Introduction to Machine Learning
Topic 9: Ensemble methods and high-dimensional problems
Lecture 9a - High-dimensional problems

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Topic overview

- **Wide data:** $N \ll p$
- **High-dimensional classification**
- **High-dimensional regression**

Wide data $N \ll p$

- **Wide data** $N \ll p$. Many variables, few data points.
  - **Genomics**
  - **Text**
- **Tall data**: $p \ll N$. Few variables, many data points.
- Tall and Wide. Supermarket scanners. Many purchases, many products.

GENOMICS - MICROARRAYS

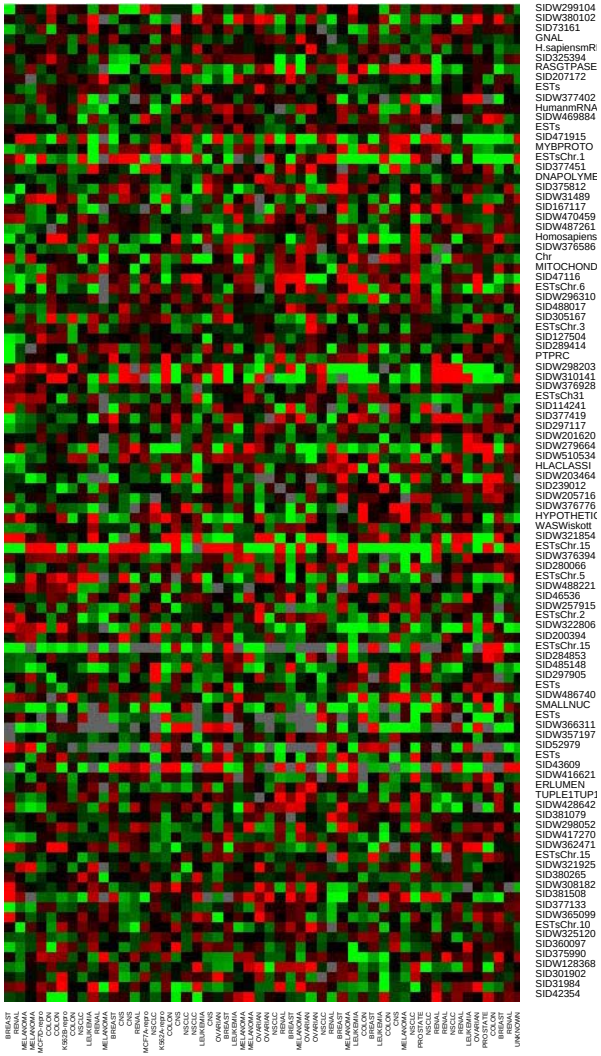


FIGURE 1.3. *DNA microarray data: expression matrix of* 6830 *genes (rows) and* 64 *samples (columns), for the human tumor data. Only a random sample of* 100 *rows are shown. The display is a heat map, ranging from bright green (negative, under expressed) to bright red (positive, over expressed). Missing values are gray. The rows and columns are displayed in a randomly chosen order.*

TEXT - DOCUMENT CLASSIFICATION

Document	has('ball')	has('EU')	has('political_arena')	wordlen	Lex. Div.	Topic
Article1	Yes	No	No	4.1	5.4	Sports
Article2	No	No	No	6.5	13.4	Sports
⋮	⋮	⋮		⋮	⋮	⋮
ArticleN	No	No	Yes	7.4	11.1	News

THE TROUBLE WITH WIDE DATA $N \ll p$

- **Linear regression** with p covariates

$$\underset{N\times 1}{\mathbf{y}} = \underset{N\times p}{\mathbf{X}}\ \underset{p\times 1}{\beta} + \underset{N\times 1}{\varepsilon}$$

,

- Least squares/**Maximum likelihood**: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- Problem: $\mathbf{X}'\mathbf{X}$ is $p \times p$ but of rank N. **Not invertible** when $N < p$.
- Too many parameters to estimate from few data points.
- **Solutions**: regularization:
  - **Dimensionality reduction**. Principal components regression.
  - **Shrinkage**. L_2-penalty (Ridge) or L_1-penalty (Lasso)
  - **Variable selection**. Forward selection. Bayesian.

