Structured Statistical Learning with Support Vector Machine for Feature Selection and Prediction

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

Predictive 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.
- Examples include
  Regression: continuous $y \in R$, and
  Classification: categorical $y \in \{1, \ldots, k\}$.

Goodness of a learning method

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

Support Vector Machine

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

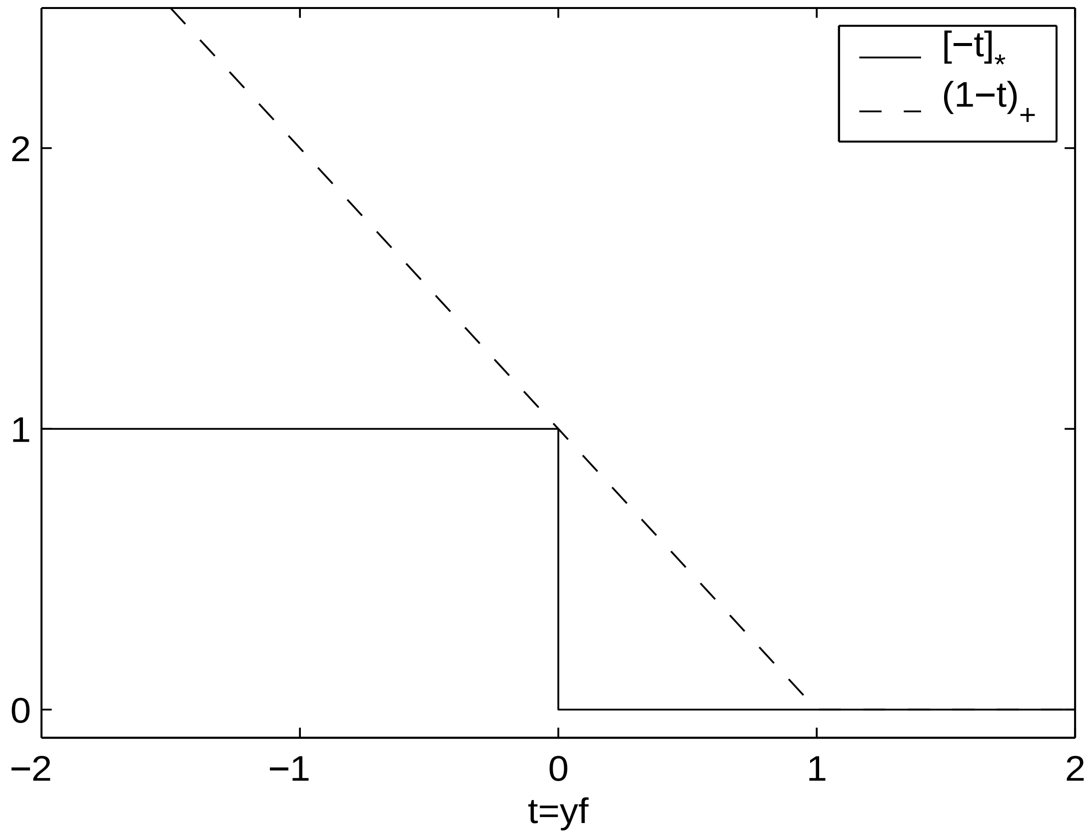

Figure: $(1 - yf(\boldsymbol{x}))_+$ is an upper bound of the misclassifi cation loss function $I(y \neq \phi(\boldsymbol{x})) = [-yf(\boldsymbol{x})]_* \leq (1 - yf(\boldsymbol{x}))_+$ where $[t]_* = I(t \geq 0)$ and $(t)_+ = \max\{t, 0\}$.

Feature Selection

- Linear SVM with ℓ_1 penalty [Bradley & Mangasarian (1998)].
- Recursive feature selection [Guyon et al. (2002)].
- Rescaling parameters [Chapelle et al. (2002)].
- Least Absolute Shrinkage and Selection Operator [Tibshirani (1996)].
- COmponent Selection and Smoothing Operator [Lin & Zhang (2003)].
- Structural modelling with sparse kernels [Gunn & Kandola (2002)].

Strategy for feature selection

- Structured representation of f.
- A sparse solution approach with ℓ_1 penalty.
- A unified treatment of the nonlinear and multiclass case.
- Not expensive additional computation.
- Systematic elaboration of f with features.

