Regularization and Variable Selection via the Elastic Net

Hui Zou and Trevor Hastie
Department of Statistics
Stanford University

Outline

- Variable selection problem
- Sparsity by regularization and the lasso
- The elastic net

Variable selection

- Want to build a model using a subset of "predictors"
- Multiple linear regression; logistic regression (GLM); Cox's partial likelihood, ...
  - model selection criteria: AIC, BIC, etc.
  - relatively small p (p is the number of predictors)
  - instability (Breiman, 1996)
- Modern data sets: high-dimensional modeling
  - microarrays (the number of genes $\simeq$ 10,000)
  - image processing
  - document classification
  - ...

Example: Leukemia classification

- *Leukemia Data,* Golub et al. Science 1999
- There are 38 training samples and 34 test samples with total $p = 7129$ genes.
- Record the expression for sample i and gene j.
- Tumors type: AML or ALL.
- Golub et al. used a Univariate Ranking method to select relevant genes.

The $p \gg n$ problem and grouped selection

- Microarrays: $p \simeq 10,000$ and $n < 100$. A typical "large p, small n" problem (West et al. 2001).
- For those genes sharing the same biological "pathway", the correlations among them can be high. We think of these genes as forming a group.
- What would an *"oracle"* do?
  - ✔ Variable selection should be *built into* the procedure.
  - ✔ Grouped selection: automatically include whole groups into the model if one variable amongst them is selected.

Sparsity via ℓ_1 penalization

- Wavelet shrinkage and Basis pursuit; Donoho et al. (1995)
- Lasso; Tibshirani (1996)
- Least Angle Regression (LARS); Efron, Hastie, Johnstone and Tibshirani (2004)
- COSSO in smoothing spline ANOVA; Lin and Zhang (2003)
- ℓ_0 and ℓ_1 relation; Donoho et al. (1999,2004)

Lasso

- Data $(\mathbf{X}, \mathbf{y})$. $\mathbf{X}$ is the $n \times p$ predictor matrix of standardized variables; and $\mathbf{y}$ is the response vector.

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 \quad \text{s.t. } \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \leq t$$

- Bias-variance tradeoff by a continuous shrinkage
- Variable selection by the ℓ_1 penalization
- Survival analysis: Cox's partial likelihood + the ℓ_1 penalty (Tibshirani 1998)
- Generalized linear models (e.g. logistic regression)
- LARS/Lasso: Efron et al. (2004).

The limitations of the lasso

- If $p > n$, the lasso selects at most n variables. The number of selected genes is bounded by the number of samples.
- Grouped variables: the lasso fails to do grouped selection. It tends to select one variable from a group and ignore the others.

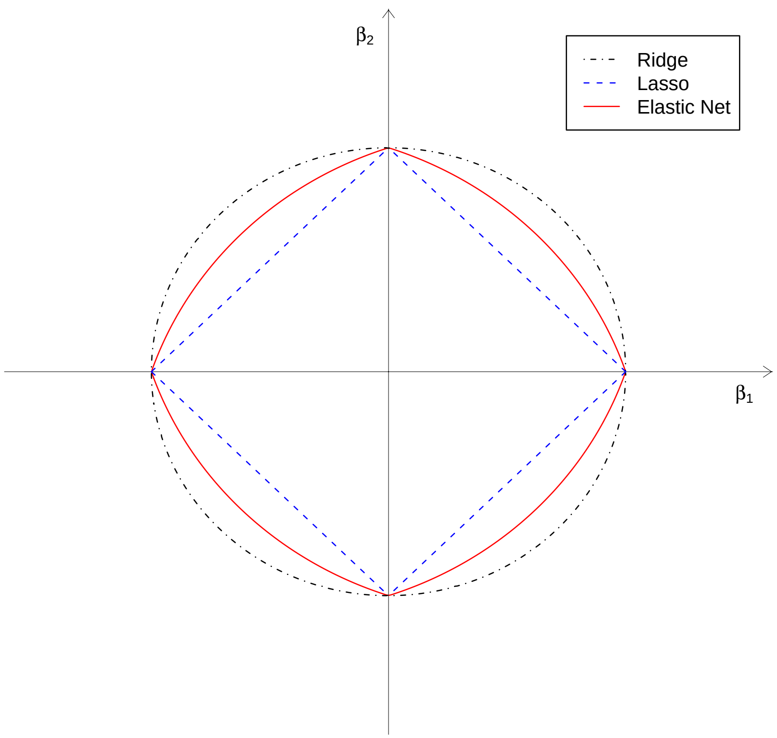
Elastic Net regularization

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1$$

- The ℓ_1 part of the penalty generates a sparse model.
- The quadratic part of the penalty
  - Removes the limitation on the number of selected variables;
  - Encourages *grouping effect*;
  - Stabilizes the ℓ_1 regularization path.

