

## Overview

# Structured learning for feature selection and prediction

Yoonkyung Lee  
Department of Statistics  
The Ohio State University

January 22-26, 2007  
Winter School at CIMAT

- ▶ Part I:  
Introduction to Kernel methods
- ▶ Part II:  
Learning with Reproducing Kernel Hilbert Spaces
- ▶ Part III:  
Structured learning for feature selection and prediction



## Outline

- ▶ Motivation
- ▶ Feature selection procedures
- ▶ Generalization of LASSO for kernel methods
- ▶ Structured MSVM with ANOVA decomposition
- ▶ Application
- ▶ Concluding remarks



## 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.



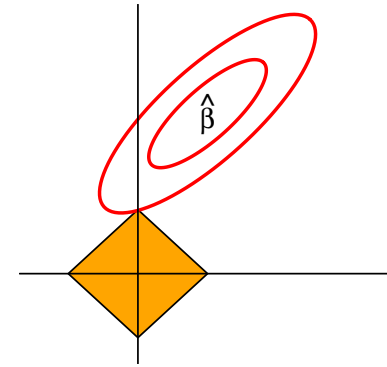
## Feature selection procedures

- ▶ **Combinatorial** approach:  
Best subset selection, Forward selection, Backward elimination, Stepwise regression  
e.g. *Guyon et al. (2002)*, Recursive feature selection
- ▶  **$\ell_1$  penalty** for simultaneous fitting and selection:  
e.g. *Bradley and Mangasarian (1998)*,  
Linear SVM with  $\ell_1$  penalty  
*Tibshirani (1996)*, LASSO  
(Least Absolute Shrinkage and Selection Operator)

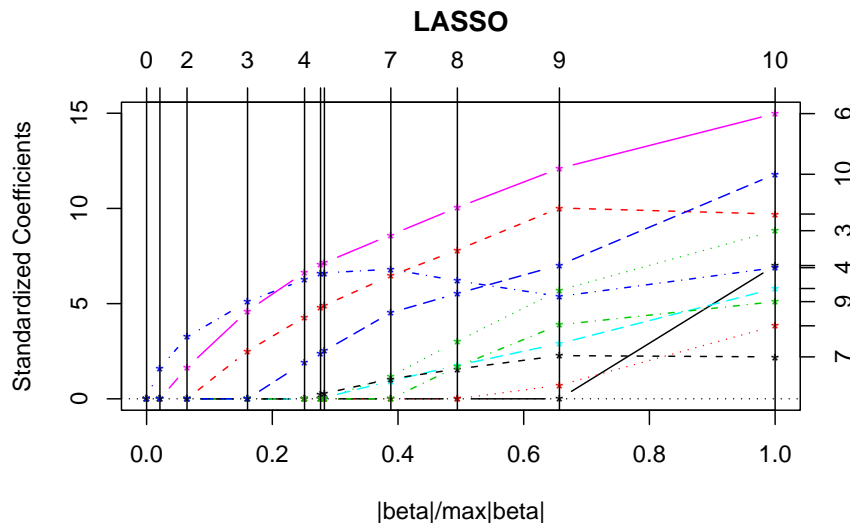


## LASSO

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



## LASSO coefficient paths



## 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  $\ell_1$  penalty
- ▶ A unified treatment for regression and classification (both linear and nonlinear cases)
- ▶ Inexpensive additional computation
- ▶ Systematic elaboration of  $f$  with features

