

Structured Multicategory Support Vector Machine with ANOVA decomposition

www.stat.ohio-state.edu/~yklee

Yoonkyung Lee

Department of Statistics

The Ohio State University

Predictive learning

- A training data set $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$.
- Functional relationship f between $\mathbf{x} = (x_1, \dots, x_p)$ and y .
 - Regression: continuous y .
 - Classification: categorical y .
- Goodness
 - Prediction accuracy for a given loss $\mathcal{L}(y, f(\mathbf{x}))$.
 - Interpretation.

Support Vector Machines

Vapnik (1995), <http://www.kernel-machines.org>

- Find $f(\mathbf{x}) = b + h(\mathbf{x})$ minimizing

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

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

- Competitive classification accuracy.
- Flexibility - implicit embedding through kernel.
- Handle high dimensional data.
- A black box unless the embedding is explicit.

Feature Selection

- The best subset selection.
- Nonnegative garrote [Breiman, *Technometrics* (1995)]
- Least Absolute Shrinkage and Selection Operator [Tibshirani, *JRSS* (1996)]
- Component Selection and Smoothing Operator [Lin & Zhang, *Technical Report* (2003)]
- Structural modelling with sparse kernels [Gunn & Kandola, *Machine Learning* (2002)]

ANOVA decomposition

Wahba (1990)

● Function:

$$f(\mathbf{x}) = b + \sum_{\alpha=1}^p f_{\alpha}(x_{\alpha}) + \sum_{\alpha < \beta} f_{\alpha\beta}(x_{\alpha}, x_{\beta}) + \dots$$

● 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 $\theta \geq 0$

$$K_{\theta}(\mathbf{x}, \mathbf{x}') = 1 + \sum_{\alpha=1}^p \theta_{\alpha} K_{\alpha}(\mathbf{x}, \mathbf{x}') + \sum_{\alpha < \beta} \theta_{\alpha\beta} K_{\alpha\beta}(\mathbf{x}, \mathbf{x}') + \dots$$

l_1 penalty on θ

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

- 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 $\theta_{\nu} \geq 0, \forall \nu$.

Structured MSVM

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

- 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

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

- By the representer theorem,

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

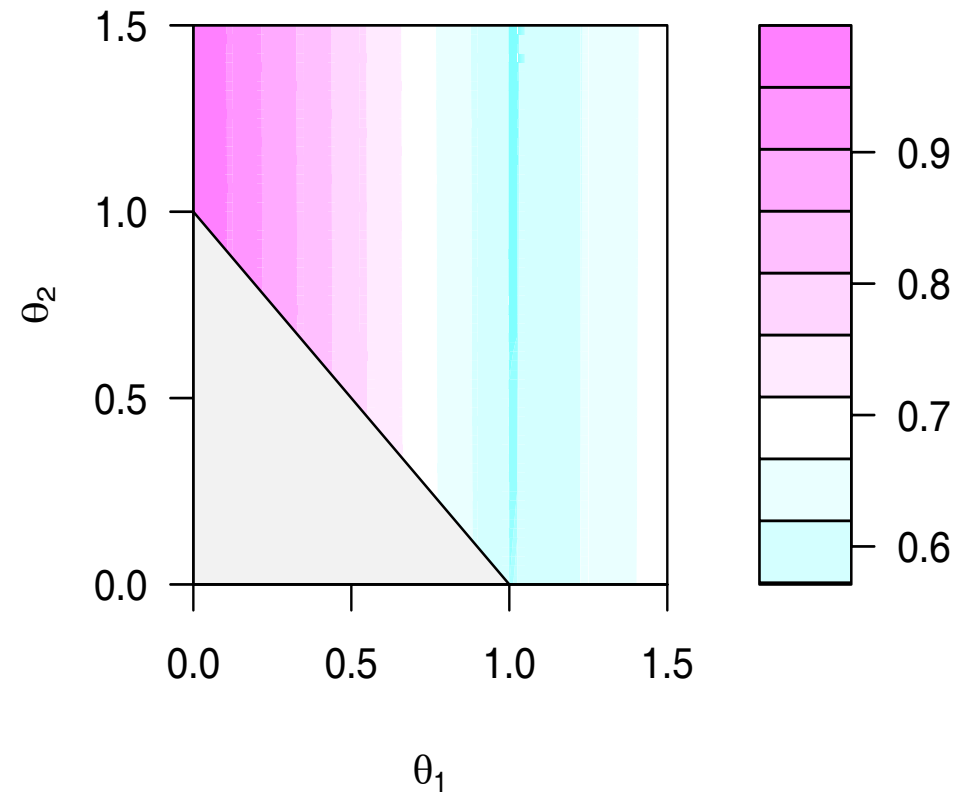
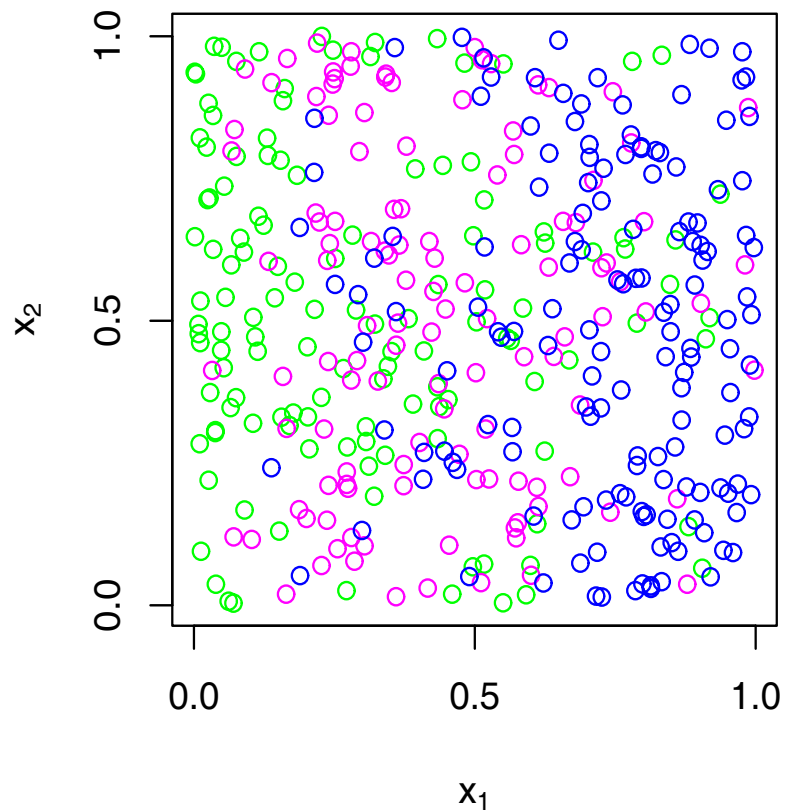
Updating Algorithm