Regression L2-shrinkage when $N \ll p$

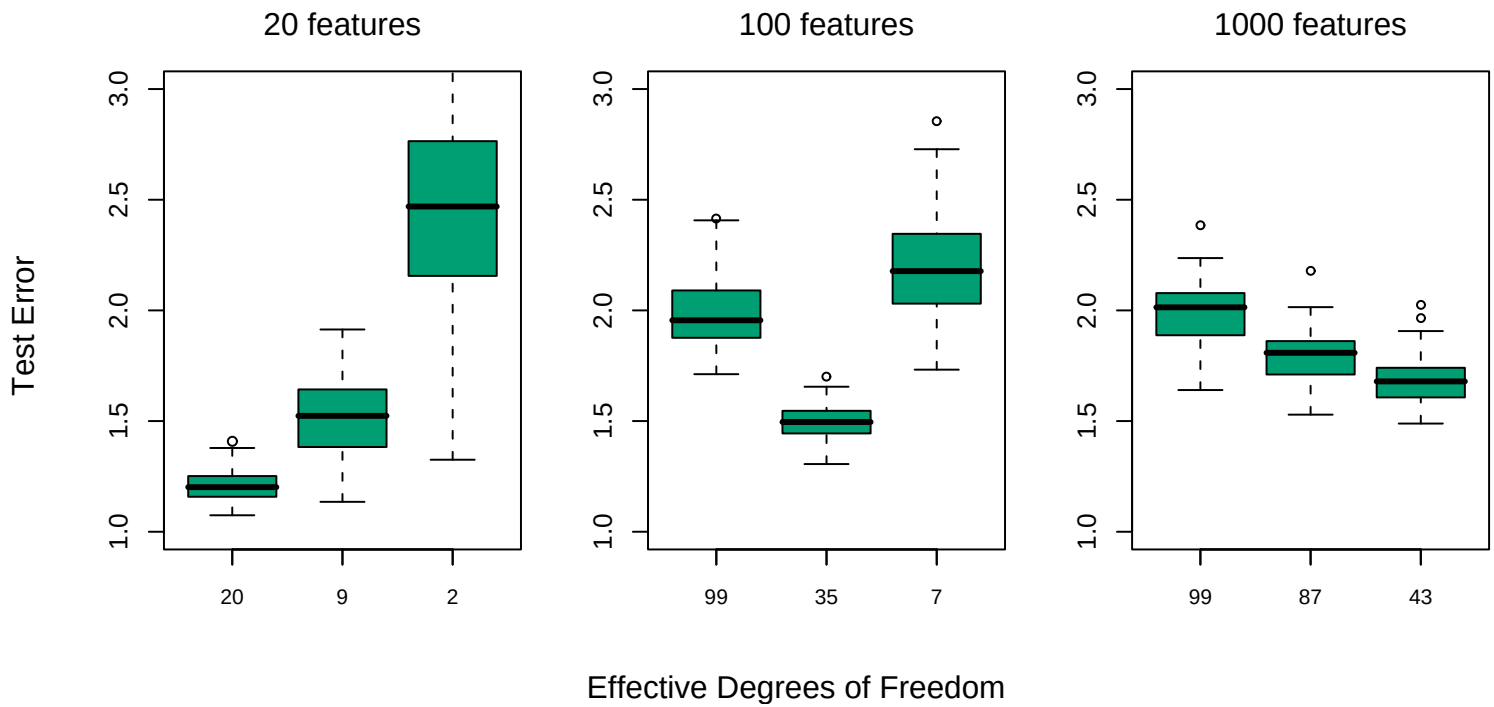


FIGURE 18.1. *Test-error results for simulation experiments. Shown are box-plots of the relative test errors over* 100 *simulations, for three different values of* p*, the number of features. The relative error is the test error divided by the Bayes error,* σ^2*. From left to right, results are shown for ridge regression with three different values of the regularization parameter* λ*:* 0.001, 100 *and* 1000*. The (average) effective degrees of freedom in the fit is indicated below each plot.*