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

## ANOVA spaces and kernels

Wahba (1990), smoothing spline ANOVA models

- ▶ Function:  $f(\mathbf{x}) = b + \sum_{\alpha=1}^p f_{\alpha}(\mathbf{x}_{\alpha}) + \sum_{\alpha < \beta} f_{\alpha\beta}(\mathbf{x}_{\alpha}, \mathbf{x}_{\beta}) + \dots$
- ▶ Functional space:  $f \in \mathcal{H} = \otimes_{\alpha=1}^p (\{1\} \oplus \bar{\mathcal{H}}_{\alpha})$ ,  
 $\mathcal{H} = \{1\} \oplus \sum_{\alpha=1}^p \bar{\mathcal{H}}_{\alpha} \oplus \sum_{\alpha < \beta} (\bar{\mathcal{H}}_{\alpha} \otimes \bar{\mathcal{H}}_{\beta}) \oplus \dots$
- ▶ Reproducing kernel (r.k.):  
 $K(\mathbf{x}, \mathbf{x}') = 1 + \sum_{\alpha=1}^p K_{\alpha}(\mathbf{x}, \mathbf{x}') + \sum_{\alpha < \beta} K_{\alpha\beta}(\mathbf{x}, \mathbf{x}') + \dots$
- ▶ Modification of r.k. by rescaling parameters  $\theta \geq 0$   
 $K_{\theta}(\mathbf{x}, \mathbf{x}') = 1 + \sum_{\alpha=1}^p \theta_{\alpha} K_{\alpha}(\mathbf{x}, \mathbf{x}') + \sum_{\alpha < \beta} \theta_{\alpha\beta} K_{\alpha\beta}(\mathbf{x}, \mathbf{x}') + \dots$

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

## Functional ANOVA decomposition

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

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

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

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

## $\ell_1$ penalty on $\theta$

- ▶ Truncating  $\mathcal{H}$  to  $\mathcal{F} = \{1\} \oplus_{\nu=1}^d \mathcal{F}_{\nu}$ , find  $f(\mathbf{x}) \in \mathcal{F}$  minimizing

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

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

- ▶ For sparsity, minimize

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

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

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

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

## Nonnegative Garrote

*Breiman, L. (1995)*,

Better Subset Regression Using the Nonnegative Garrote

- ▶ Starting with the full LSE, it both shrinks and zeroes coefficients.
- ▶ Given  $\hat{\beta}^{LS}$ , take  $(c_1, \dots, c_p)$  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

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

## One-step update for structured regression

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

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left( y_i - \hat{b} - \sum_{\nu=1}^d \theta_\nu \left[ \sum_{j=1}^n \hat{c}_j K_\nu(\mathbf{x}_j, \mathbf{x}_i) \right] \right)^2 \\ & + \lambda \sum_{\nu} \theta_\nu \sum_{i,j=1}^n \hat{c}_i \hat{c}_j K_\nu(\mathbf{x}_i, \mathbf{x}_j) \\ & \text{subject to } \theta_\nu \geq 0, \forall \nu, \text{ and } \sum_{\nu} \theta_\nu \leq s \end{aligned}$$

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

## SVM when $k > 2$

*Lee, Lin & Wahba, JASA (2004)*

- ▶  $\mathbf{y} = (y^1, \dots, y^k)$ : class code with  $y^j = 1$  and  $-1/(k-1)$  elsewhere, if  $y = j$ .
- ▶ Find  $\mathbf{f} = (f^1, \dots, f^k) = (b^1 + h^1(\mathbf{x}), \dots, b^k + h^k(\mathbf{x}))$  with  $h^j \in \mathcal{H}_K$  and the sum-to-zero constraint minimizing

$$\frac{1}{n} \sum_{i=1}^n \sum_{j \neq y_i} (f^j(\mathbf{x}_i) - y_i^j)_+ + \frac{\lambda}{2} \sum_{j=1}^k \|h^j\|^2.$$

- ▶ Classification rule:  $\phi(\mathbf{x}) = \arg \max_j [f^j(\mathbf{x})]$

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍 ↻

## Structured MSVM with ANOVA decomposition

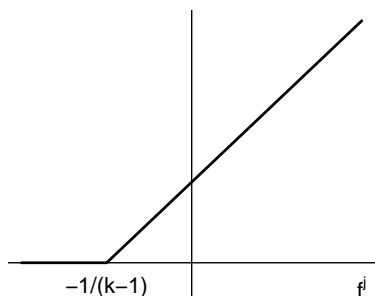


Figure: MSVM component loss  $(f^j - y^j)_+$  where  $y^j = -1/(k-1)$ .