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

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

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

Related to kernel learning

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

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

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

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

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

- $\boldsymbol{y} = (y^1, \ldots, y^k)$: class code with $y^j = 1$ and $-1/(k-1)$ elsewhere, if $y = j$ and $\boldsymbol{L}(\boldsymbol{y})$: misclassification cost.
- By the representer theorem, $\hat{f}^j(\boldsymbol{x}) = \hat{b}^j + \sum_{i=1}^{n} \hat{c}_i^j \left[\sum_{\nu=1}^{d} \theta_\nu K_\nu(\boldsymbol{x}_i, \boldsymbol{x})\right]$.

Updating Algorithm

Letting $\boldsymbol{C} = (\{b^j\}, \{c_i^j\})$ 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 majority rule to the complete overfit to data as λ decreases.
- θ-step solutions range from the constant model to the full model with all the variables as λ_θ decreases.
- Any computational shortcut to get the entire regularization path?
  e.g. Least Angle Regression [Efron et al. (2004)] and SVM solution path [Hastie et al. (2004)].

c-step regularization path

- Extension of the binary SVM solution path [Hastie et al. (2004)].
- By the Karush-Kuhn-Tucker (KKT) complementarity conditions, the MSVM solution at λ satisfies that for i, j

$$\alpha_i^j \, (f_i^j - y_i^j - \xi_i^j) = 0$$
$$(L_{cat(i)}^j - \alpha_i^j) \, \xi_i^j = 0$$
$$0 \leq \alpha_i^j \leq L_{cat(i)}^j \text{ and } \xi_i^j \geq 0$$

where $f_i^j = \hat{f}_\lambda^j(\boldsymbol{x}_i)$ and $cat(i)$: the category of y_i, thus $(L_{cat(i)}^1, \ldots, L_{cat(i)}^k) = \boldsymbol{L}(\boldsymbol{y}_i)$.

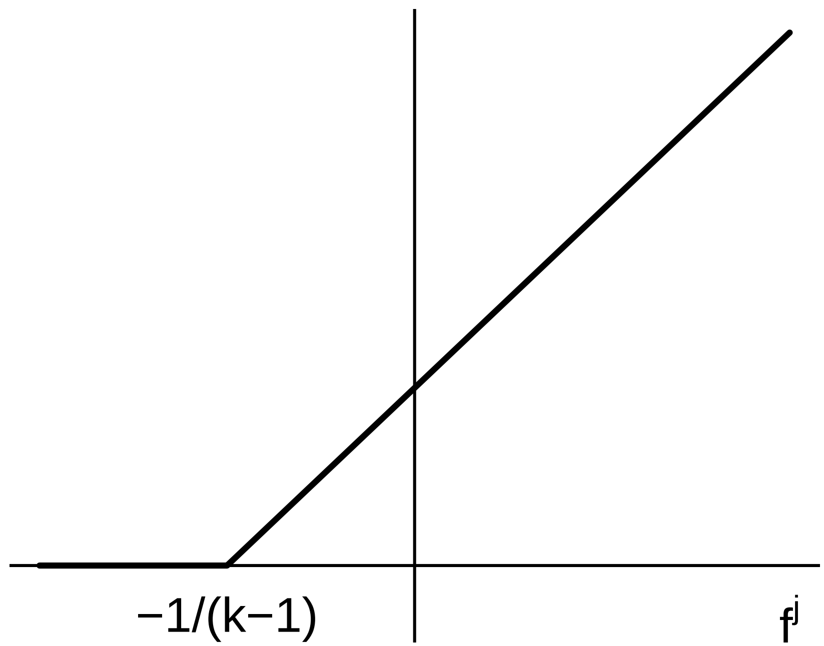

Figure: MSVM component loss $(f^j - y^j)_+$ where $y^j = -1/(k-1)$.

$$\mathcal{E} = \{(i,j)|\ f_i^j - y_i^j = 0,\ \xi_i^j = 0,\ 0 \le \alpha_i^j \le L^j_{cat(i)}\}$$ Elbow set,

$$\mathcal{U} = \{(i,j)|\ f_i^j - y_i^j > 0,\ \xi_i^j > 0,\ \alpha_i^j = L^j_{cat(i)}\}$$ Upper set,

$$\mathcal{L} = \{(i,j)|\ f_i^j - y_i^j < 0,\ \xi_i^j = 0,\ \alpha_i^j = 0\}$$ Lower set.