DIAGONAL COVARIANCE LDA FOR $N \ll p$ PROBLEMS

- LDA requires the estimation of class conditional distributions $N(\mu_1, \Sigma), ..., N(\mu_2, \Sigma)$.
- LDA makes the simplifying assumption of having the **same covariance matrix** Σ in each class ...
- ... but Σ is $p \times p$ and symmetric. $p(p+1)/2$ elements. Impossible if when $N \ll p$.
- **Diagonal covariance LDA**: implying assumption: $\Sigma = \text{Diag}(\sigma_1^2, ..., \sigma_p^2)$. Special case of Naive Bayes.
- Discriminant score for Diag LDA:

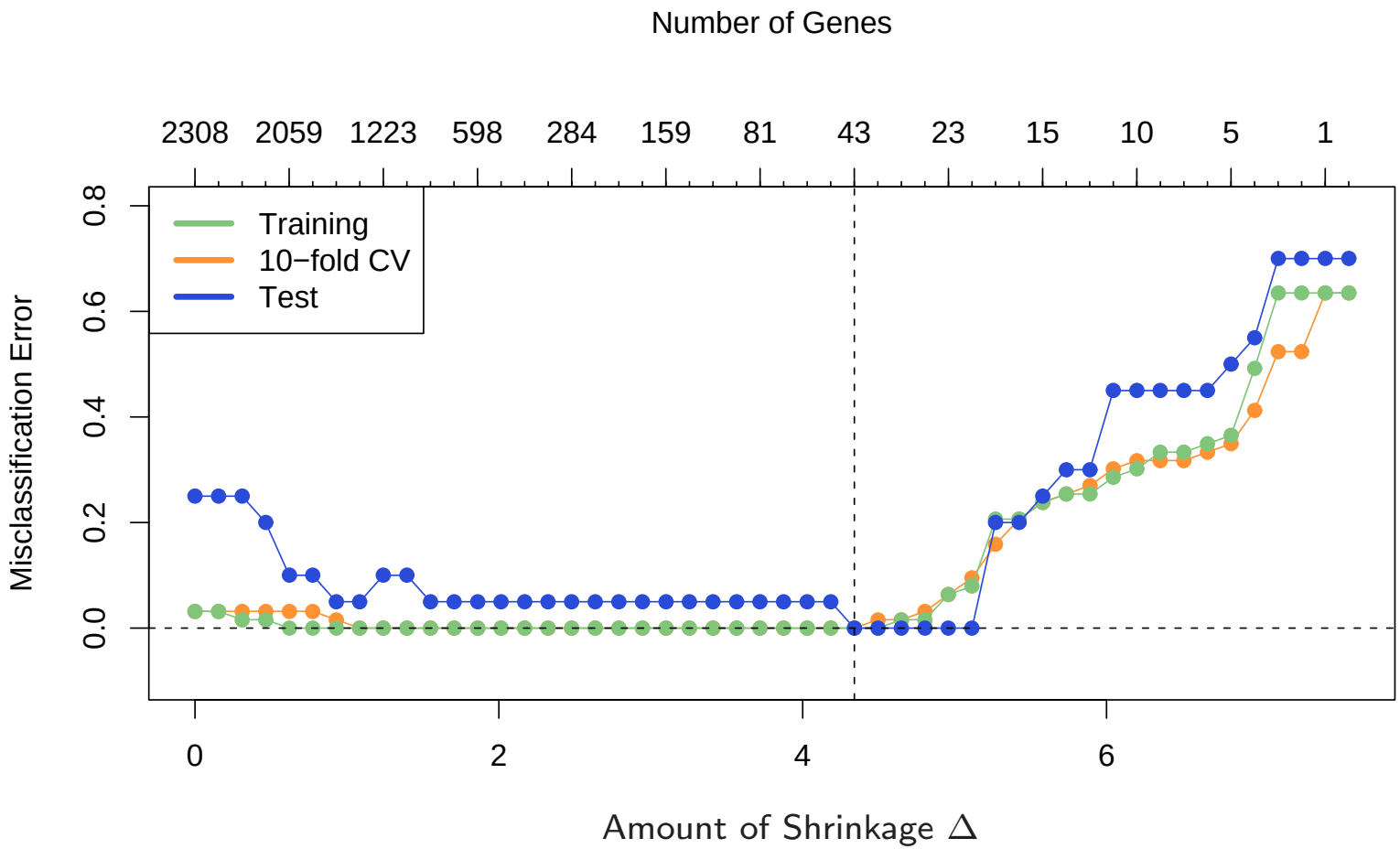
$$\delta_k(\mathbf{x}^\star) = -\sum_{j=1}^{p} \frac{\left(x_j^\star - \bar{x}_{kj}\right)^2}{s_j^2} + 2\log \pi_k$$

- Classification rule: assign to class with highest score: $C(\mathbf{x}^\star) = \ell$ if $\delta_\ell(\mathbf{x}^\star) = \max_k \delta_k(\mathbf{x}^\star)$.

