The Solution Path of Multicategory Support Vector Machines

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

Joint work with Zhenhuan Cui

Method of Regularization

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

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

- $\mathcal{F}$: a class of candidate functions
- $J(f)$: complexity of the model f
- $\lambda > 0$: a regularization parameter
- Solution path: $\lambda \to \hat{f}_\lambda$

Classification

- A training data set $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$
- $\boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$
- $y \in \mathcal{Y} = \{1, \ldots, k\}$
- Learn a rule $\phi : \mathbb{R}^p \to \mathcal{Y}$ from the training data, which can be generalized to novel cases.

Support Vector Machine

Vapnik (1995)

- $y_i \in \{-1, 1\}$ for $k = 2$
- Find $f(\boldsymbol{x}) = b + h(\boldsymbol{x})$ with $h \in \mathcal{H}_K$ minimizing

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

  where K is the reproducing kernel of $\mathcal{H}_K$, and λ is a regularization parameter.
- Classification rule: $\phi(\boldsymbol{x}) = sign\ [f(\boldsymbol{x})]$

Hinge loss

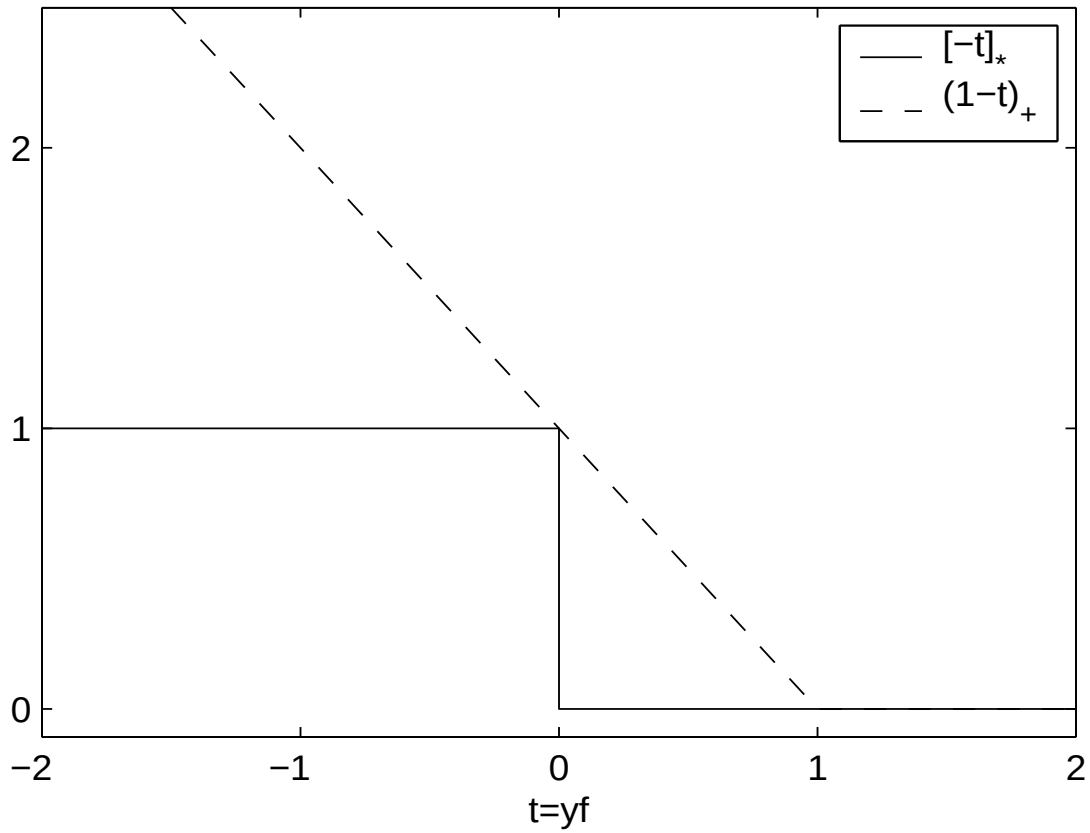

Figure: $(1 - yf(\boldsymbol{x}))_+$ is an upper bound of the misclassification loss function $I(y \neq \phi(\boldsymbol{x})) = [-yf(\boldsymbol{x})]_* \leq (1 - yf(\boldsymbol{x}))_+$ where $[t]_* = I(t \geq 0)$ and $(t)_+ = \max\{t, 0\}$.

SVM computation

- Quadratic programming problem
- Getting solutions for a fixed λ:

  e.g. Sequential Minimal Optimization [Platt (1999)], SVM light [Joachims (1999)], LIBSVM [Hsu and Lin (2002)], http://www.kernel-machines.org/software.html

- Getting the entire regularization path ranging from the simplest majority rule to the complete overfit to data as λ decreases.