Denoting the objective function by $\Phi(\boldsymbol{\theta}, \mathbf{b}, \mathbf{C})$,

- Initialize $\boldsymbol{\theta}^{(0)} = (1, \dots, 1)^t$ and $(\mathbf{b}^{(0)}, \mathbf{C}^{(0)}) = \operatorname{argmin} \Phi(\boldsymbol{\theta}^{(0)}, \mathbf{b}, \mathbf{C})$.
- At the m -th step ($m = 1, 2, \dots$)
 - $\boldsymbol{\theta}$ -step:
Find $\boldsymbol{\theta}^{(m)}$ minimizing $\Phi(\boldsymbol{\theta}, \mathbf{b}^{(m-1)}, \mathbf{C}^{(m-1)})$ with (\mathbf{b}, \mathbf{C}) fixed.
 - \mathbf{c} -step:
Find $(\mathbf{b}^{(m)}, \mathbf{C}^{(m)})$ minimizing $\Phi(\boldsymbol{\theta}^{(m)}, \mathbf{b}, \mathbf{C})$ with $\boldsymbol{\theta}$ fixed.

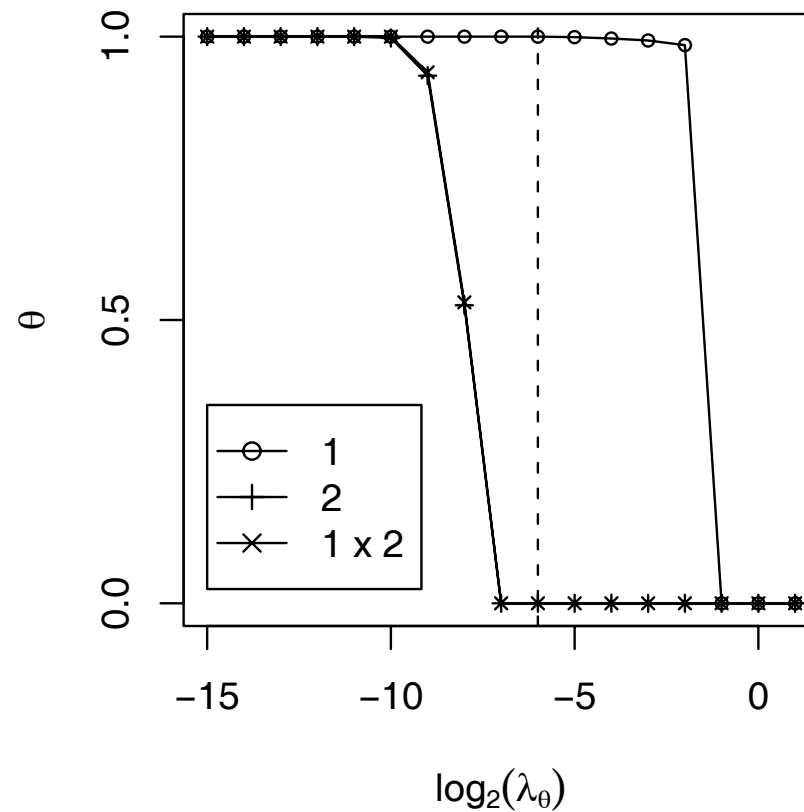
A toy example

- scatter plot and visualization of the θ -step



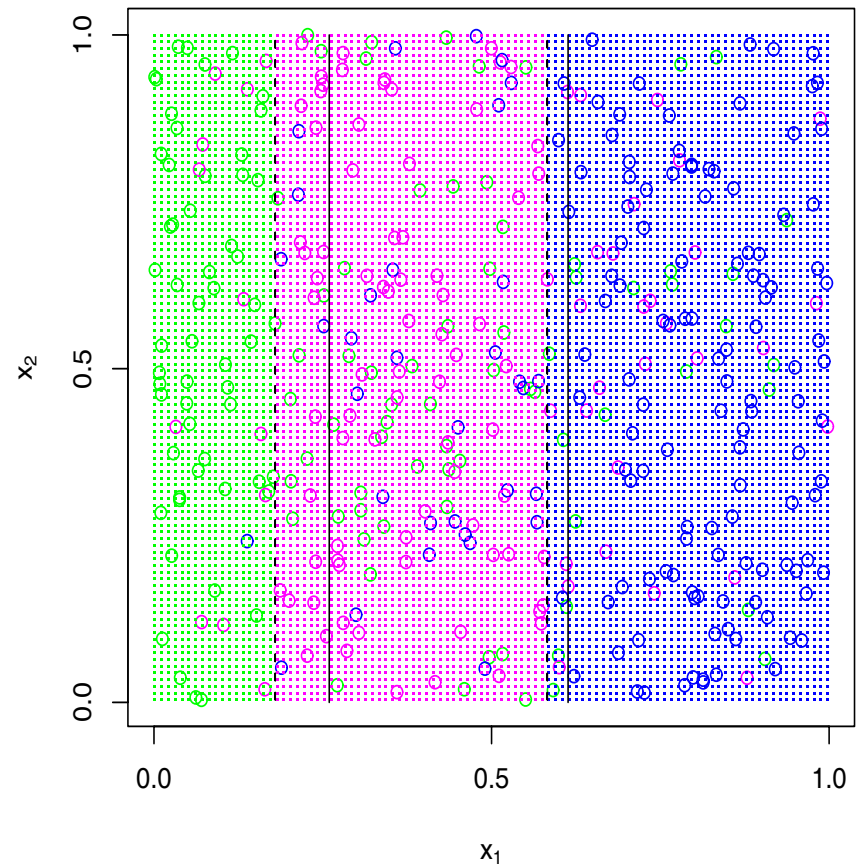
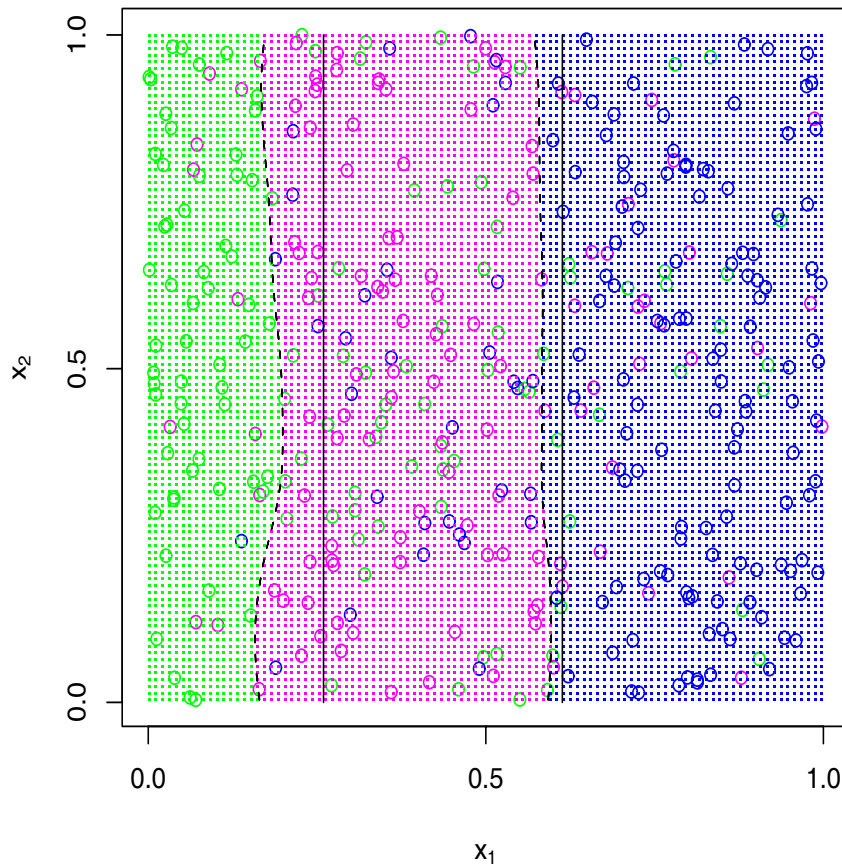
The trajectory of θ