Characterization of the entire solution path

- Keep track of the events that change the elbow set.
- $\lambda_0 > \lambda_1 > \lambda_2 > \ldots$, a decreasing sequence of breakpoints of λ at which the elbow set $\mathcal{E}$ changes.
- Piecewise linearity of the solution:

  The coefficient path of the MSVM is linear in $1/\lambda$ on the interval $(\lambda_{\ell+1}, \lambda_\ell)$.

- Construct the path sequentially by solving a system of linear equations.

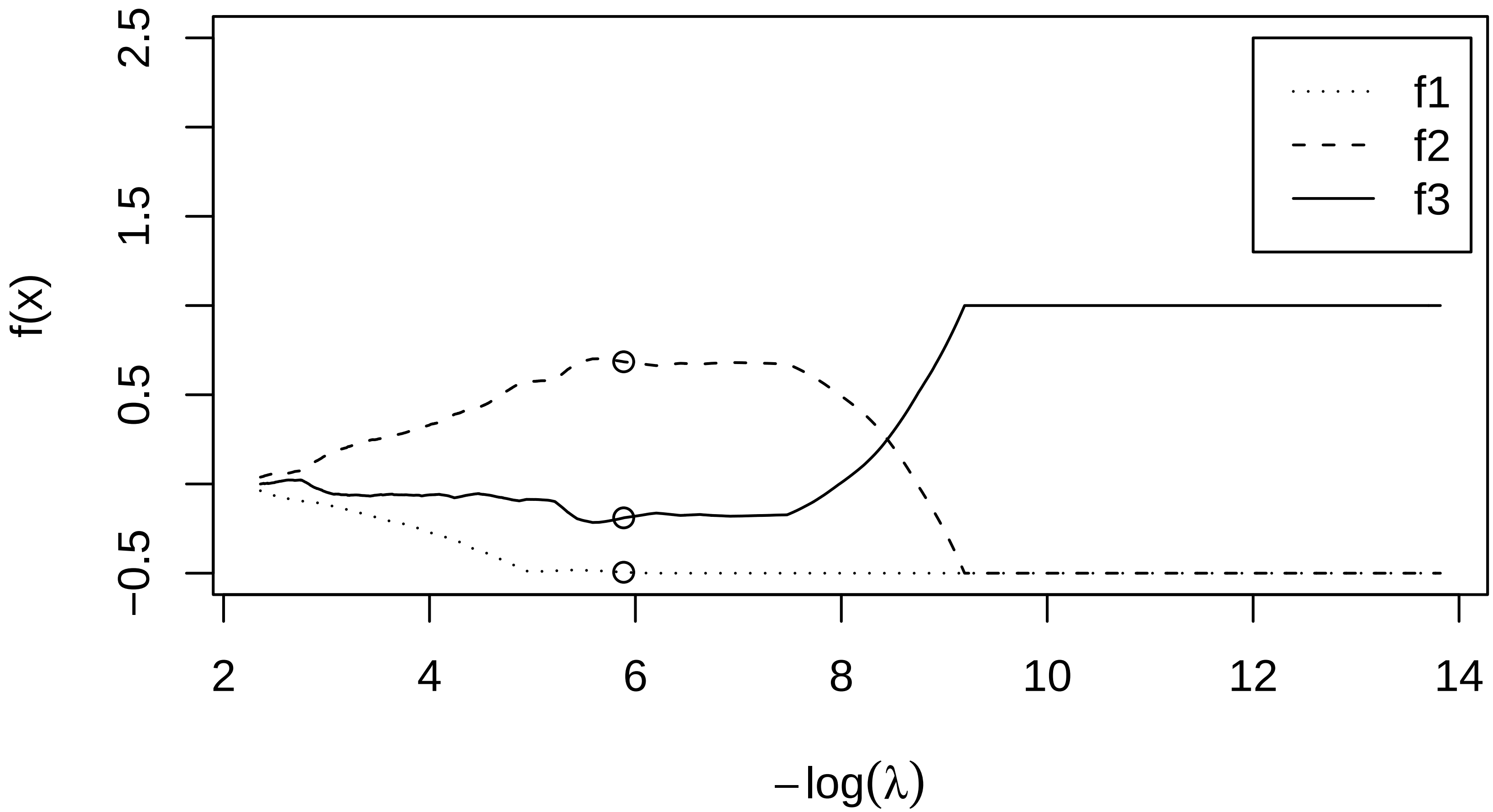

Figure: The entire paths of $\hat{f}^1_\lambda(\boldsymbol{x}_i)$, $\hat{f}^2_\lambda(\boldsymbol{x}_i)$, and $\hat{f}^3_\lambda(\boldsymbol{x}_i)$ for an outlying instance $\boldsymbol{x}_i$ from class 3. The circles correspond to λ with the minimum test error rate.

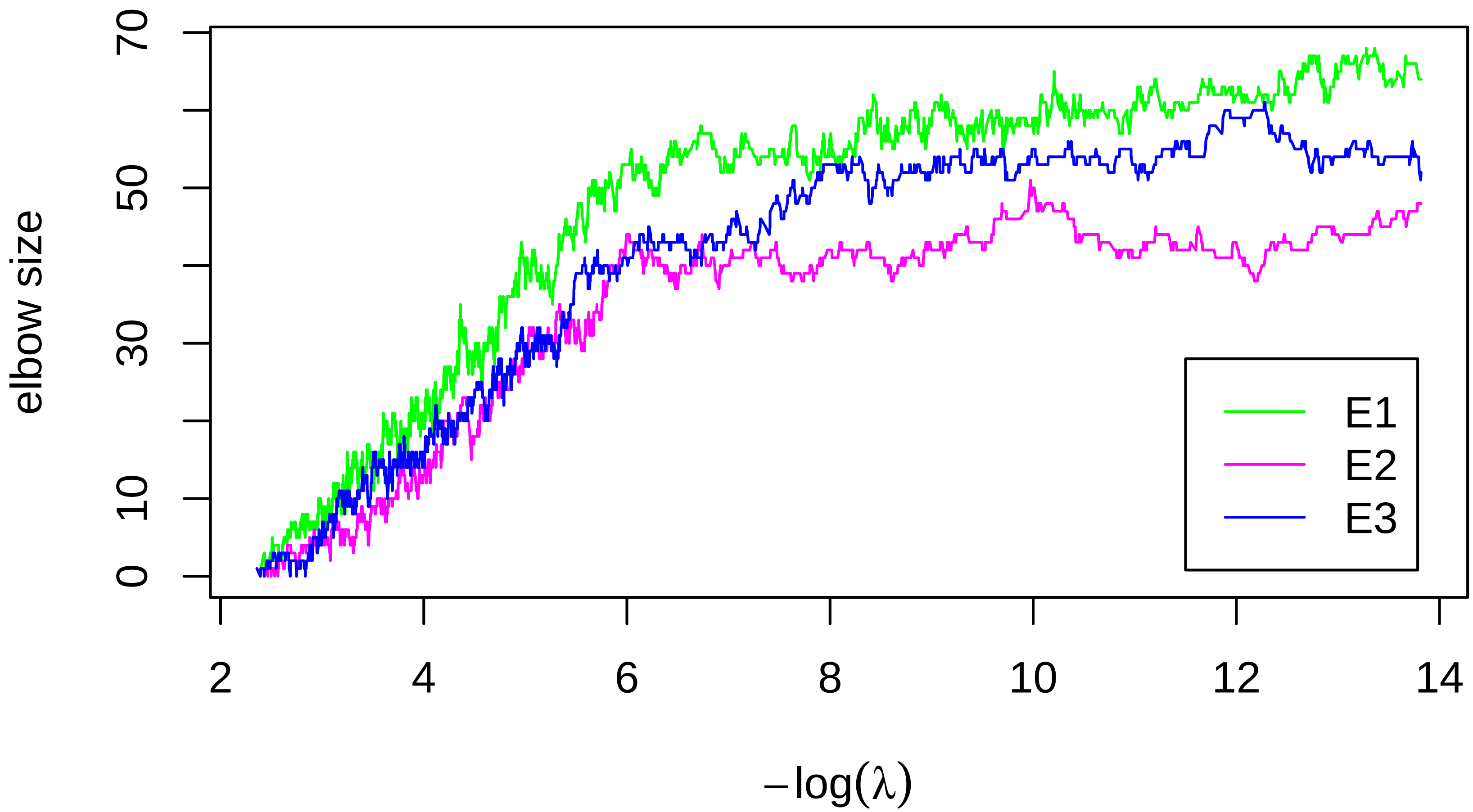

Figure: The size of elbow set $\mathcal{E}_\ell^j$ for three classes as a function λ.

