
  `http://www.kernel-machines.org` for other implementations in Matlab and Fortran
  `svmpath` for binary SVM solution path
  `msvmpath` for multicategory SVM solution path (in progress)
- Other path-finding algorithms:
  `lars` for LASSO, LAR, stagewise fitting
  `glmpath` for l_1 regularized generalized linear models
  `lpRegPath` for parametric linear programming with linear loss and l_1 penalty (in progress)

Acknowledgments

- Grace Wahba and Yi Lin (Statistics, University of Wisconsin)
- Cheol-Koo Lee (Genomics, Korea University)
- Zhenhuan Cui and Yonggang Yao (former students now at SAS)

- For preprints or reprints of my work, visit http://www.stat.osu.edu/∼yklee

- E-mail: yklee@stat.osu.edu

References

Peter J. Bickel and Bo Li.
Regularization in statistics.
Test, 15(2):271–344, Dec 2006.

Leo Breiman.
Statistical modeling: The two cultures.
Statistical Science, 16(3):199–215, aug 2001.

N. Cristianini and J. Shawe-Taylor.
An Introduction to Support Vector Machines.
Cambridge University Press, 2000.

T. Hastie, R. Tibshirani, and J. Friedman.
The Elements of Statistical Learning.
Springer Verlag, New York, 2001.

B. Schölkopf and A. Smola.
Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond.
MIT Press, 2002.

V. Vapnik.
The Nature of Statistical Learning Theory.
Springer Verlag, New York, 1995.

G. Wahba.
Spline Models for Observational Data.
Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.