A Sparse Solution Approach to Gene Selection for Cancer Diagnosis Using Microarray Data

Yoonkyung Lee
Department of Statistics
The Ohio State University
http://www.stat.ohio-state.edu/∼yklee

May 13, 2005

Microarray Data

- Relative amount of mRNAs of tens of thousands of genes. (e.g. Affymetrix Human U133a GeneChips: 22K genes)
- Gene expression profiles as fingerprints at the molecular level.
- Potential for cancer diagnosis or prognosis.
- Examples
  - Golub et al. (1999) *Science*: acute leukemia classifi cation.
  - Perou et al. (2000) *Nature*: breast cancer subclass identifi cation.
  - Van't Veer et al. (2002) *Nature*: breast cancer prognosis.

Objectives and related issues

- Accurate diagnosis or prognosis.
- Interpretability.
- Subset selection or dimension reduction.
- Systematic approach to gene (variable) selection.
- Assessment of the variability in analysis results.

Small Round Blue Cell Tumors of Childhood

- Khan et al. (2001) in *Nature Medicine*
- Tumor types: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS).
- Number of genes : 2308
- Class distribution of data set

Data set	EWS	BL(NHL)	NB	RMS	total
Training set	23	8	12	20	63
Test set	6	3	6	5	20
Total	29	11	18	25	83

Classifi cation problem

- A training data set $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$, where $y_i \in \{1, \ldots, k\}$.
- Learn a functional relationship f between $\boldsymbol{x} = (x_1, \ldots, x_p)$ and y from the training data, which can be generalized to novel cases.

Measures of marginal association for gene selection

Pick genes with the largest marginal association.

- Two-class: two-sample t-test statistic, correlation, and etc.
- Multi-class: Dudoit et al. (2000)
  For gene ℓ, the ratio of between classes sum of squares to within class sum of squares is defined as

$$\frac{BSS(\ell)}{WSS(\ell)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} I(y_i = j)(\bar{x}_{\cdot\ell}^{(j)} - \bar{x}_{\cdot\ell})^2}{\sum_{i=1}^{n}\sum_{j=1}^{k} I(y_i = j)(x_{i\ell} - \bar{x}_{\cdot\ell}^{(j)})^2}.$$