Geometry of the elastic net

2-dimensional illustration $\alpha = 0.5$



The elastic net penalty

$$J(\beta) = \alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1$$

(with $\alpha = \frac{\lambda_2}{\lambda_2+\lambda_1}$)

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 \text{ s.t. } J(\beta) \leq t.$$

- Singularities at the vertexes (necessary for sparsity)
- Strict convex edges. The strength of convexity varies with α (grouping)

A simple illustration: elastic net vs. lasso

- Two independent "hidden" factors $\mathbf{z_1}$ and $\mathbf{z_2}$

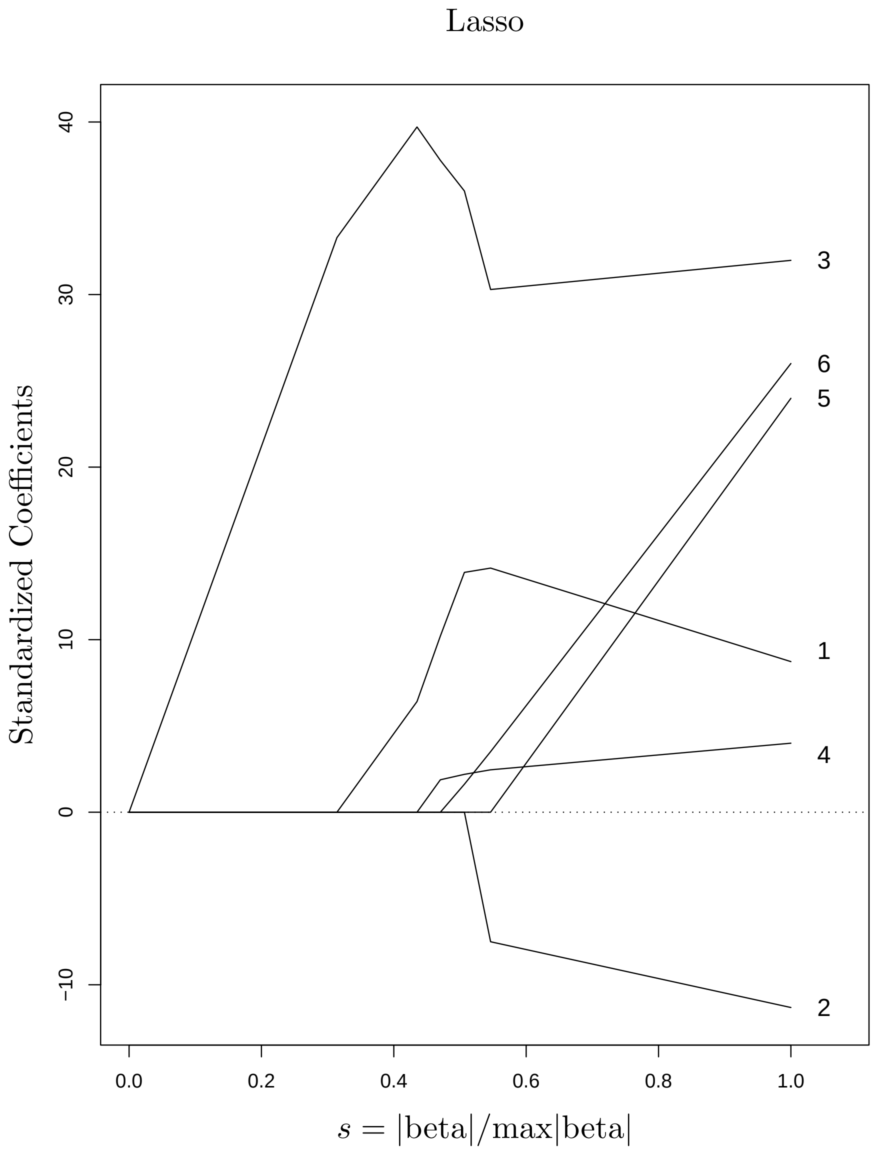
$$\mathbf{z}_1 \sim U(0,20), \quad \mathbf{z}_2 \sim U(0,20)$$