- two-way interaction spline kernel with λ_θ tuned by GCKL



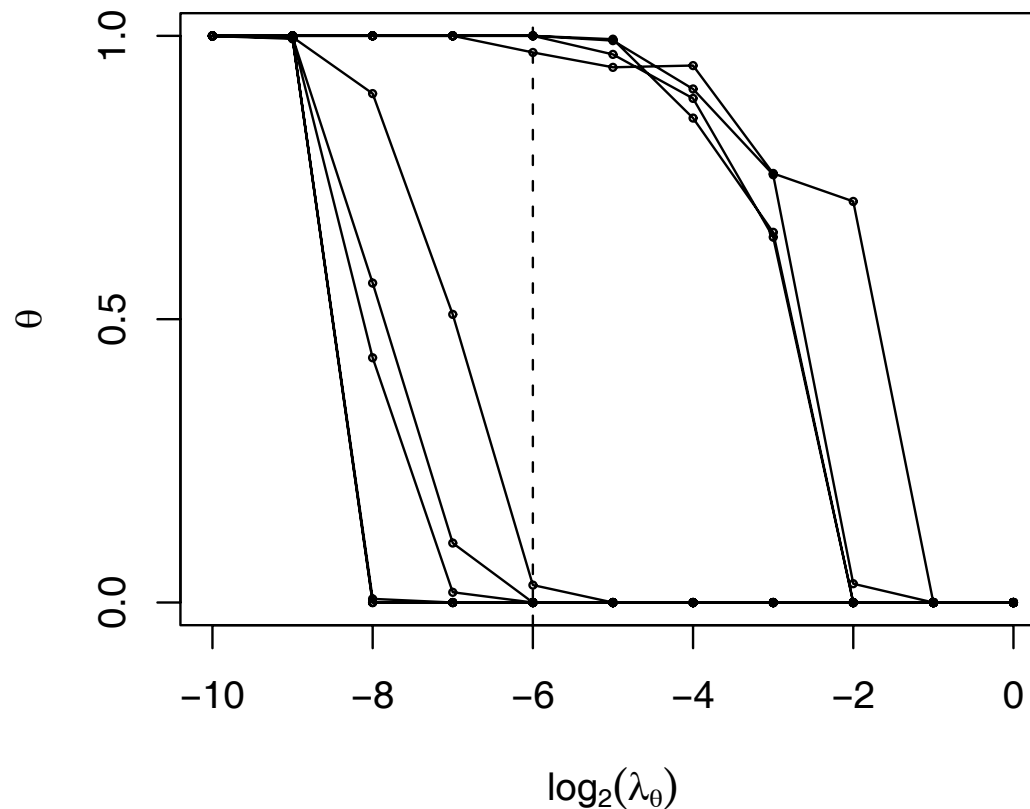
Classification boundaries

- ordinary MSVM (0.3970) vs. structured MSVM (0.3967)