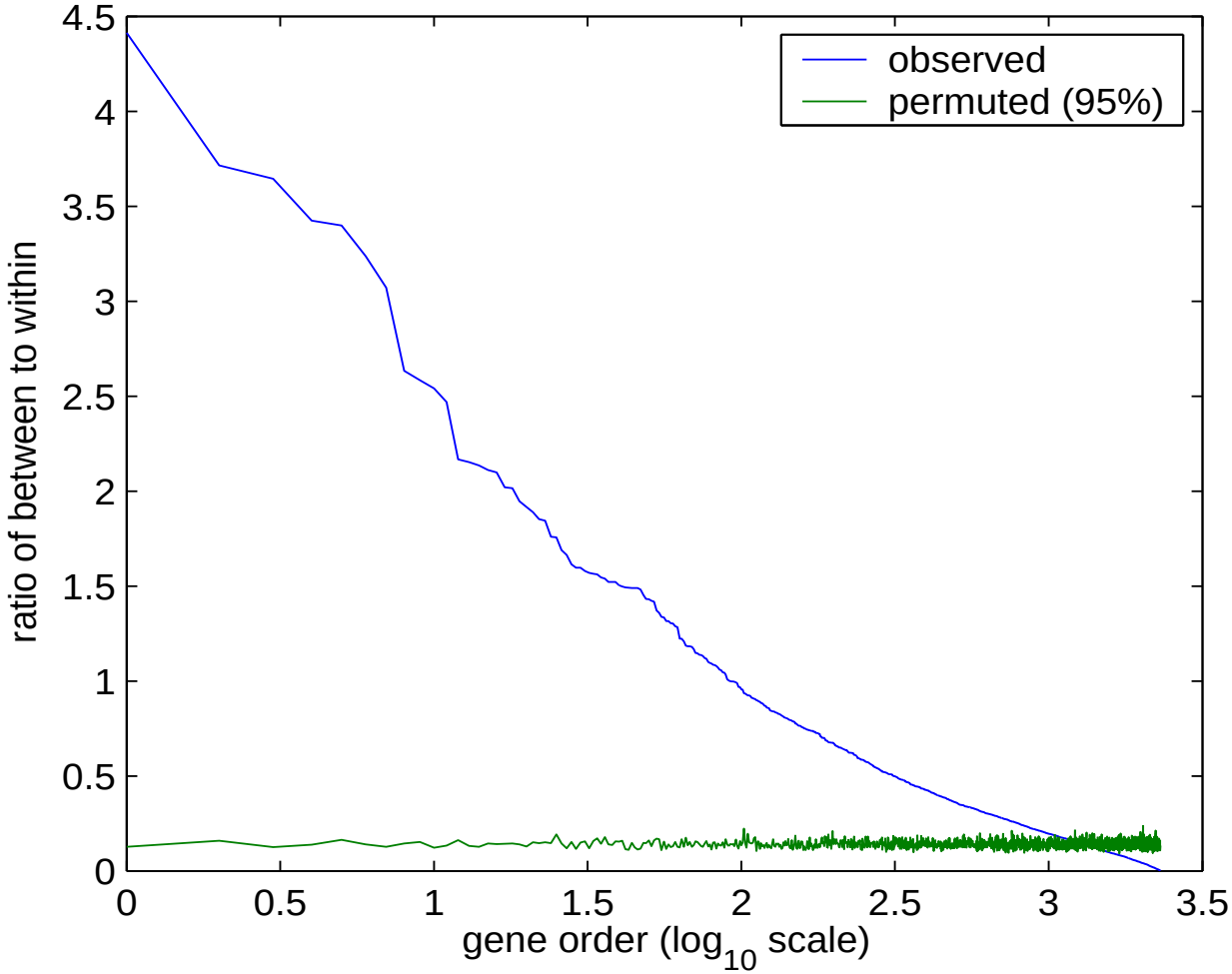

Figure: Observed ratios of between-class SS to within-class SS and the 95 percentiles of the corresponding ratios for expression levels with randomly permuted class labels.

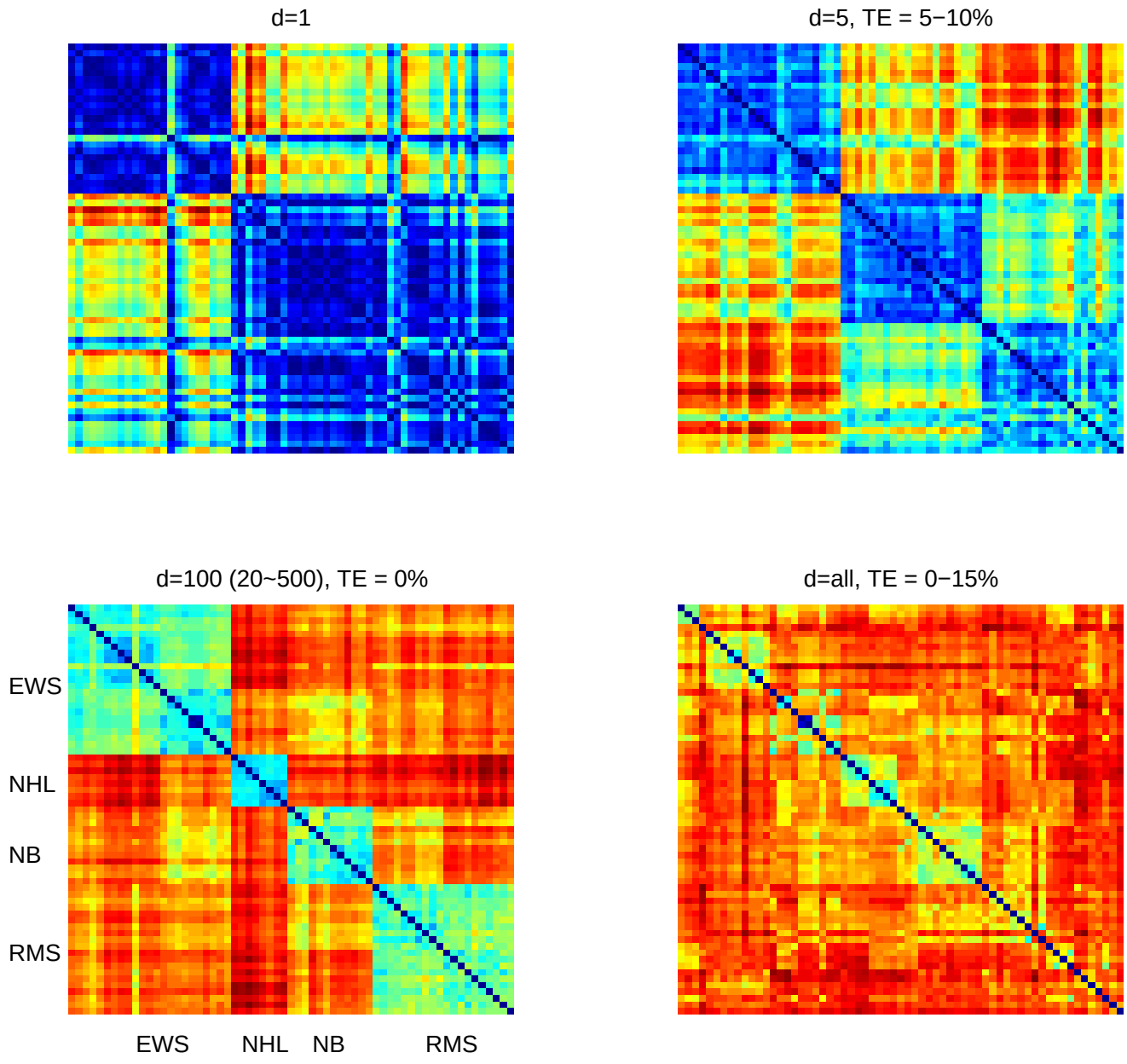

Figure: Pairwise distance matrices for the training data as the numbers of genes included change, and test error rates of MSVM with Gaussian kernel.