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

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 (%)

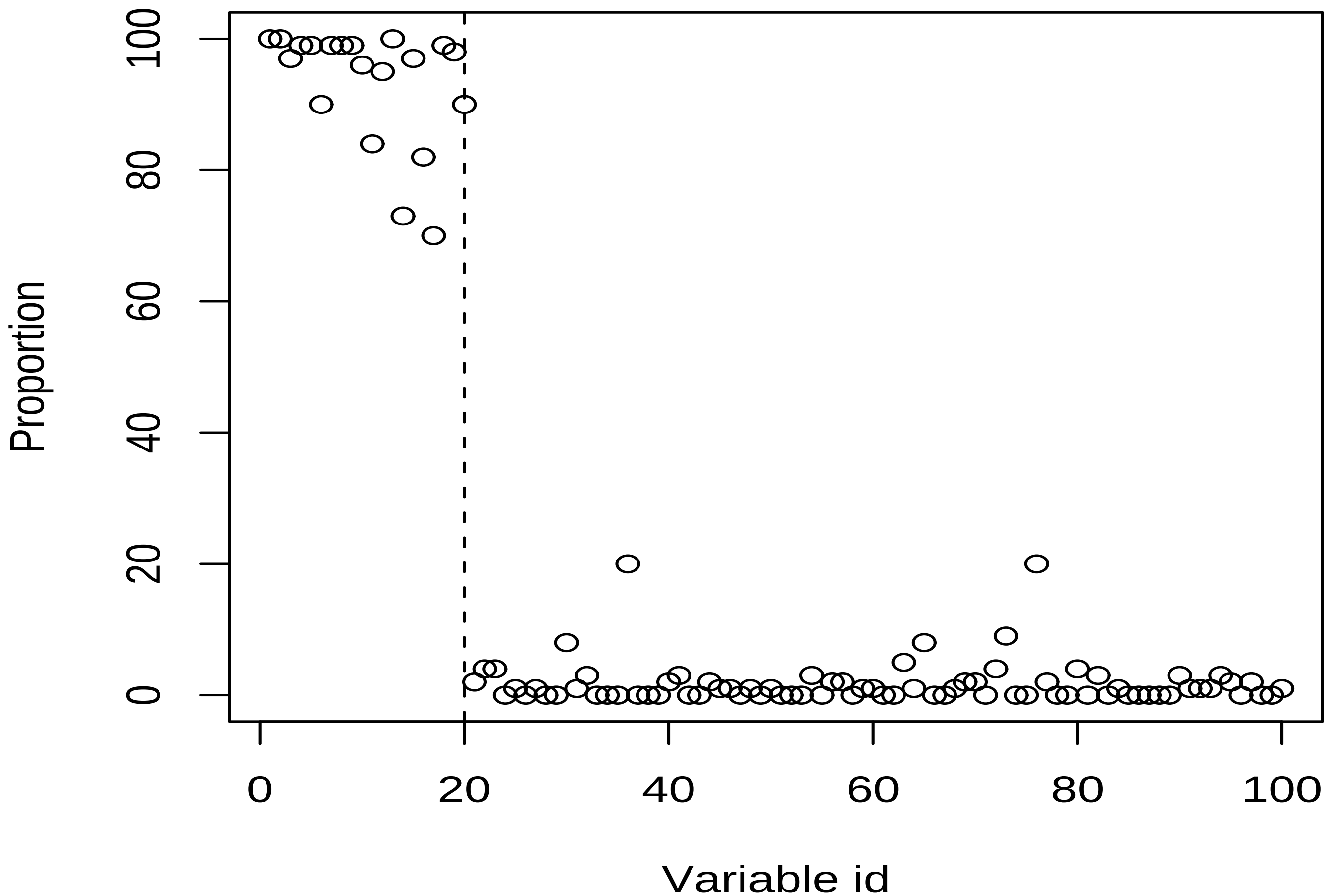

Figure: The proportion of inclusion (%) of each gene in the final classifiers over 100 runs. The dotted line delimits informative variables from noninformative ones. 10-fold CV was used for tuning.