  SVM solution path [Hastie et al. (2004)]

SVM when $k > 2$

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

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

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

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

MSVM hinge loss

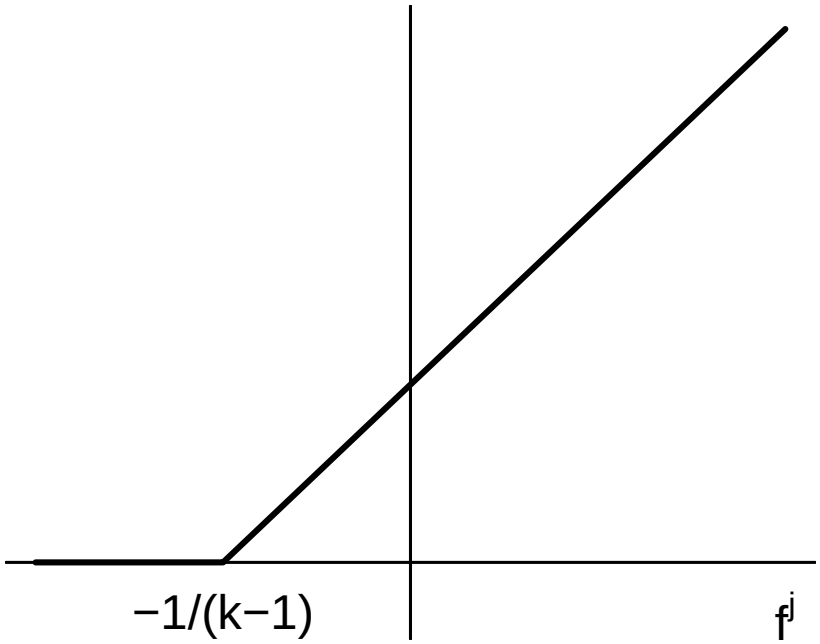

Figure: MSVM component loss $(f^j - y^j)_+$ for $y \neq j$, where $y^j = -1/(k-1)$.

Optimization problem for MSVM

- By the representer theorem, $\hat{f}^j(\boldsymbol{x}) = b^j + \sum_{i=1}^n c_i^j K(\boldsymbol{x}_i, \boldsymbol{x})$
- $\boldsymbol{c}^j = (c_1^j, \ldots, c_n^j)^\top$, $K_n = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))$ and $f_i^j = \hat{f}^j(\boldsymbol{x}_i)$
- Primal problem: minimize

$$L_P(\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{\xi}) = \frac{1}{n} \sum_{i=1}^n \sum_{j \neq y_i} \xi_i^j + \frac{\lambda}{2} \sum_{j=1}^k (\boldsymbol{c}^j)^\top K_n \boldsymbol{c}^j$$

subject to

$$\begin{aligned} f_i^j - y_i^j &\leq \xi_i^j && \text{for } i, j \\ \xi_i^j &\geq 0 && \text{for } i, j \\ \textstyle\sum_{j=1}^k (b^j \boldsymbol{e} + K_n \boldsymbol{c}^j) &= 0. \end{aligned}$$

Optimization problem for MSVM (cont'd)

- α_i^j: Lagrange multiplier for $f_i^j - y_i^j \leq \xi_i^j$
  Set $\alpha_i^j = 0$ for $j = y_i$.
- $\alpha^j = (\alpha_1^j, \ldots, \alpha_n^j)^\top$ and $\boldsymbol{y}^j = (y_1^j, \ldots, y_n^j)^\top$
- Dual problem: maximize

$$L_D(\boldsymbol{\alpha}) = -\frac{1}{2n}\sum_{j=1}^{k}(\boldsymbol{\alpha}^j - \bar{\boldsymbol{\alpha}})^\top K_n(\boldsymbol{\alpha}^j - \bar{\boldsymbol{\alpha}}) - \lambda\sum_{j=1}^{k}(\boldsymbol{\alpha}^j)^\top \boldsymbol{y}^j$$

$$\text{subject to} \qquad 0 \leq \alpha_i^j \leq 1 \qquad \text{for } i, j$$
$$(\boldsymbol{\alpha}^j - \bar{\boldsymbol{\alpha}})^\top \boldsymbol{e} = 0 \quad \text{for } j = 1, \ldots, k.$$

Optimality conditions

- By the Karush-Kuhn-Tucker (KKT) complementarity conditions, the MSVM solution at λ satisfies

$$\alpha_i^j \left(f_i^j - y_i^j - \xi_i^j\right) = 0,$$
$$\left(1 - \alpha_i^j\right) \xi_i^j = 0.$$

  Also, $0 \leq \alpha_i^j \leq 1$ and $\xi_i^j \geq 0$.
- Categorize (i, j) into three sets depending on $f_i^j - y_i^j$.