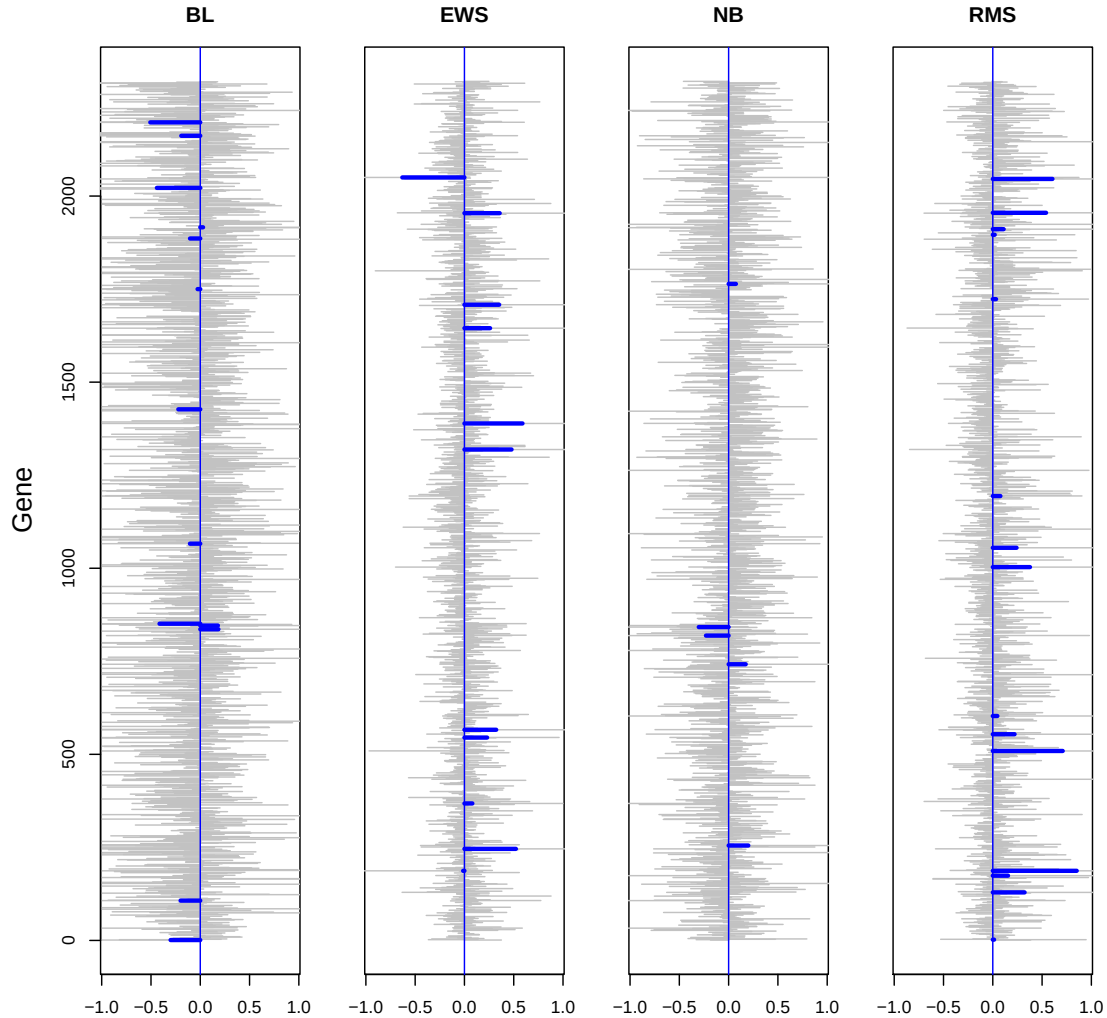
Nearest shrunken centroids (NSC)