The original data with 2308 genes

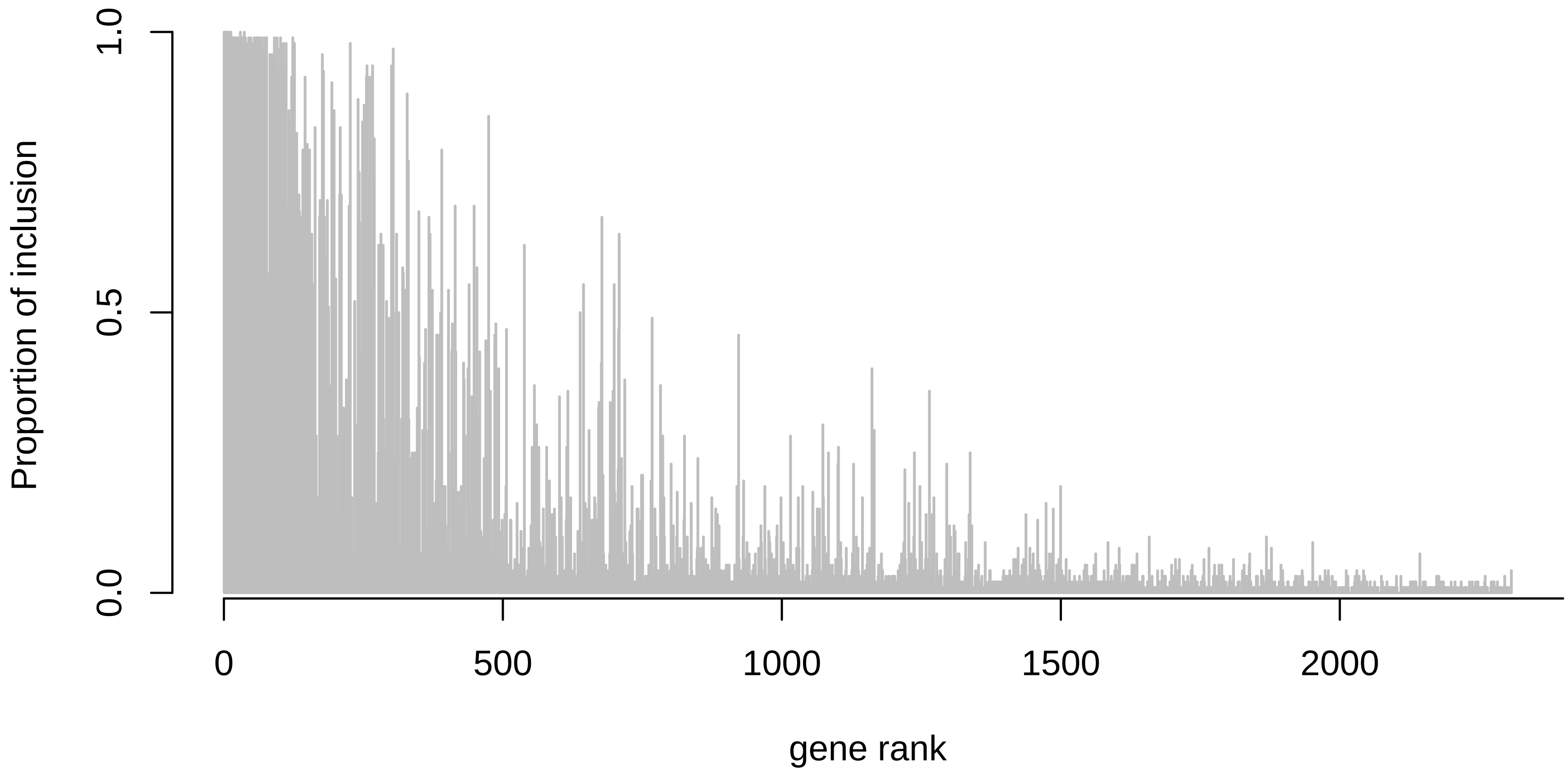

Figure: The proportion of selection of each gene in one-step updated SMSVMs for 100 bootstrap samples. Genes are presented in the order of marginal rank in the original sample.