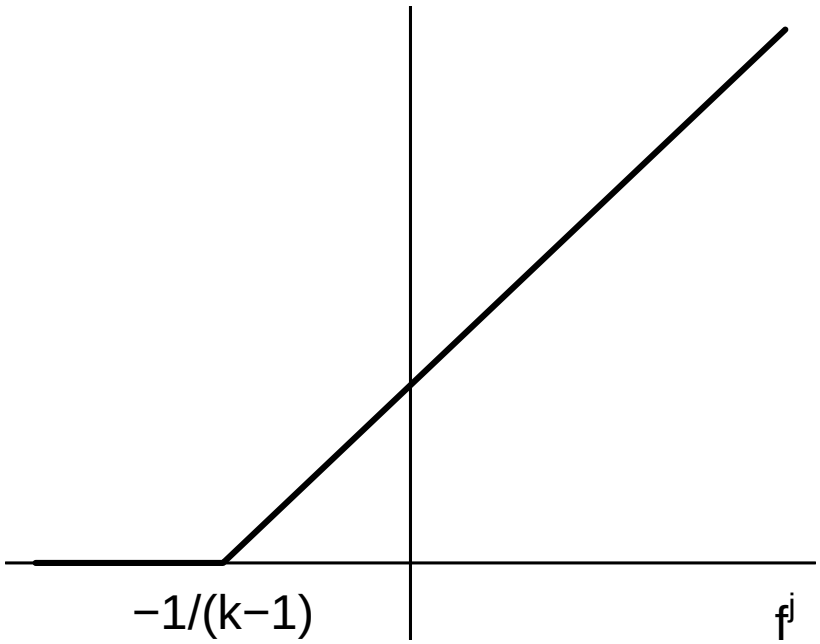

Figure: MSVM component loss $(f^j - y^j)_+$ where $y^j = -1/(k-1)$.

$\mathcal{E} = \{(i,j) |\ f_i^j - y_i^j = 0,\ \xi_i^j = 0,\ 0 \leq \alpha_i^j \leq 1\}$, Elbow set
$\mathcal{U} = \{(i,j) |\ f_i^j - y_i^j > 0,\ \xi_i^j > 0,\ \alpha_i^j = 1\}$, 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

- Initialize α_i^j's for sufficiently large λ so that they correspond to the majority rule.
- 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.
- Construct the path sequentially by solving a system of linear equations for α_i^j with $(i, j) \in \mathcal{E}_\ell$.

Piecewise linearity of the SVM solution

- The coefficient c_i^j path of the MSVM is linear in $1/\lambda$ on the interval $(\lambda_{\ell+1}, \lambda_\ell)$.
- The path of b^j is also linear in $1/\lambda$ on the interval $(\lambda_{\ell+1}, \lambda_\ell)$ if there is at most one empty elbow set $\mathcal{E}_\ell^j$ at λ_ℓ. Otherwise, the path of b^j for the empty $\mathcal{E}_\ell^j$ can be arbitrary.

Generating the joints $\{\lambda_\ell\}$

- Given λ_ℓ, consider the following possible events.
  1. An index (i, j) in $\mathcal{E}_\ell$ leaves the elbow set, and α_i^j $(0 \leq \alpha_i^j \leq 1)$ becomes either 0 or 1.
  2. An index (i, j) in $\mathcal{L}_\ell$ or $\mathcal{U}_\ell$ joins the elbow set, and $\hat{f}^j(\boldsymbol{x}_i)$ is then y_i^j.
- The two potential events propose candidate values of $\lambda_{\ell+1}$.
- The next break point $\lambda_{\ell+1}$ is determined by the largest $\lambda < \lambda_\ell$ among the potential values.