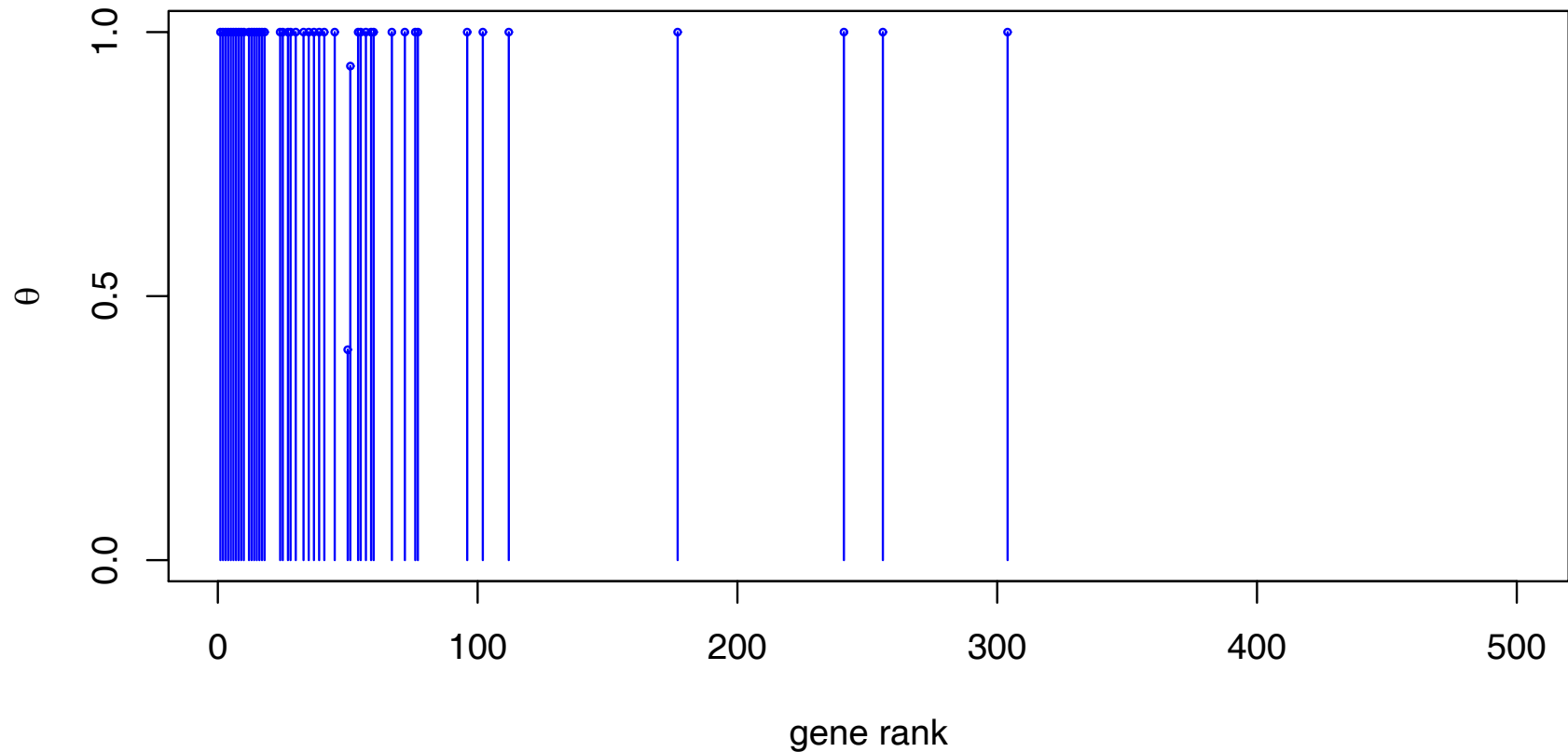
An “apple” example

- $P(Y|X_1, \dots, X_{10})$ depends only on $\sum_{i=1}^4 X_i^2$.
- additive spline kernel with 5-fold CV
($\hat{\theta}_1 = 0.9707$, $\hat{\theta}_2 = \hat{\theta}_3 = \hat{\theta}_4 = 1$, and $\hat{\theta}_5 = 0.0308$).



Gene selection: microarray data

- 2308 genes and four tumor types [Khan et al. *Nature Medicine* (2001)]
- 46 positive rescaling parameters out of 500.



Concluding remarks

- Integrate feature selection with learning classification rule.
- Enhance interpretation without compromising prediction accuracy.
- Characterize the solution path for effective computation and tuning.
- Tailor the structure of component penalty for refined selection.

- Joint work with Yuwon Kim (SNU), Ja-Yong Koo (Inha Univ.), and Sangjun Lee (SNU) in Korea.
- Manuscript to be posted at www.stat.ohio-state.edu/~yklee.