- Weakness of Diagonal covariance LDA: it uses all features in the classification rule. Doesn't tell us which features are useful for the classification. Interpretation.
- **Nearest Shrunken Centroids** (**NSC**) shrinks the class means $\bar{x}_{kj}$ for feature j toward global mean $\bar{x}_j$.
- Basic idea: small deviations from global mean (normalized $|\bar{x}_{kj} - \bar{x}_j| < \Delta$) are set to zero.
- Some features are set equal to global mean for all classes $\rightarrow$ that feature does not have classification power.
- **Hard or smooth threshold**. Δ is a **regularization** parameter. Larger Δ gives simpler model (more features drop out of the model).

Nearest shrunken centroids (NSC)



Nearest shrunken centroids (NSC)



Centroids: Average Expression Centered at Overall Centroid

Nearest shrunken centroids (NSC)

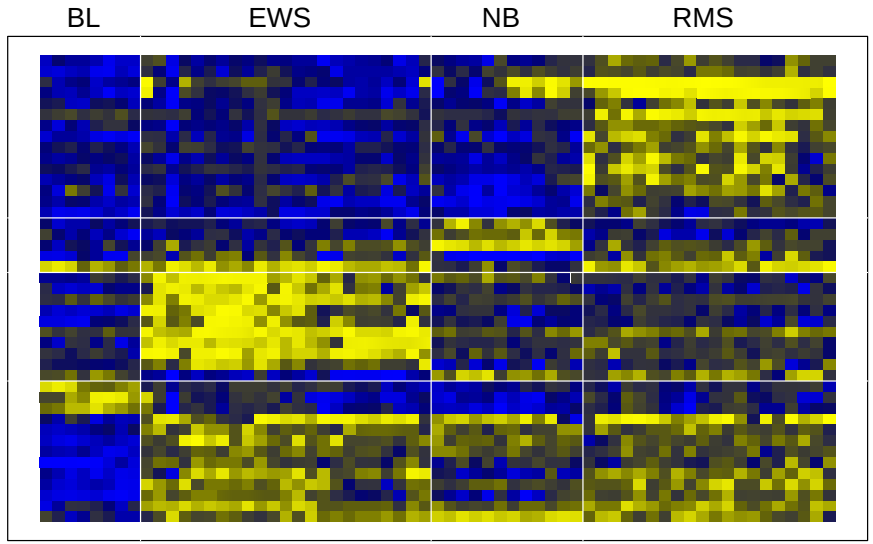


FIGURE 18.3. *Heat-map of the chosen 43 genes. Within each of the horizontal partitions, we have ordered the genes by hierarchical clustering, and similarly for the samples within each vertical partition. Yellow represents over- and blue under-expression.*

Regularized Discriminant Analysis (RDA)

- Estimate full $p \times p$ covariance matrix $\hat{\Sigma}$, but **shrink toward diagonal** covariance matrix:

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\text{Diag}(\hat{\Sigma})$$

  where $0 \leq \gamma \leq 1$ is the **shrinkage parameter**.

- Note that $\hat{\Sigma}$ has rank $N < p$, so shrinkage toward diagonal is essential.

- RDA is like **ridge regression** (where $\mathbf{X}'\mathbf{X}$ is shrunk toward the identity matrix).