Support Vector Machines

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

- $y_i \in \{-1, 1\}$.
- Find $f(\boldsymbol{x}) = b + h(\boldsymbol{x})$ with $h \in \mathcal{H}_K$ minimizing

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

  Then $\hat{f}(\boldsymbol{x}) = \hat{b} + \sum_{i=1}^{n} \hat{c}_i K(\boldsymbol{x}_i, \boldsymbol{x})$, where K: a bivariate positive definite function called a reproducing kernel.
- Classification rule: $\phi(\boldsymbol{x}) = sign[f(\boldsymbol{x})]$.

Hinge loss

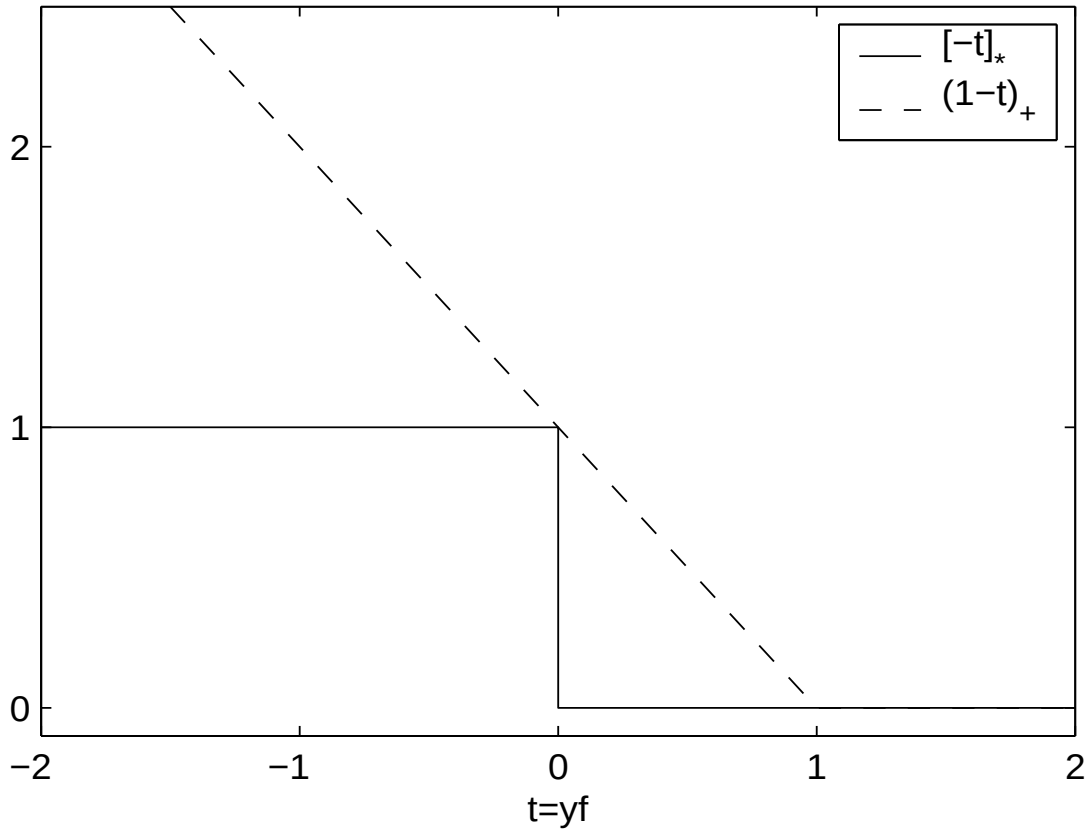

Figure: $(1 - yf(\boldsymbol{x}))_+$ is an upper bound of the misclassifi cation loss function $I(y \neq \phi(\boldsymbol{x})) = [-yf(\boldsymbol{x})]_* \leq (1 - yf(\boldsymbol{x}))_+$ where $[t]_* = I(t \geq 0)$ and $(t)_+ = \max\{t, 0\}$.