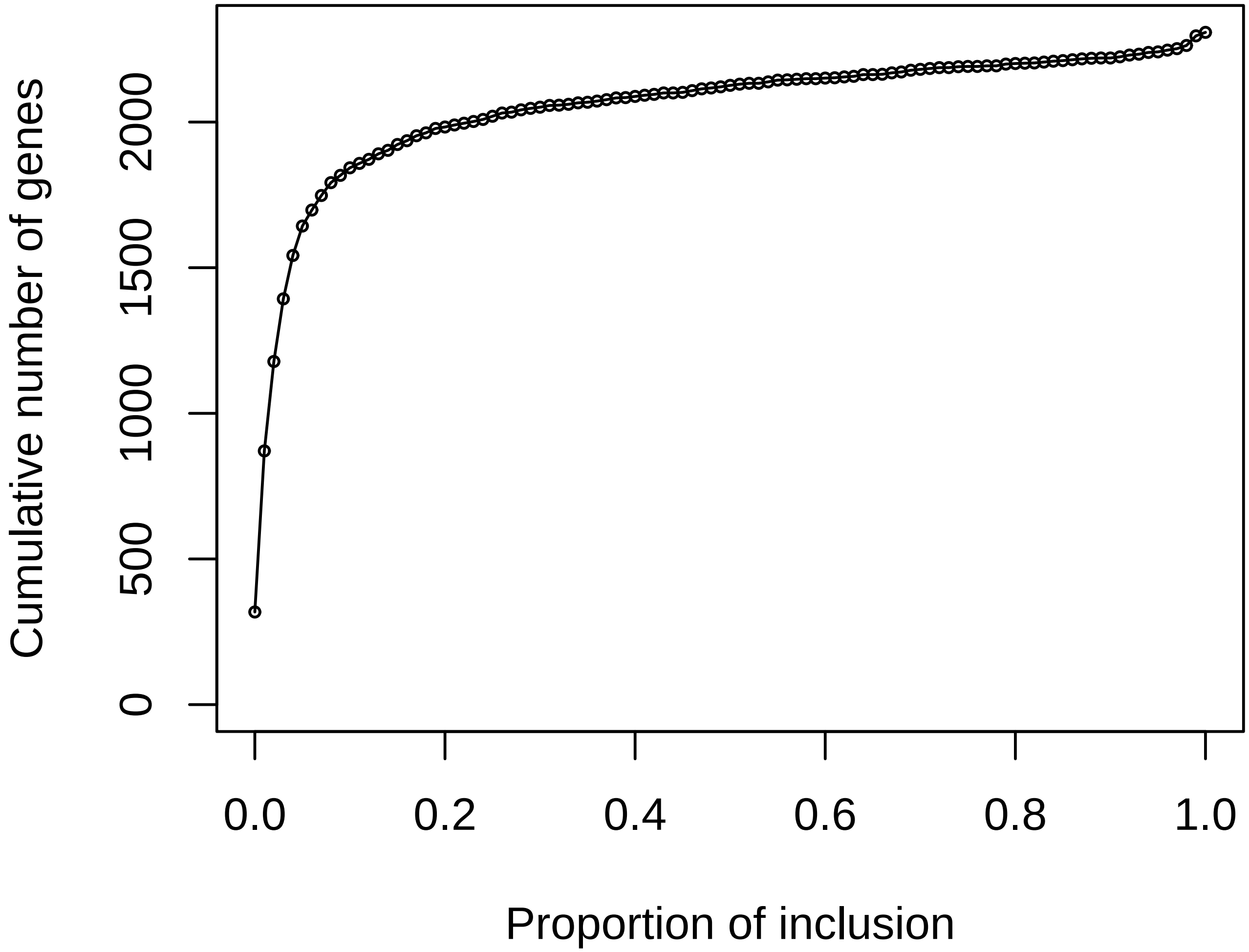

Figure: The number of genes selected less often than or as frequently as a given proportion in 100 runs.

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 SVM using ℓ_1 type penalty for general case.
- Enhance interpretation without compromising prediction accuracy.
- Construct the entire solution path of c-step regularization via the optimality conditions.
- Further streamline the c-step fitting process by early stopping and basis thinning.
- Characterize the solution path of θ-step for effective computation and tuning.

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.