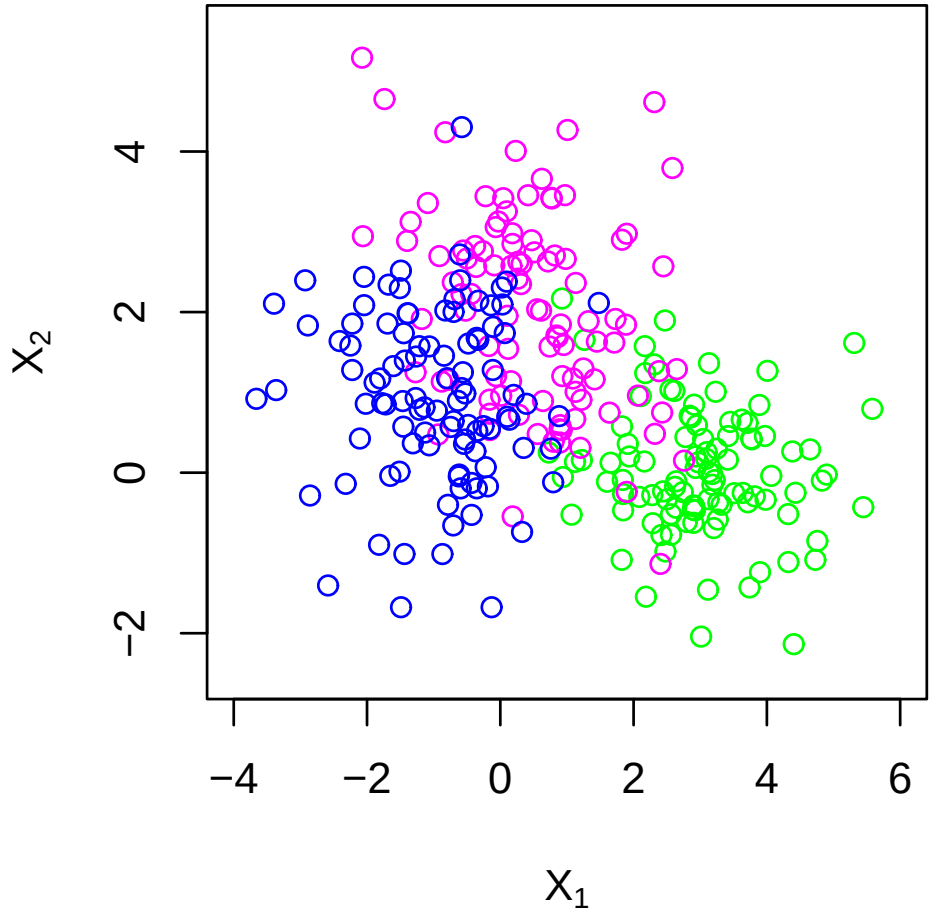

Figure: A scatter plot of the training data. Class 1: green, Class 2: magenta, and Class 3: blue.