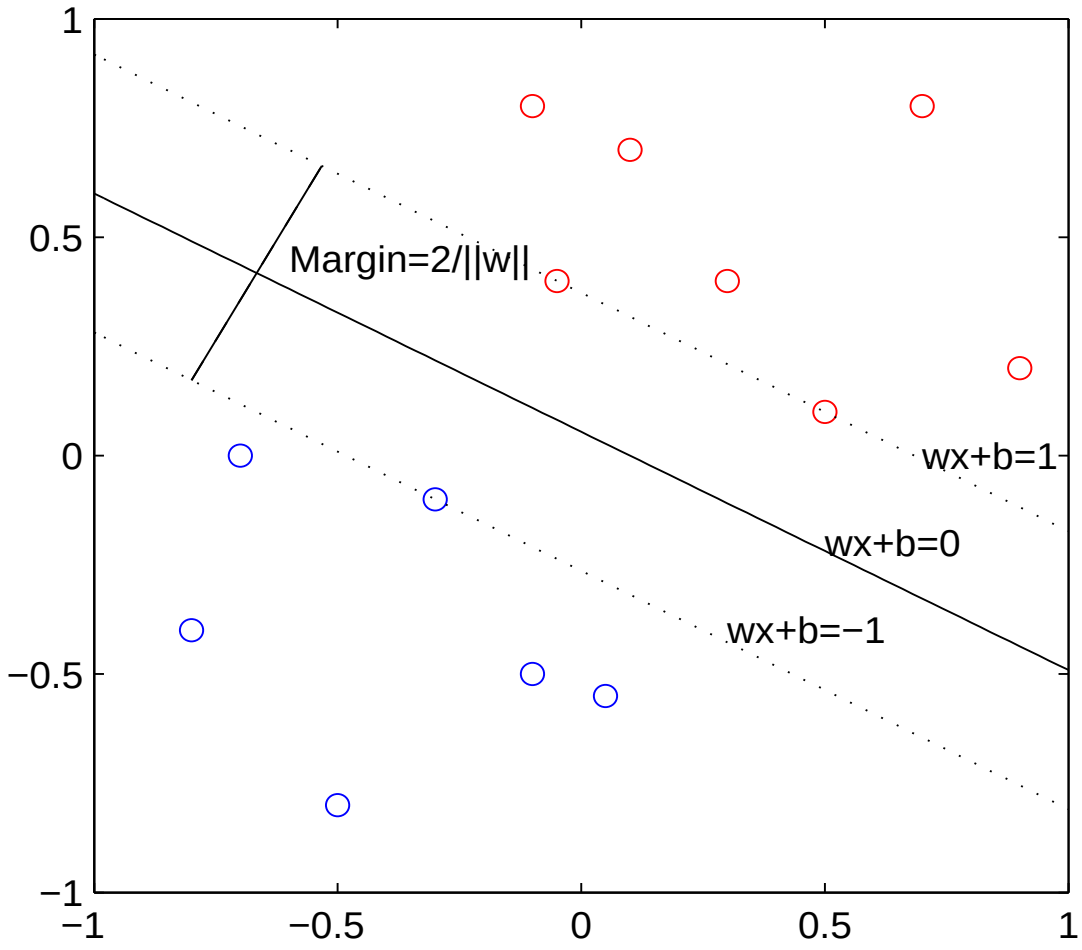

Figure: Linear Support Vector Machine example in separable case. Red and blue circles indicate positive and negative examples, respectively. The solid line is the separating hyperplane with the maximum margin in this example.

Characteristics of Support Vector Machines

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

Variable (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)]

Functional ANOVA decomposition

Wahba (1990)

- 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.):
  $K(\boldsymbol{x}, \boldsymbol{x}') = 1 + \sum_{\alpha=1}^{p} K_\alpha(\boldsymbol{x}, \boldsymbol{x}') + \sum_{\alpha<\beta} K_{\alpha\beta}(\boldsymbol{x}, \boldsymbol{x}') + \cdots$
- Modification of r.k. by rescaling parameters $\boldsymbol{\theta} \geq 0$
  $K_\theta(\boldsymbol{x}, \boldsymbol{x}') = 1 + \sum_{\alpha=1}^{p} \theta_\alpha K_\alpha(\boldsymbol{x}, \boldsymbol{x}') + \sum_{\alpha<\beta} \theta_{\alpha\beta} K_{\alpha\beta}(\boldsymbol{x}, \boldsymbol{x}') + \cdots$

l_1 penalty on $\boldsymbol{\theta}$

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

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

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

Structured MSVM with ANOVA decomposition

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

- 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

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{L}(\boldsymbol{y}_i) \cdot (\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{y}_i)_+ + \frac{\lambda}{2}\sum_{j=1}^{k}\left(\sum_{\nu=1}^{d} \theta_\nu^{-1} \|P^\nu h^j\|^2\right)$$

$$+\lambda_\theta \sum_{\nu=1}^{d} \theta_\nu \text{ subject to } \theta_\nu \geq 0, \text{ for } \nu = 1, \ldots, d.$$

- By the representer theorem, $\hat{f}^j(\boldsymbol{x}) = \hat{b}^j + \sum_{i=1}^{n} \hat{c}_i^j \sum_{\nu=1}^{d} \theta_\nu K_\nu(\boldsymbol{x}_i, \boldsymbol{x})$.

Updating Algorithm

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

- Initialize $\boldsymbol{\theta}^{(0)} = (1, \ldots, 1)^t$ and $(\boldsymbol{b}^{(0)}, \boldsymbol{C}^{(0)}) = \operatorname{argmin} \Phi(\boldsymbol{\theta}^{(0)}, \boldsymbol{b}, \boldsymbol{C})$.
- At the m-th step ($m = 1, 2, \ldots$)
  - θ-step:
    Find $\boldsymbol{\theta}^{(m)}$ minimizing $\Phi(\boldsymbol{\theta}, \boldsymbol{b}^{(m-1)}, \boldsymbol{C}^{(m-1)})$ with $(\boldsymbol{b}, \boldsymbol{C})$ fi xed.
  - c-step:
    Find $(\boldsymbol{b}^{(m)}, \boldsymbol{C}^{(m)})$ minimizing $\Phi(\boldsymbol{\theta}^{(m)}, \boldsymbol{b}, \boldsymbol{C})$ with $\boldsymbol{\theta}$ fi xed.
- One-step update can be used in practice.