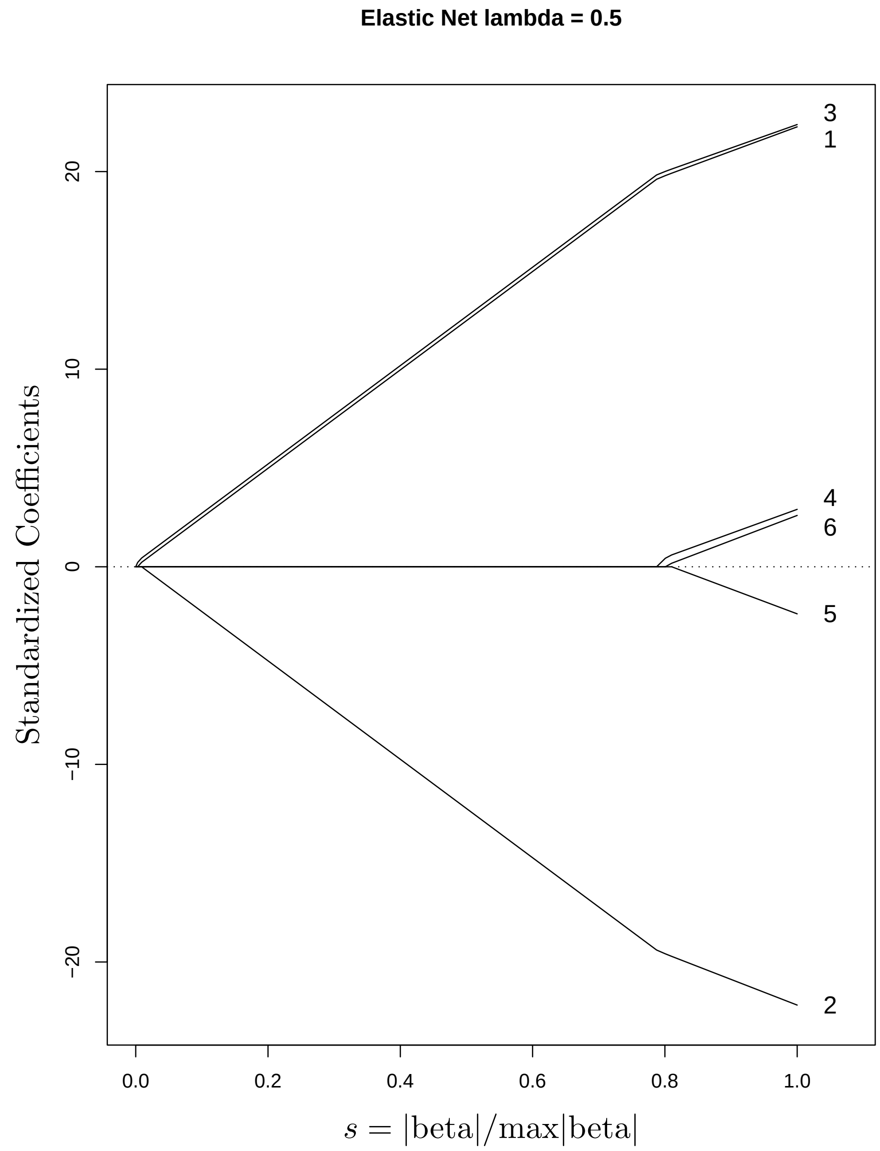
- Generate the response vector $\mathbf{y} = \mathbf{z}_1 + 0.1 \cdot \mathbf{z}_2 + N(0,1)$
- Suppose only observe predictors