Test error rates

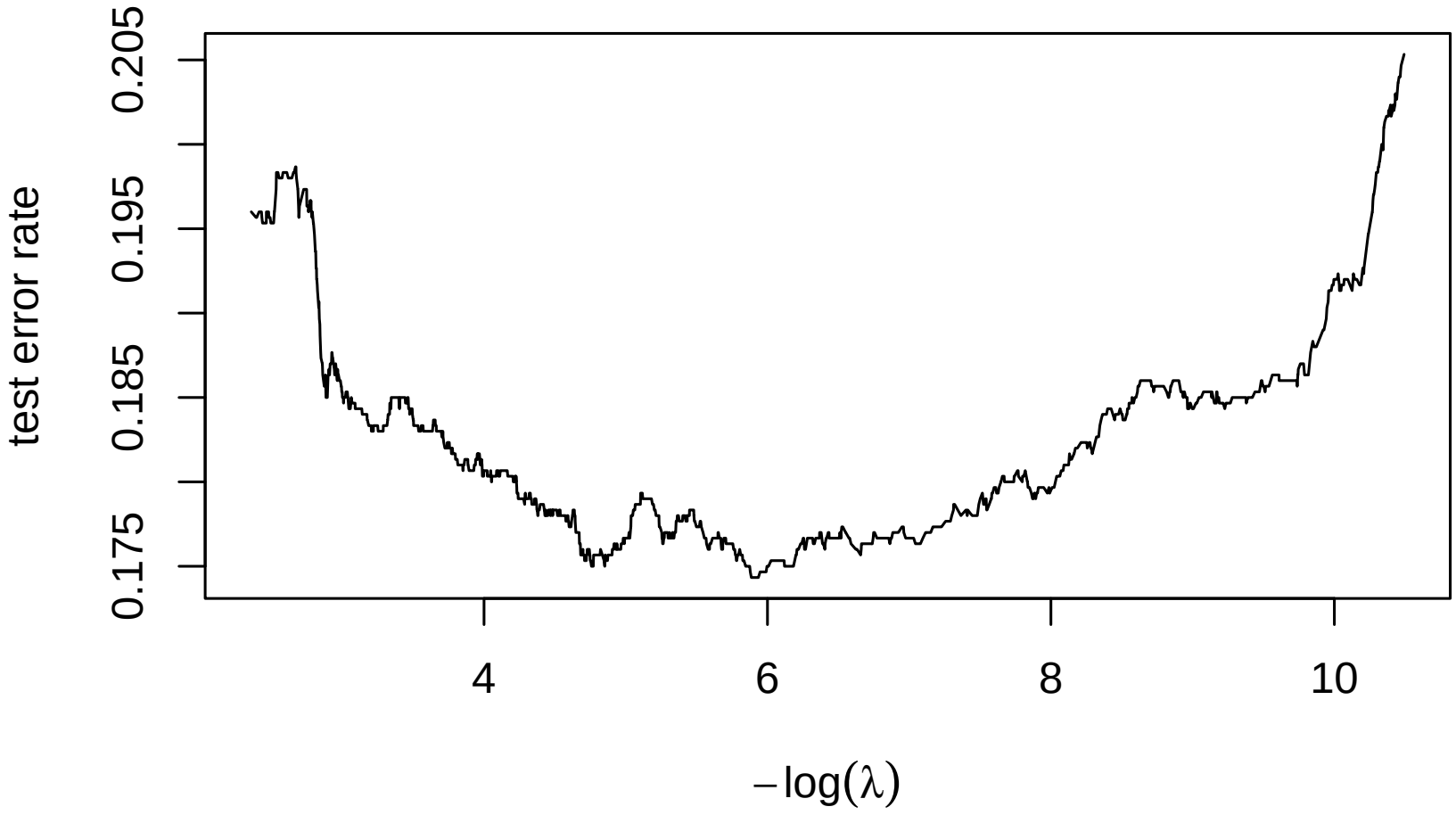

Figure: Test error rate as a function of λ.