A toy example: visualization of the θ-step

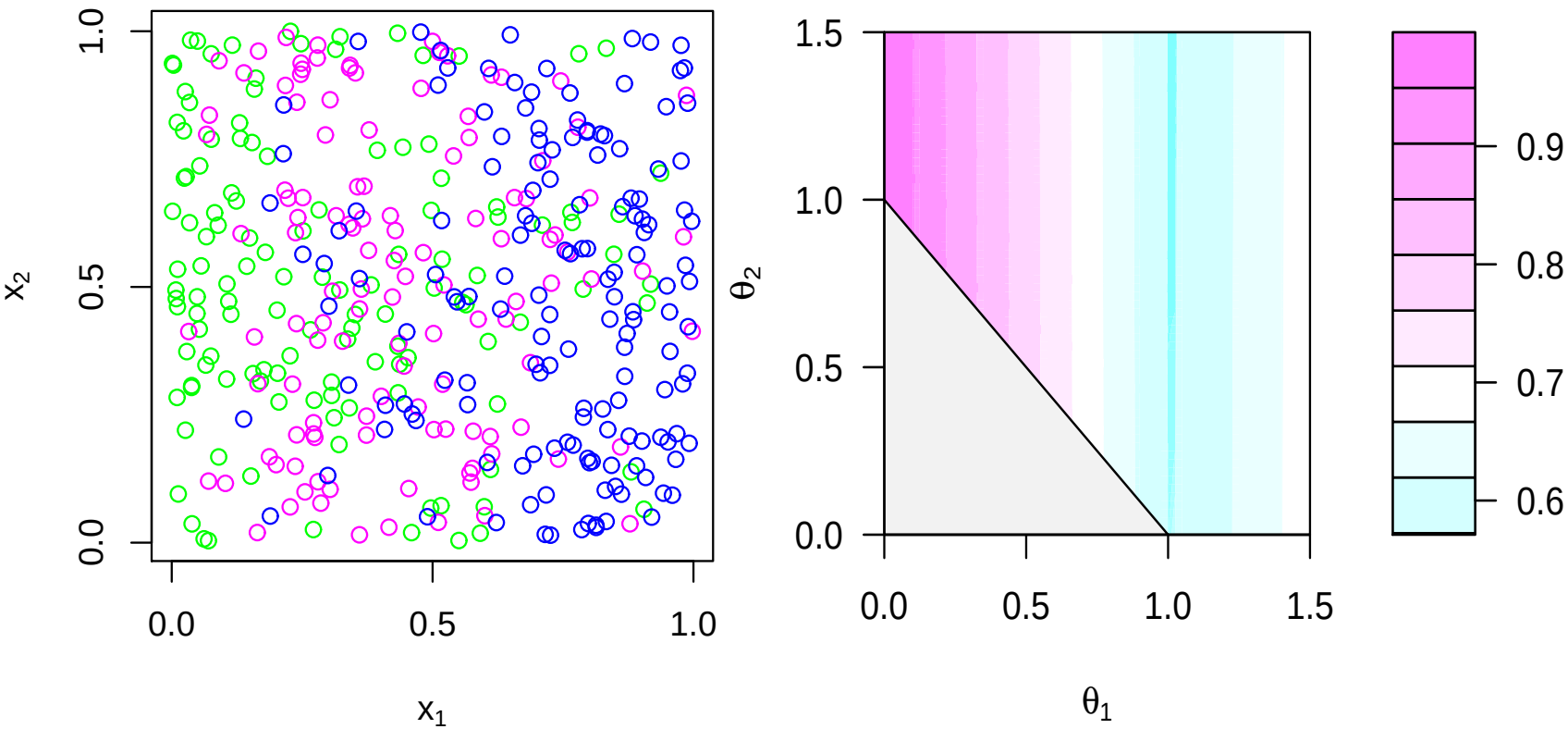

Figure: Left: the scatter plot of two covariates with class labels distinguished by color. Right: the feasible region of (θ_1, θ_2) in gray and a level plot of the θ-step objective function.