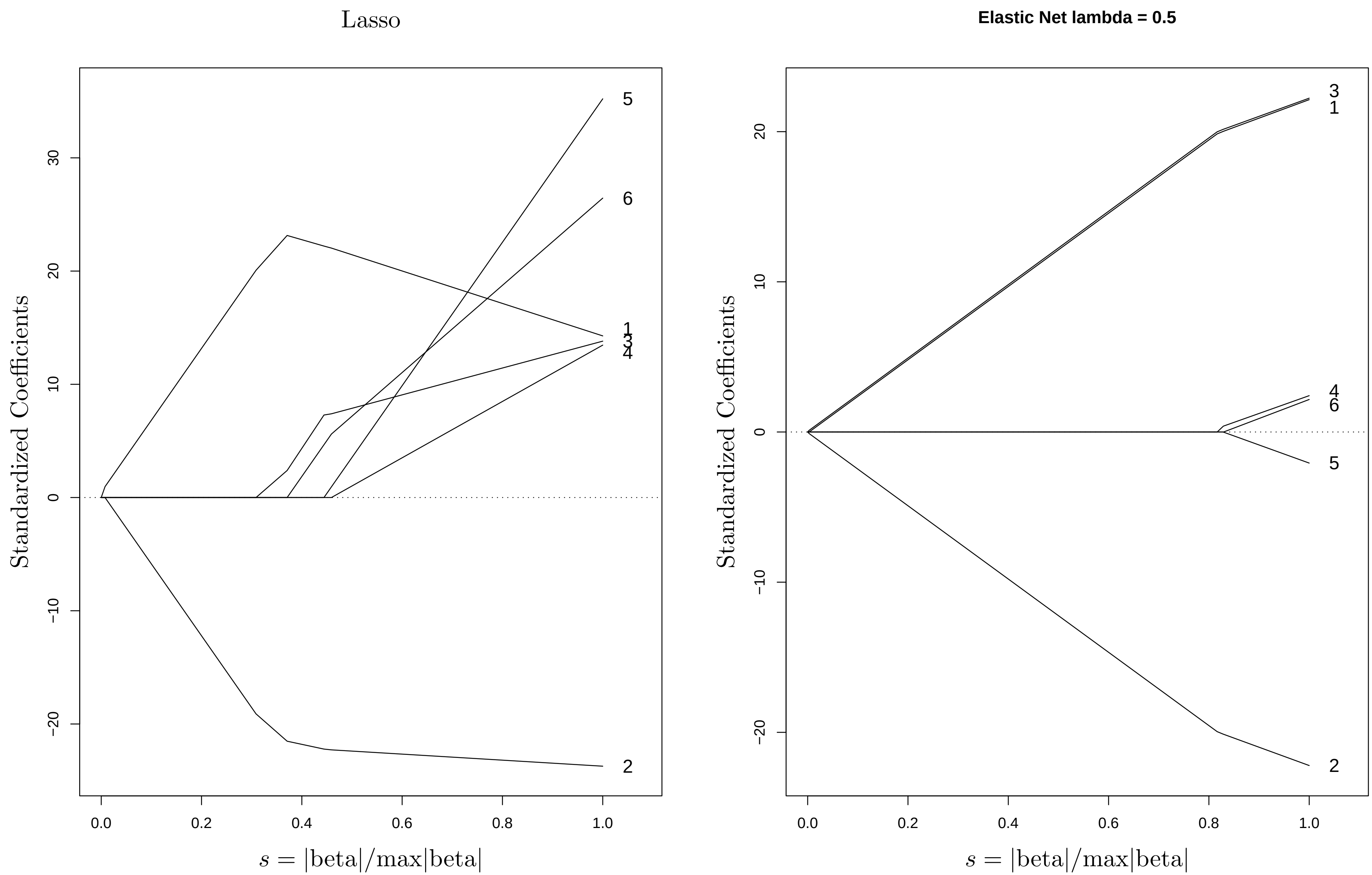
$$\mathbf{x}_1 = \mathbf{z}_1 + \epsilon_1, \quad \mathbf{x}_2 = -\mathbf{z}_1 + \epsilon_2, \quad \mathbf{x}_3 = \mathbf{z}_1 + \epsilon_3$$

$$\mathbf{x}_4 = \mathbf{z}_2 + \epsilon_4, \quad \mathbf{x}_5 = -\mathbf{z}_2 + \epsilon_5, \quad \mathbf{x}_6 = \mathbf{z}_2 + \epsilon_6$$

- Fit the model on $(\mathbf{X}, \mathbf{y})$
- An "oracle" would identify $\mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$ (the $\mathbf{z}_1$ group) as the most important variables.


Lasso
Standardized Coefficients
40
30
20
10
0
−10
0.0
0.2
0.4
0.6
0.8
1.0
3
6
5
1
4
2
s = |beta|/max|beta|


Elastic Net lambda = 0.5
Standardized Coefficients
20
10
0
−10
−20
0.0
0.2
0.4
0.6
0.8
1.0
3
1
4
6
5
2
s = |beta|/max|beta|


Lasso
Standardized Coefficients
s = |beta|/max|beta|
30
20
10
0
−10
−20
0.0
0.2
0.4
0.6
0.8
1.0
5
6
1
3
4
2
Elastic Net lambda = 0.5
Standardized Coefficients
s = |beta|/max|beta|
20
10
0
−10
−20
0.0
0.2
0.4
0.6
0.8
1.0
3
1
4
6
5
2

Results on the grouping effect

Regression

Let $\rho_{ij} = \widehat{\text{cor}}(\mathbf{x}_i, \mathbf{x}_j)$. Suppose $\hat{\beta}_i(\lambda_1)\hat{\beta}_j(\lambda_1) > 0$, then $\frac{1}{|\mathbf{y}|}|\hat{\beta}_i(\lambda_1) - \hat{\beta}_j(\lambda_1)| \leq \frac{\sqrt{2}}{\lambda_2}\sqrt{1-\rho_{ij}}$.