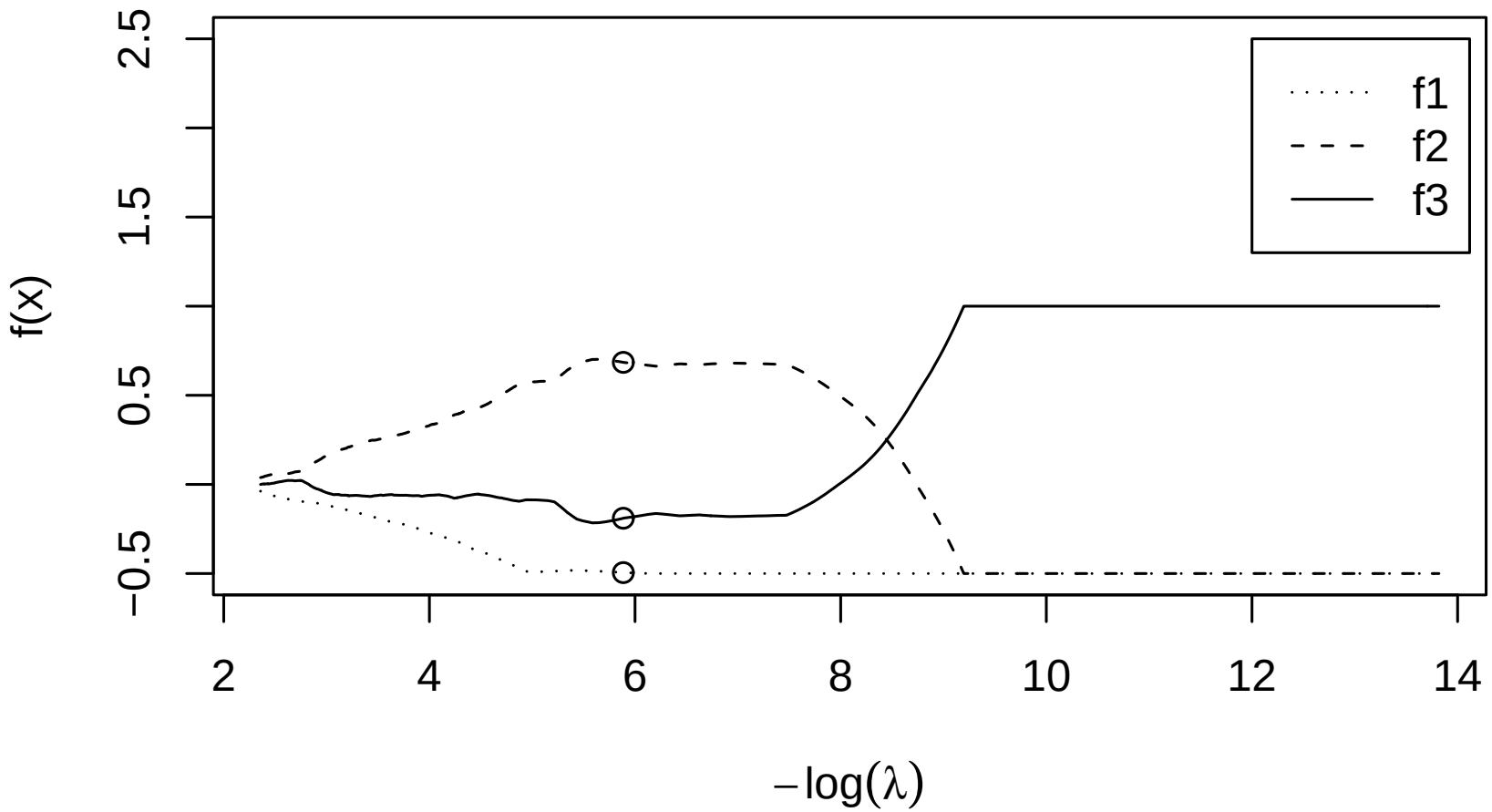

Figure: The entire paths of $\hat{f}^1_\lambda(\boldsymbol{x}_i)$, $\hat{f}^2_\lambda(\boldsymbol{x}_i)$, and $\hat{f}^3_\lambda(\boldsymbol{x}_i)$ for an outlying instance $\boldsymbol{x}_i$ from class 3. The circles correspond to λ with the minimum test error rate.

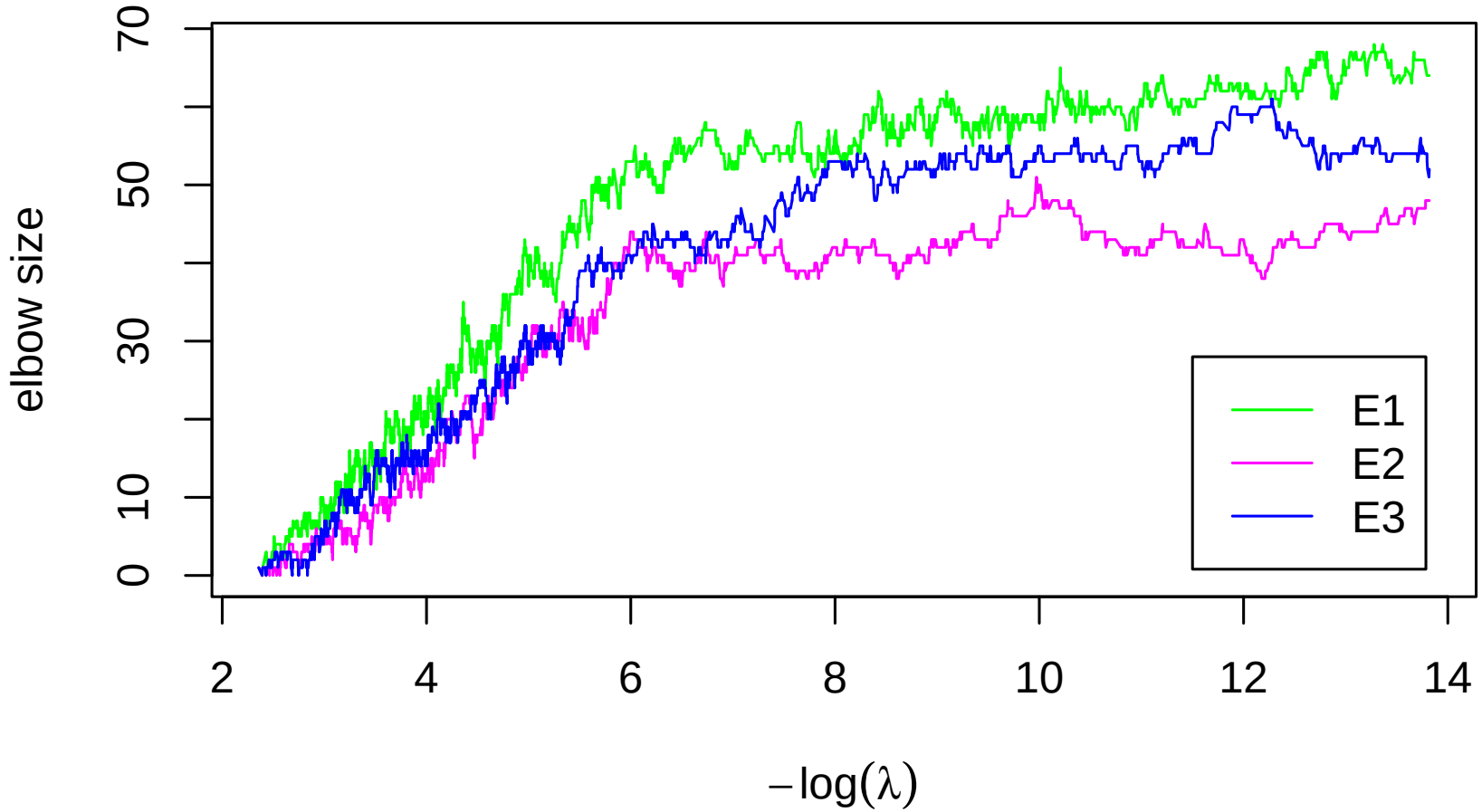

Figure: The size of elbow set $\mathcal{E}_\ell^j$ for three classes as a function λ.