The trajectory of $\boldsymbol{\theta}$ as a function of λ_θ

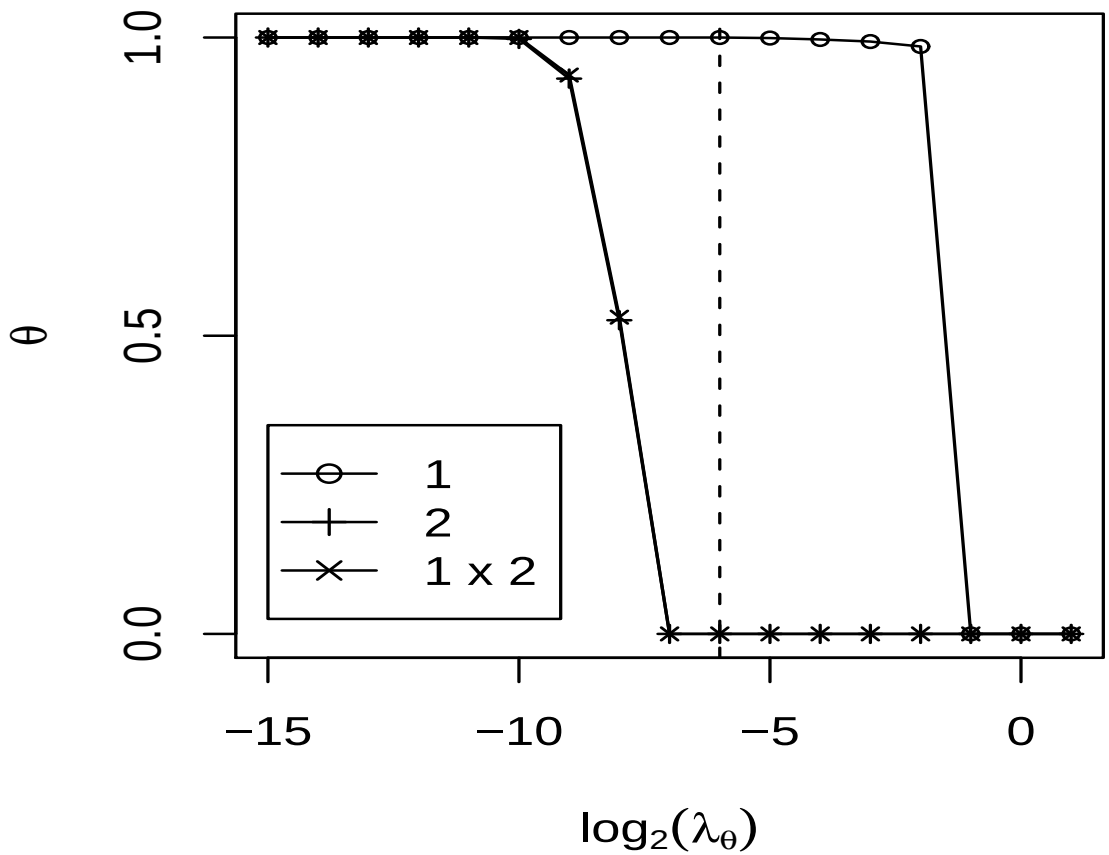

Figure: The trajectories of θ_1; $\circ$, θ_2; $+$, and θ_{12}; $\times$ with the two-way interaction spline kernel with λ_θ tuned by GCKL. The overlap of $+$ and $\times$ indicates that the trajectories of θ_2 and θ_{12} are almost indistinguishable.

A synthetic miniature data set

- It consists of 100 genes from Khan et al. (63 training and 20 test cases)
- Use the F-ratio for each gene based on the training cases only.
- The top 20 genes as variables truly associated with the class.
- The bottom 80 genes with the class label randomly jumbled as irrelevant variables.
- 100 replicates by bootstrapping samples from this miniature data set keeping the class proportions the same as the original data.

The proportion of gene inclusion (%)