Classification Let ϕ be a margin-based loss function, i.e., $\phi(y, f) = \phi(yf)$ and $y \in \{1, -1\}$. Consider

$$\hat{\beta} = \arg\min_{\beta} \sum_{k=1}^{n} \phi\left(y_k \mathbf{x}_k^T \beta\right) + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1$$

Assume that ϕ is Lipschitz, i.e., $|\phi(t_1) - \phi(t_2)| \leq M\,|t_1 - t_2|$, then $\forall$ a pair of (i, j), we have

$$\left|\hat{\beta}_i - \hat{\beta}_j\right| \leq \frac{M}{\lambda_2} \sum_{k=1}^{n} |\mathbf{x}_{k,i} - \mathbf{x}_{k,j}| \leq \frac{\sqrt{2}M}{\lambda_2}\sqrt{1-\rho_{ij}}.$$

Elastic net with scaling correction

$$\hat{\beta}_{\text{enet}} \stackrel{\text{def}}{=} (1+\lambda_2)\hat{\beta}$$

- Keep the grouping effect and overcome the double shrinkage by the quadratic penalty.
- Consider $\widehat{\Sigma} = \mathbf{X}^T\mathbf{X}$ and $\widehat{\Sigma}_{\lambda_2} = (1-\gamma)\widehat{\Sigma} + \gamma\mathbf{I}$, $\gamma = \frac{\lambda_2}{1+\lambda_2}$. $\widehat{\Sigma}_{\lambda_2}$ is a shrunken estimate for the correlation matrix of the predictors.
- Decomposition of the ridge operator: $\hat{\beta}_{\text{ridge}} = \frac{1}{1+\lambda_2}\widehat{\Sigma}_{\lambda_2}^{-1}\mathbf{X}^T\mathbf{y}$.
- We can show that

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \beta^T\widehat{\Sigma}\beta - 2\mathbf{y}^T\mathbf{X}\beta + \lambda_1\|\beta\|_1$$

$$\hat{\beta}_{\text{enet}} = \arg\min_{\beta} \beta^T\widehat{\Sigma}_{\lambda_2}\beta - 2\mathbf{y}^T\mathbf{X}\beta + \lambda_1\|\beta\|_1$$

- With orthogonal predictors, $\hat{\beta}_{\text{enet}}$ reduces to the (minimax) optimal soft-thresholding estimator.

Computation

- The elastic net solution path is *piecewise linear*.
- Given a fixed λ_2, a stage-wise algorithm called LARS-EN efficiently solves the *entire* elastic net solution path.
  - At step k, efficiently updating or downdating the Cholesky factorization of $\mathbf{X}^T_{\mathcal{A}_{k-1}}\mathbf{X}_{\mathcal{A}_{k-1}} + \lambda_2\mathbf{I}$, where $\mathcal{A}_k$ is the active set at step k.
  - Only record the non-zero coefficients and the active set at each LARS-EN step.
  - Early stopping, especially in the $p \gg n$ problem.
- `R` package: *elasticnet*

Simulation example 1: 50 data sets consisting of 20/20/200 observations and 8 predictors. $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 3$. $\text{cor}(\mathbf{x}_i, \mathbf{x}_j) = (0.5)^{|i-j|}$.

Simulation example 2: Same as example 1, except $\beta_j = 0.85$ for all j.

Simulation example 3: 50 data sets consisting of 100/100/400 observations and 40 predictors. $\beta = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})$ and $\sigma = 15$; $\text{cor}(x_i, x_j) = 0.5$ for all i, j.

Simulation example 4: 50 data sets consisting of 50/50/400 observations and 40 predictors. $\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})$ and $\sigma = 15$.

$$
\begin{aligned}
\mathbf{x}_i &= Z_1 + \epsilon_i^x, \quad Z_1 \sim N(0,1), \quad i = 1, \cdots, 5, \\
\mathbf{x}_i &= Z_2 + \epsilon_i^x, \quad Z_2 \sim N(0,1), \quad i = 6, \cdots, 10, \\
\mathbf{x}_i &= Z_3 + \epsilon_i^x, \quad Z_3 \sim N(0,1), \quad i = 11, \cdots, 15, \\
\mathbf{x}_i &\sim N(0,1), \qquad \mathbf{x}_i \quad \text{i.i.d} \qquad i = 16, \ldots, 40.
\end{aligned}
$$