Discriminative classification models, $N \ll p$

- **Multi-class logistic regression**

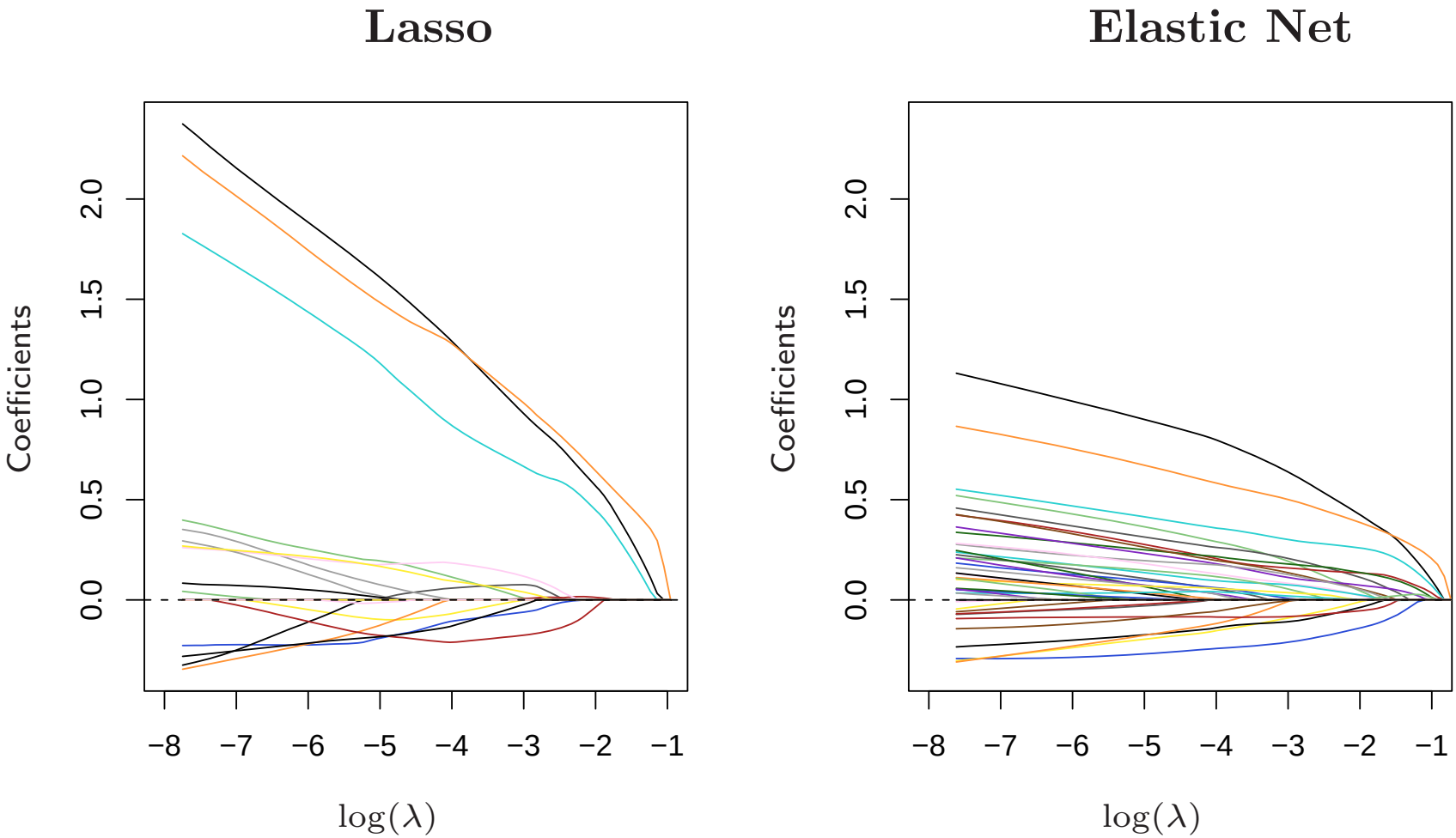
$$\Pr(G = k|\mathbf{x}) = \frac{\exp\left(\beta_{k0} + \mathbf{x}'\beta_k\right)}{\sum_{\ell=1}^{K} \exp\left(\beta_{\ell 0} + \mathbf{x}'\beta_\ell\right)}$$

- Regularization (L2):

$$\max_{\{\beta_{0k},\beta_k\}_{k=1}^{K}} \left[\Pr\left(g_i|x_i\right) - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} \beta_{kj}^2 \right]$$

- Similar regularization can be done with **SVMs** and other models.
- L1=**Lasso** regularization is another option. Penalty: $\lambda \sum_{k=1}^{K} \sum_{j=1}^{p} |\beta_{kj}|$. Severe regularization since at most N selected (non-zero) features when $N < p$.
- **Elastic net** penalty: $\lambda \sum_{k=1}^{K} \sum_{j=1}^{p} \left(\alpha |\beta_{kj}| + (1-\alpha)\beta_{kj}^2\right)$. Number of selected features can be larger than N. Choose λ and α by cross-validation.