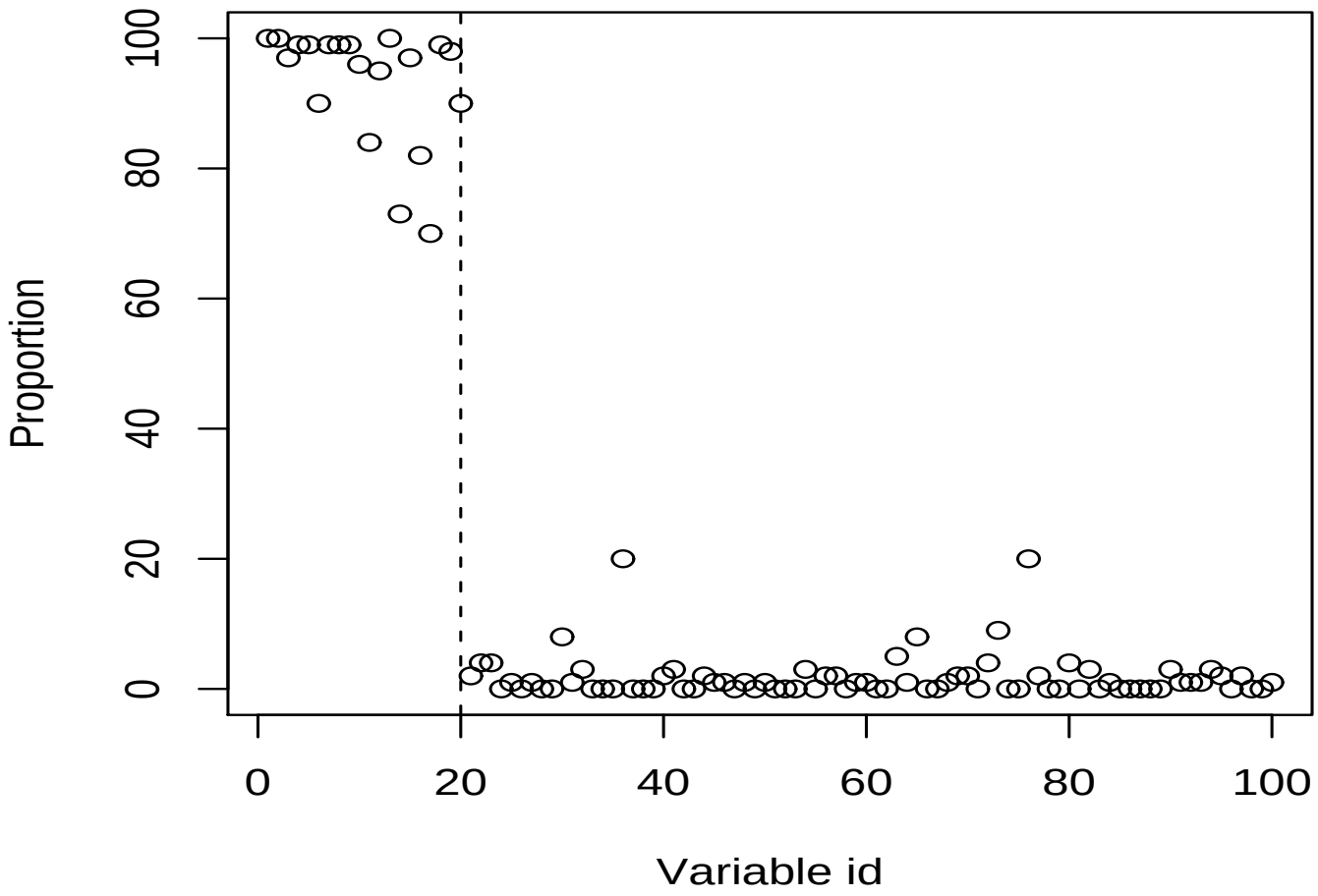

Figure: The proportion of inclusion (%) of each gene in the fi nal classifi ers over 100 runs. The dotted line delimits informative variables from noninformative ones. 10-fold CV was used for tuning.

Summary of the synthetic miniature data analysis with Structured MSVM

- 13 informative genes were selected more than 95% of the runs.
- 74 noninformative genes were picked up less than 5% of the runs.
- Gene selection resulted in decrease in the test error rate over the 20 test cases on average of 0.0095 (from 0.0555 to 0.0460) with standard error 0.00413.

The original data with 2308 genes

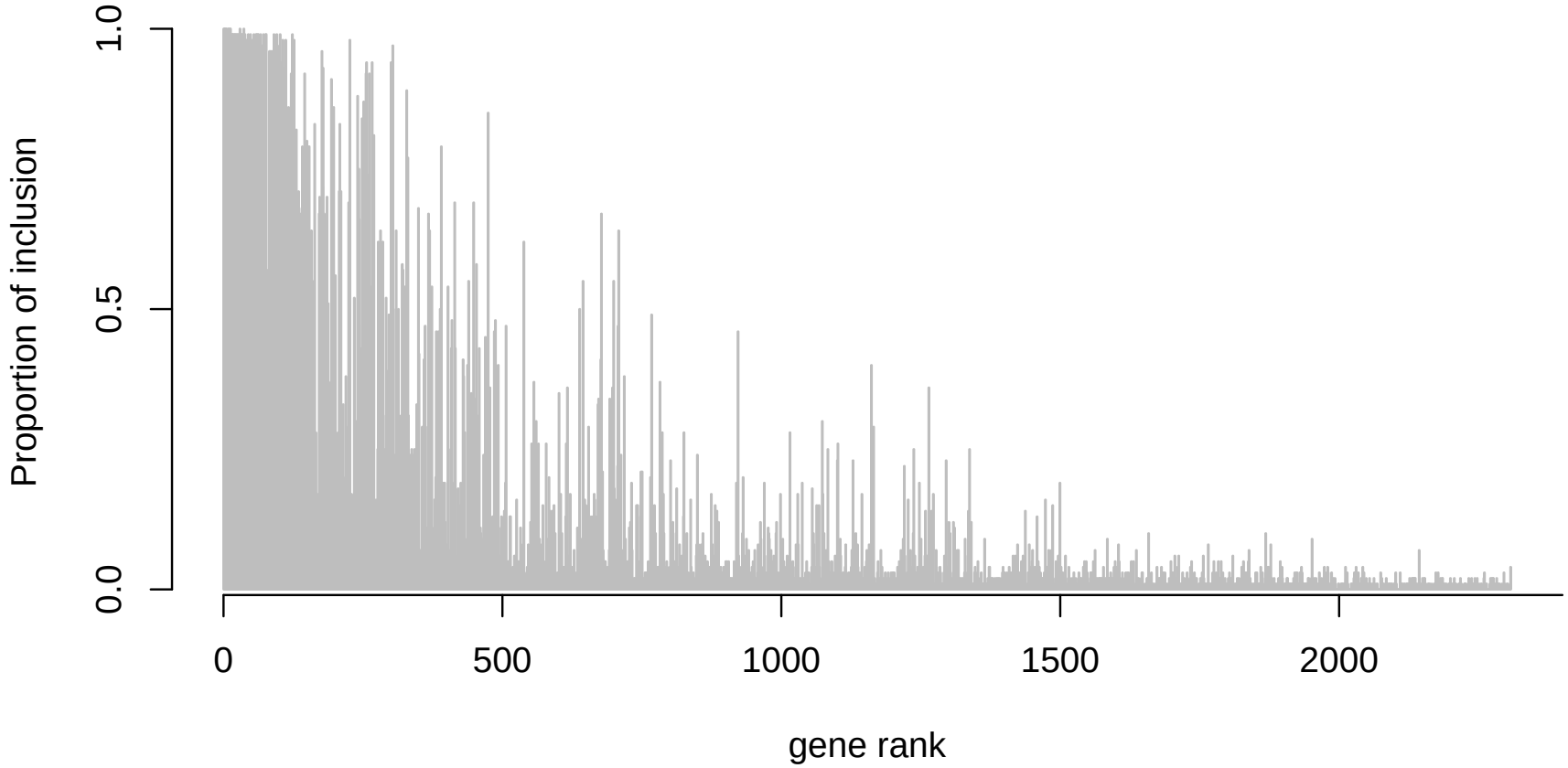

Figure: The proportion of selection of each gene in one-step updated SMSVMs for 100 bootstrap samples. Genes are presented in the order of marginal rank in the original sample.