Median MSE for the simulated examples

Method	Ex.1	Ex.2	Ex.3	Ex.4
Ridge	4.49 (0.46)	2.84 (0.27)	39.5 (1.80)	64.5 (4.78)
Lasso	3.06 (0.31)	3.87 (0.38)	65.0 (2.82)	46.6 (3.96)
Elastic Net	2.51 (0.29)	3.16 (0.27)	56.6 (1.75)	34.5 (1.64)
No re-scaling	5.70 (0.41)	2.73 (0.23)	41.0 (2.13)	45.9 (3.72)

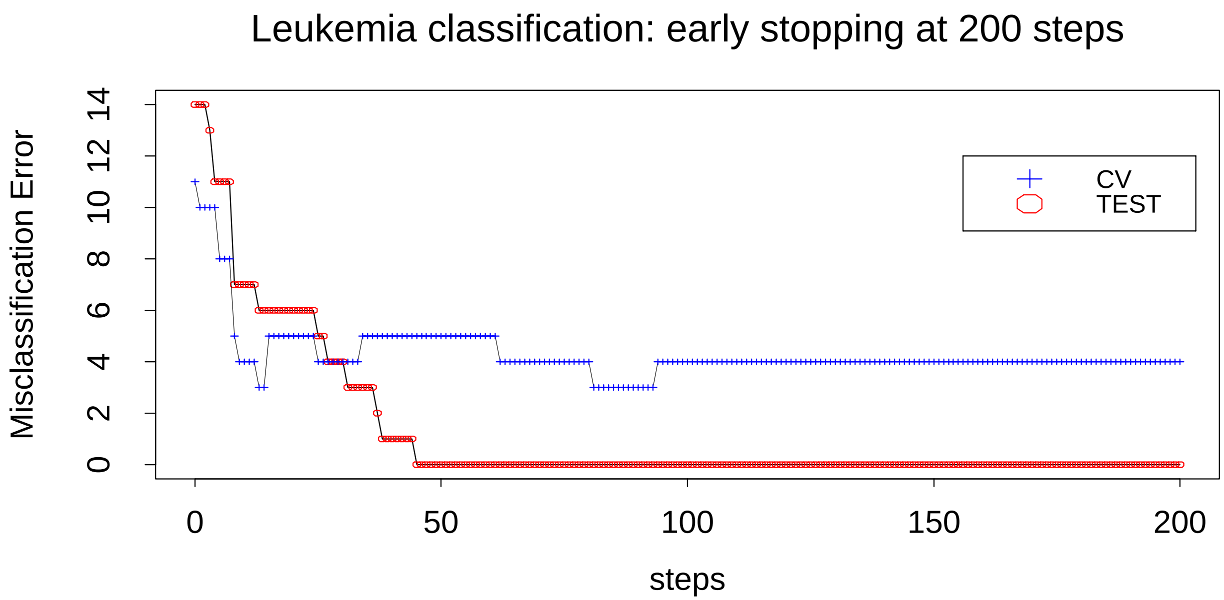
Variable selection results

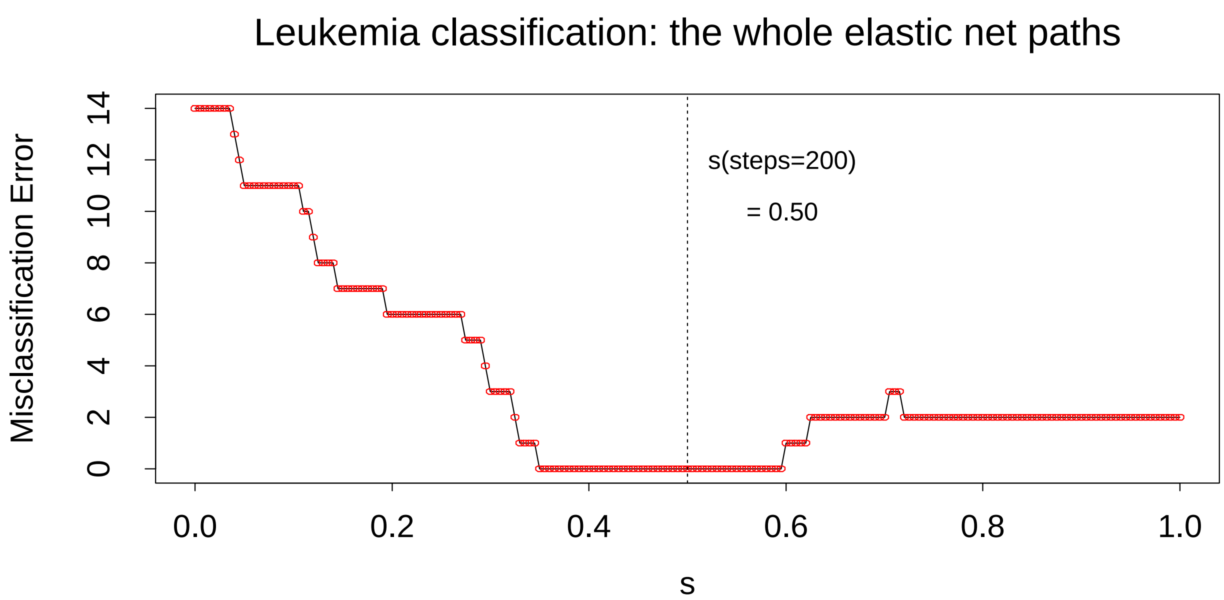
Method	Ex.1	Ex.2	Ex.3	Ex.4
Lasso	5	6	24	11
Elastic Net	6	7	27	16