Discriminative classification models, $N \ll p$



SURVIVAL DATA WITH CENSORING

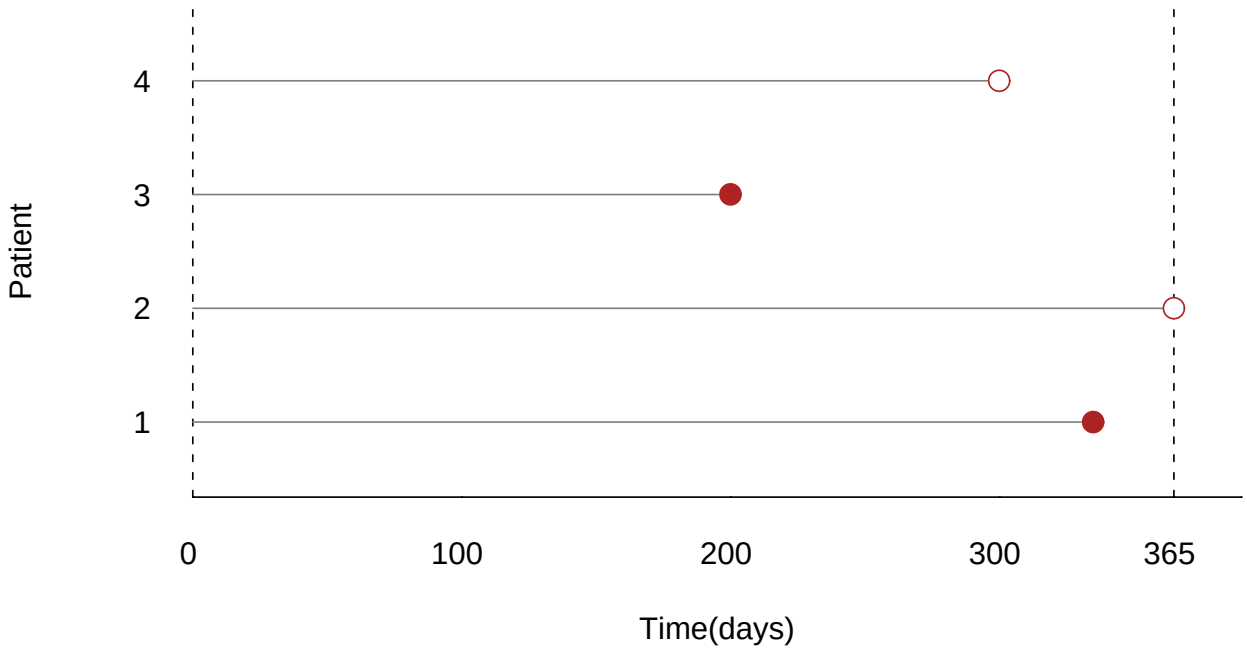
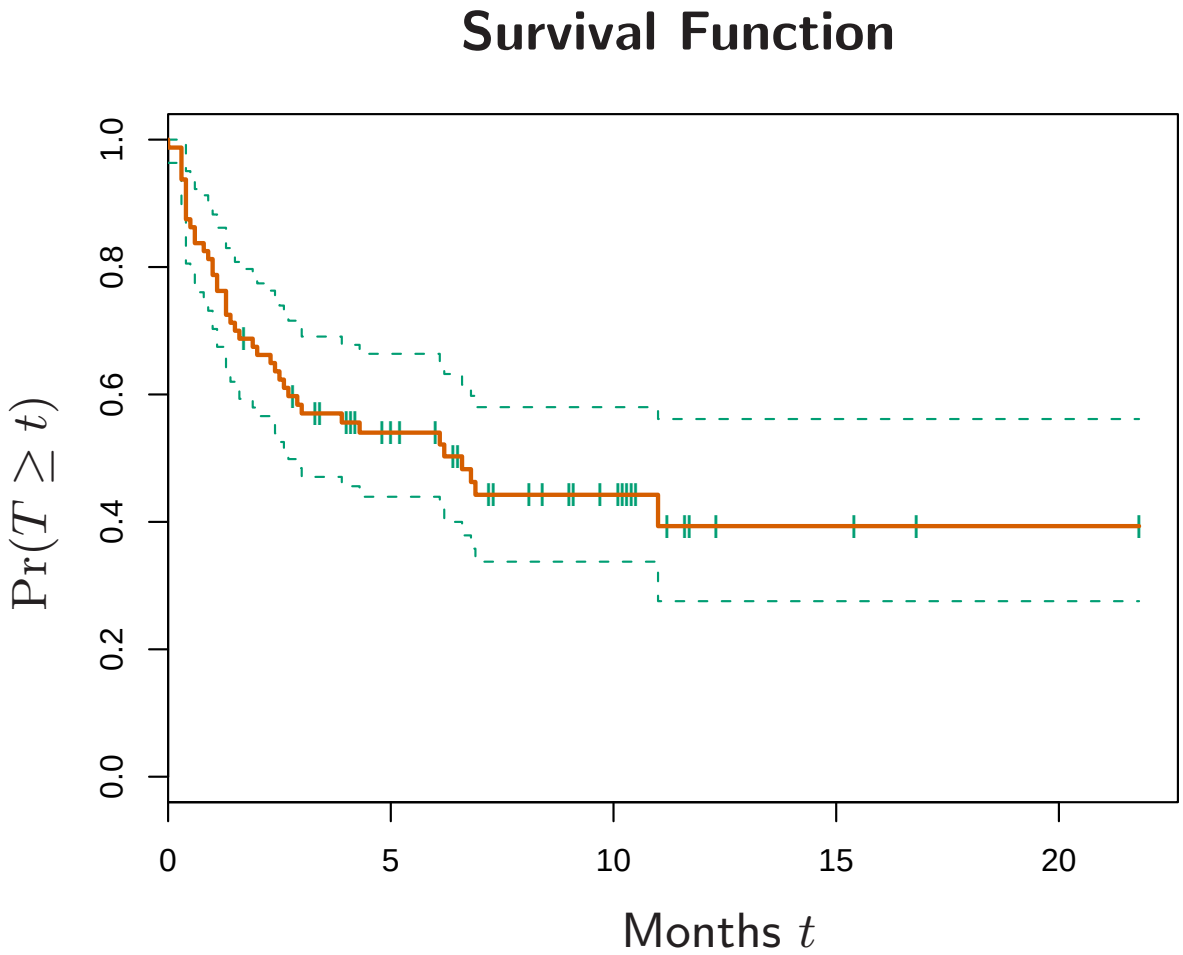