Lee et al., *Biometrika* (2006)

- ▶ Find  $\mathbf{f} = (f^1, \dots, f^k) = (b^1 + h^1(\mathbf{x}), \dots, b^k + h^k(\mathbf{x}))$  with the sum-to-zero constraint minimizing

$$\frac{1}{n} \sum_{i=1}^n \mathbf{L}(\mathbf{y}_i) \cdot (\mathbf{f}(\mathbf{x}_i) - \mathbf{y}_i)_+ + \frac{\lambda}{2} \sum_{j=1}^k \left( \sum_{\nu=1}^d \theta_{\nu}^{-1} \|P^{\nu} h^j\|^2 \right) + \lambda_{\theta} \sum_{\nu=1}^d \theta_{\nu} \text{ subject to } \theta_{\nu} \geq 0, \text{ for } \nu = 1, \dots, d.$$

- ▶  $\mathbf{L}(\mathbf{y})$ : misclassification cost
- ▶ By the representer theorem,  $\hat{f}^j(\mathbf{x}) = \hat{b}^j + \sum_{i=1}^n \hat{c}_i^j \left[ \sum_{\nu=1}^d \theta_{\nu} K_{\nu}(\mathbf{x}_i, \mathbf{x}) \right]$ .

## Updating Algorithm

Letting  $\mathbf{C} = (\{b^j\}, \{c_i^j\})$  and denoting the objective function by  $\Phi(\theta, \mathbf{C})$ ,

- ▶ Initialize  $\theta^{(0)} = (1, \dots, 1)^t$  and  $\mathbf{C}^{(0)} = \text{argmin } \Phi(\theta^{(0)}, \mathbf{C})$ .
- ▶ At the  $m$ -th iteration ( $m = 1, 2, \dots$ )

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

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

- ▶ One-step update can be used in practice.

## Two-way regularization

- ▶  $c$ -step solutions range from the simplest majority rule to the complete overfit to data as  $\lambda$  decreases.
- ▶  $\theta$ -step solutions range from the constant model to the full model with all the variables as  $\lambda_{\theta}$  decreases.



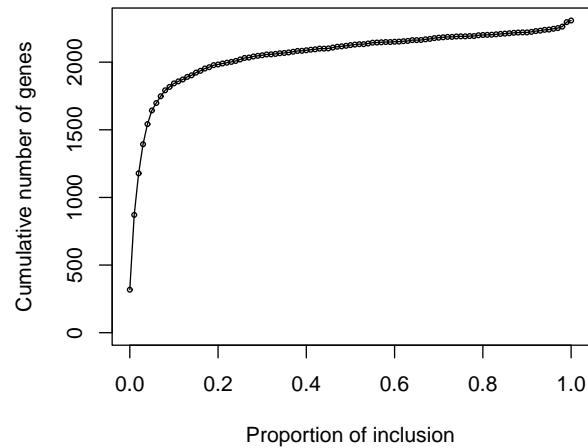


Figure: The number of genes selected less often than or as frequently as a given proportion in 100 runs.

## Summary of the full data analysis

- ▶ The empirical distribution of the number of genes included in one-step updates contained the middle 50% of values between 212 and 228 with median 221.
- ▶ 67 genes were consistently selected for more than 95% of the time.
- ▶ About 2000 genes were selected less than 20% of the time.
- ▶ Gene selection led to reduction in test error rates by 0.0230 on average (from 0.0455 to 0.0225) with standard error of 0.00484.
- ▶ It also reduced the variance of test error rates.

## Concluding remarks

- ▶ Integrate feature selection with kernel methods using  $\ell_1$  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.