Further generalization

- The path finding algorithm can be extended to the case with unequal misclassification costs ($\boldsymbol{L}(\boldsymbol{y}) = (L_y^1, \ldots, L_y^k)$) or different data weights (w_i).

- $\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{y}_i) = w_i \sum_{j \neq y_i} L_{y_i}^j (f_i^j - y_i^j)_+$
- $0 \leq \alpha_i^j \leq w_i L_{y_i}^j$

  Complementarity conditions:

$$\alpha_i^j \, (f_i^j - y_i^j - \xi_i^j) = 0 \text{ and } (w_i L_{y_i}^j - \alpha_i^j) \, \xi_i^j = 0$$

- Initialize α_i^j's for sufficiently large λ so that they correspond to the rule minimizing the total weighted costs $\sum_{i=1, y_i \neq j}^n w_i L_{y_i}^j$.
- Piecewise linearity of the coefficient path remains true.

Computational complexity

- Proportional to the number of break points λ_ℓ.
- At each λ_ℓ, it takes $O(|\mathcal{E}_\ell|^3 + (k-1)n|\mathcal{E}_\ell^{\mathcal{I}}|)$ operations (a system of linear equations needs to be solved, and the next break point has to be determined.)
- $|\mathcal{E}_\ell|$, intermediate function evaluations, and the number of break points tend to be proportional to $(k-1)$.
- $\mathcal{E}_\ell$ may grow too big for large data sets and flexible kernel functions.

Scaling down data sets

- Basis thinning:
  Recall $\hat{f}^j(\boldsymbol{x}) = b^j + \sum_{i=1}^n c_i^j K(\boldsymbol{x}_i, \boldsymbol{x})$.

  Choose $\mathcal{I}^* \subset \{1, \ldots, n\}$ with $|\mathcal{I}^*| \ll n$ such that
  $span\{K(\boldsymbol{x}_{i^*}, \cdot), i^* \in \mathcal{I}^*\} \approx span\{K(\boldsymbol{x}_i, \cdot), i = 1, \ldots, n\}$.
  - Random sampling
  - Statistically equivalent blocks [Wilks (1962)]
  - Cluster analysis: $\{(\tilde{\boldsymbol{x}}_{i^*}, \tilde{y}_{i^*})$ with weight $w_{i^*}\}$
- Data squashing [Dumouchel et al. (1999)]:
  Binning and generating pseudo data with weights matching the moments of the full data.