FIGURE 18.11. *Censored survival data. For illustration there are four patients. The first and third patients die before the study ends. The second patient is alive at the end of the study (365 days), while the fourth patient is lost to follow-up before the study ends. For example, this patient might have moved out of the country. The survival times for patients two and four are said to be "censored."*

KAPLAN-MEIER SURVIVAL CURVES



Principal Components regression

- Compute the **principal components** of $\mathbf{X}$ ($N \times p$) by performing the eigenvalue decomposition

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\Lambda\mathbf{V}'$$

- Let $\mathbf{Z} = \mathbf{XV}$ be the $N \times p$ matrix with **principal component scores**.
- **PC regression**: Regress $\mathbf{y}$ on the m PCs with largest eigenvalues. Choose m by cross-validation.
- Problem: PCs are designed to **capture most of the variation in $\mathbf{X}$**, but is unrelated to the linear combinations predictive power with respect to $\mathbf{y}$.
- **Supervised PC regression** and **Partial Least Squares** (**PLS**) use the correlation between $\mathbf{y}$ and the linear combinations to construct linear combinations.

Supervised Principal Components regression

Algorithm 18.1 *Supervised Principal Components.*

1. Compute the standardized univariate regression coefficients for the outcome as a function of each feature separately.

2. For each value of the threshold θ from the list $0 \leq \theta_1 < \theta_2 < \cdots < \theta_K$:

   (a) Form a reduced data matrix consisting of only those features whose univariate coefficient exceeds θ in absolute value, and compute the first m principal components of this matrix.

   (b) Use these principal components in a regression model to predict the outcome.

3. Pick θ (and m) by cross-validation.

PC REGRESSION FOR SURVIVAL ANALYSIS