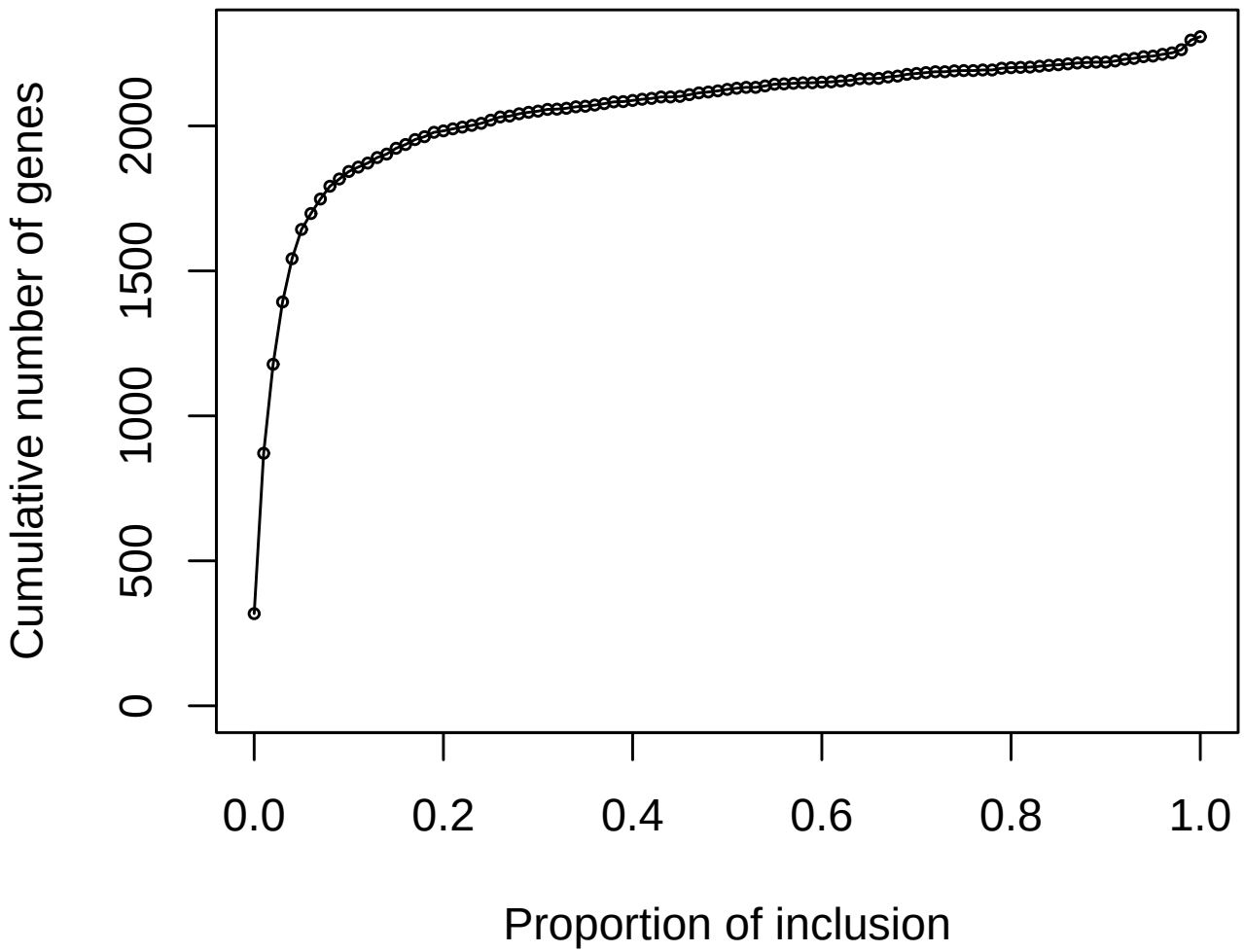

Figure: The number of genes selected less often than or as frequently as a given proportion in 100 runs.

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 learning classification rules.
- Enhance interpretation without compromising prediction accuracy.
- Characterize the solution path of SMSVM for effective computation and tuning.
  - Efron et al. LAR (2004) and Hastie et al. SVM solution path (2004).
  - c-step: the entire spectrum of solutions from the simplest majority rule to the complete overfi t to data.
  - θ-step: the entire spectrum of solutions from the constant model to the full model with all the variables.

The following papers are available from www.stat.ohio-state.edu/∼yklee.

- *Structured Multicategory Support Vector Machine with ANOVA decomposition*, Lee, Y., Kim, Y., Lee, S., and Koo, J.-Y., Technical Report No. 743, The Ohio State University, 2004.
- *Characterizing the Solution Path of Multicategory Support Vector Machines*, Lee, Y. and Cui, Z., Technical Report No. 754, The Ohio State University, 2005.