Statistically equivalent blocks

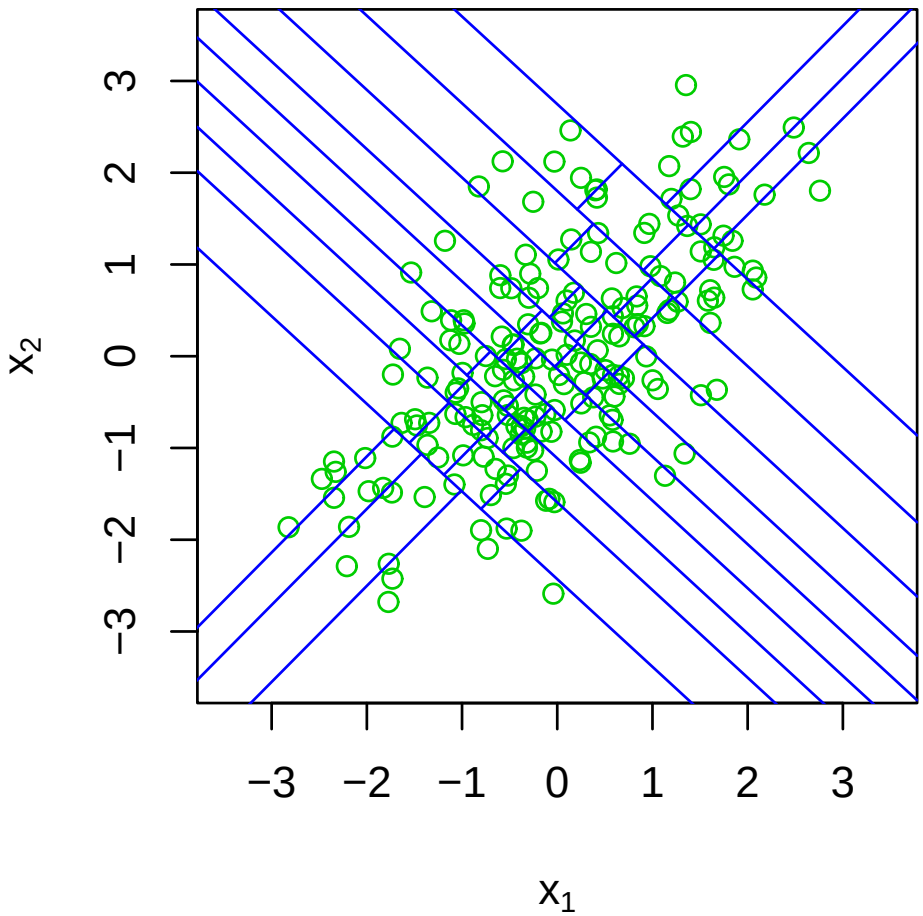

Figure: $n = 200$, 40 (10×4) statistically equivalent blocks obtained by using the principal components as the ordering functions.

Clusters

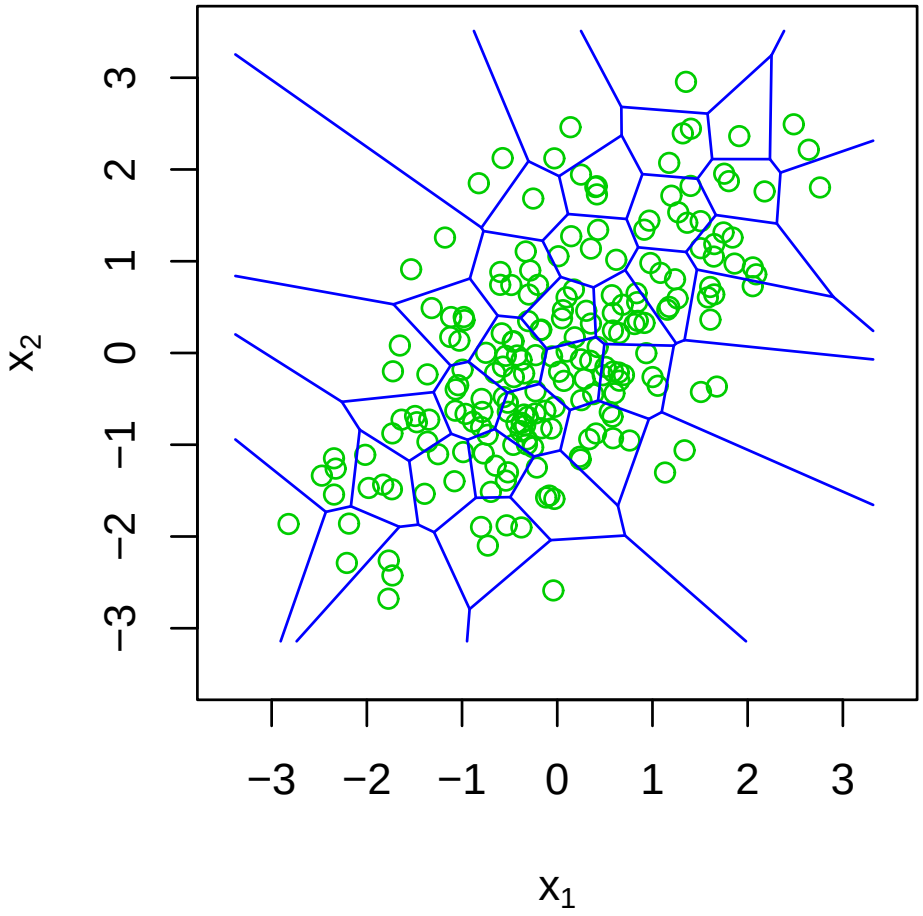

Figure: $n = 200$, 40 clusters formed by the k-means algorithm and the Voronoi tessellation of the cluster centers.

Finding an approximate solution path

- Scale down the data by choosing a representative subsample for each class.
- Find the regularized solution path for the reduced data.
- “Out-of-bag sample” as in bootstrap can be used for tuning.

Comparison of data reduction methods

- The three-class problem with sample size 120×3 and reduction factor 6 (20×3)
- Number of replicates: 100
- Test error rates over 1200×3 cases (the Bayes rate ≈ 0.1773)

Method	Random sampling	SEB	Cluster
Mean	0.19345	0.18307	0.18216
SD	0.01452	0.00768	0.00720

- The quality of approximate solutions would depend on reduction factor, data partitioning scheme, dimensionality, and the Bayes rate.

Summary

- Characterize the solution path of MSVM for effective computation and tuning.
- Direct use of the complementarity conditions and the piecewise linearity of the coefficient path.
- Find an approximate solution path to cope with large data sets.
- Data geometry (manifold learning, nonlinear PCA) can guide data reduction procedures.