Leukemia classification example

Method	10-fold CV error	Test error	No. of genes
Golub UR	3/38	4/34	50
SVM RFE	2/38	1/34	31
PLR RFE	2/38	1/34	26
NSC	2/38	2/34	21
Elastic Net	2/38	0/34	45

UR: univariate ranking (Golub et al. 1999)
RFE: recursive feature elimination (Guyon et al. 2002)
SVM: support vector machine (Guyon et al. 2002)
PLR: penalized logistic regression (Zhu and Hastie 2004)
NSC: nearest shrunken centroids (Tibshirani et al. 2002)


Leukemia classification: early stopping at 200 steps
Misclassification Error
0
2
4
6
8
10
12
14
0
50
100
150
200
steps
CV
TEST


Leukemia classification: the whole elastic net paths
Misclassification Error
0
2
4
6
8
10
12
14
0.0
0.2
0.4
0.6
0.8
1.0
s
s(steps=200)
= 0.50

Effective degrees of freedom

- Effective *df* describes the model complexity.
- *df* is very useful in estimating the prediction accuracy of the fitted model.
- *df* is well studied for linear smoothers: $\hat{\boldsymbol{\mu}} = \mathbf{S}\mathbf{y}$, $df(\hat{\boldsymbol{\mu}}) = \mathrm{tr}(\mathbf{S})$.
- For the ℓ_1 related methods, the *non-linear* nature makes the analysis difficult.
- Conjecture by Efron et al. (2004): *Starting at step 0, let m_k be the index of the last model in the Lasso sequence containing exact k predictors. Then $df(m_k) \doteq k$.*

Elastic Net: degrees of freedom

- $df = \mathrm{E}[\widehat{df}]$, where $\widehat{df}$ is an unbiased estimate for df, and

$$\widehat{df} = \mathrm{Tr}\left(\mathbf{H}_{\lambda_2}(\mathcal{A})\right)$$

where $\mathcal{A}$ is the active set and

$$\mathbf{H}_{\lambda_2}(\mathcal{A}) = \mathbf{X}_{\mathcal{A}}\left(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}} + \lambda_2\mathbf{I}\right)^{-1}\mathbf{X}_{\mathcal{A}}^T.$$

- For the lasso ($\lambda_2 = 0$),

$$\widehat{df}(\text{lasso}) = \text{the number of nonzero coefficients.}$$

- Proof: SURE+LARS+convex analysis

Elastic Net: other applications

- Sparse PCA
  - Obtain (modified) principal components with sparse loadings.
- Kernel elastic net
  - Generate a class of kernel machines with support vectors.

Sparse PCA

- $\mathbf{X}_{n \times p}$ and $\mathbf{x}_i$ is the i-th row vector of $\mathbf{X}$.
- α and β are p-vectors.

SPCA: the leading sparse PC

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \|\mathbf{x}_i - \alpha\beta^T \mathbf{x}_i\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1$$

subject to $\|\alpha\|^2 = 1$.

$\hat{v} = \frac{\hat{\beta}}{\|\hat{\beta}\|}$, the loadings.

- A large λ_1 generates sparse loadings.
- The equivalence theorem: consider the SPCA with $\lambda_1 = 0$
  1. $\forall \lambda_2 > 0$, SPCA $\equiv$ PCA;
  2. When $p > n$, SPCA $\equiv$ PCA if only if $\lambda_2 > 0$.

Sparse PCA (cont.)

- $\mathbf{A}_{p\times k} = [\alpha_1, \cdots, \alpha_k]$ and $\mathbf{B}_{p\times k} = [\beta_1, \cdots, \beta_k]$

SPCA: the first k sparse PCs

$$\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\|^2 + \lambda_2 \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{1j} \|\beta_j\|_1$$

subject to $\mathbf{A}^T\mathbf{A} = \mathbf{I}_{k\times k}$.

Let $\hat{v}_j = \frac{\hat{\beta}_j}{\|\hat{\beta}_j\|}$, for $j = 1, \ldots, k$.

- Solution:
  - **B** given **A**: k independent elastic net problems.
  - **A** given **B**: exact solution by SVD.

SPCA algorithm

1. Let $\mathbf{A}$ start at $\mathbf{V}[\ ,1:k]$, the loadings of the first k ordinary principal components.
2. Given a fixed $\mathbf{A} = [\alpha_1, \cdots, \alpha_k]$, solve the following elastic net problem for $j = 1, 2, \ldots, k$

   $$\beta_j = \arg\min_{\beta} (\alpha_j - \beta)^T \mathbf{X}^T \mathbf{X} (\alpha_j - \beta) + \lambda_2 \|\beta\|^2 + \lambda_{1,j} \|\beta\|_1$$

3. For a fixed $\mathbf{B} = [\beta_1, \cdots, \beta_k]$, compute the SVD of $\mathbf{X}^T\mathbf{X}\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, then update $\mathbf{A} = \mathbf{U}\mathbf{V}^T$.
4. Repeat steps 2–3, until convergence.
5. Normalization: $\hat{v}_j = \frac{\beta_j}{\|\beta_j\|}$, $j = 1, \ldots, k$.

Sparse PCA: pitprops data example

- There are 13 measured variables. First introduced by Jeffers (1967) who tried to interpret the first 6 principal components.
- A classic example showing the difficulty of interpreting principal components.
- The original data have 180 observations. The sample correlation matrix (13×13) is sufficient in our analysis.

	PCA			SPCA		
topdiam	-.404	.218	-.207	-.477		
length	-.406	.186	-.235	-.476		
moist	-.124	.541	.141		.785	
testsg	-.173	.456	.352		.620	
ovensg	-.057	-.170	.481	.177		.640
ringtop	-.284	-.014	.475			.589
ringbut	-.400	-.190	.253	-.250		.492
bowmax	-.294	-.189	-.243	-.344	-.021	
bowdist	-.357	.017	-.208	-.416		
whorls	-.379	-.248	-.119	-.400		
clear	.011	.205	-.070			
knots	.115	.343	.092		.013	
diaknot	.113	.309	-.326			-.015
variance	32.4	18.3	14.4	28.0	14.0	13.3

Kernel Machines

- Binary classification: $y \in \{1, -1\}$.
- Take a margin-based loss function $\phi(y, f) = \phi(yf)$.
- A kernel matrix $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. We consider $\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \hat{\alpha}_i k(\mathbf{x}_i, \mathbf{x})$ with

$$\hat{\alpha} = \arg\min_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})) + \lambda_2 \alpha^T \mathbf{K} \alpha$$

- SVMs uses $\phi(y, f) = (1 - yf)_+$, the *hinge loss* (Wahba, 2000).
  - ✔ maximizes the margin
  - ✔ directly approximates the Bayes rule (Lin, 2002)
  - ✔ only a fraction of α are non-zero: support vectors
  - ✘ no estimate for $p(y|\mathbf{x})$

Kernel elastic net

- Take $\phi(y, f) = \log(1 + \exp(-yf))$. We consider $\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \hat{\alpha}_i k(\mathbf{x}_i, \mathbf{x})$ with

$$\hat{\alpha} = \arg\min_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})) + \lambda_2 \alpha^T \mathbf{K} \alpha + \lambda_1 \sum_{i=1}^{n} |\alpha_i|$$

✔ estimates $p(y|\mathbf{x})$

- KLR: $\lambda_1 = 0$, *no support vectors*

✔ a large λ_1 generates *genuine support vectors*

✔ combines margin maximization with boosting

  – λ_1 is the main tuning parameter: the regularization method in boosting (Rosset, Zhu and Hastie, 2004).

  – small positive λ_2: the limiting solution ($\lambda_1 \to 0$) is close to the margin-maximization classifier.

Summary

- The elastic net performs simultaneous regularization and variable selection.
- Ability to perform grouped selection
- Appropriate for the $p \gg n$ problem
- Analytical results on the *df* of the elastic net/lasso
- Interesting implications in other areas: sparse PCA and new support kernel machines

References

- Zou, H. and Hastie, T. (2004) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B.* To appear.
- Zou, H., Hastie, T. and Tibshirani, R. (2004). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics.* Tentatively accepted.
- Zou, H., Hastie, T. and Tibshirani, R. (2004). On the "Degrees of Freedom" of the Lasso. Submitted to *Annals of Statistics.*

http://www-stat.stanford.edu/~hzou