Linear Methods for Regression

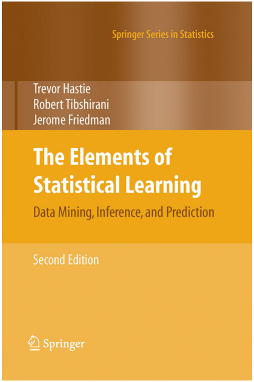
Stéphane Canu
stephane.canu@insa-rouen.fr

APPC - Advanced Machine Learning

11 septembre 2023

Outline

Chapter 3: Linear Methods for Regression, page 43



	AGE	SEX	BMI	BP	$\cdots$ Serum Measurements $\cdots$						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Table 1. Diabetes study. 442 diabetes patients were measured on 10 baseline variables. prediction model was desired for the response variable, a measure of disease progression one yea after baseline.

$$\mathbb{E}(Y|X) = f(X) = \beta_0^\star + \beta_1^\star X_1 + \cdots + \beta_p^\star X_p = \beta_0^\star + \beta^\star X$$

X the explanatory variables (independent variable or set of predictors)

Y the response to be predicted (dependent variable)

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ the parameters (unknown)

β_0 the intercept (also unknown)

p the number of variables of the model

More notations

$$\mathbb{E}(Y|X) = f(X) = \beta_0^\star + \sum_{j=1}^{p} \beta_j^\star X_j$$

$\beta^\star$ the true parameters (the good)

$\widehat{\beta}$ the estimated parameters (the bad)

β the free parameter (the ugly)

$$\widehat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} J(\beta, data)$$

The parameter estimation error (unknown)

$$\|\beta^\star - \widehat{\beta}\|^2$$

The prediction error (unknown)

$$\|Y - \widehat{\mathbf{y}}\|^2 \text{ with } \widehat{\mathbf{y}} = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j X_j$$

Outline

Data: $N = 446$ and $p = 10 + 1$ (Diabetes)

$$X = \begin{pmatrix} age & sex & \dots & x_{p=10} \\ 59 & 2 & \dots & 87 \\ \vdots & \vdots & \dots & \vdots \\ 36 & 1 & \dots & 92 \end{pmatrix}; \ \mathbf{y} = \begin{pmatrix} diab. \\ 151 \\ \vdots \\ 57 \end{pmatrix}$$

	AGE	SEX	BMI	BP	$\cdots$ Serum	Measurements				$\cdots$	Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Table 1. Diabetes study. 442 diabetes patients were measured on 10 baseline variables. prediction model was desired for the response variable, a measure of disease progression one yea after baseline.

Example of explanatory variables X_j

- quantitative inputs
- transformations of quantitative inputs (e.g. $\log, \exp, \sqrt{\ldots}$)
- basis expansions (e.g. $X_2 = X_1^2, X_3 = X_1^3 \ldots$)
- interactions (e.g. $X_3 = X_1 X_2$)
- binary variable
- numeric or “dummy” coding of the levels of qualitative inputs (one hot vector)

The model is linear with respect to β

not necessarily with respect to inputs

The least squares estimation principle

Linear model

$$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

Minimize the residual sum-of-squares (RSS)

$$\widehat{\beta}^{\text{ls}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}}\, RSS(\beta, data)$$

$$\begin{aligned} RSS(\beta, data) &= \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 \\ &= \sum_{i=1}^{n} \big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\big)^2 \\ &= \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 \qquad \text{with} \qquad \hat{y}_i = f(x_i), i = 1, n \end{aligned}$$

The intercept as a parameter

X_1	X_2	X_3	...	X_10	Y
-0.6416	-0.3934	-0.7313	...	0.9820	0.5524
0.8415	2.1730	0.3442	...	1.6411	0.7336
0.6521	0.5006	3.4307	...	-0.5842	-1.6001
0.1848	-0.2535	0.3812	...	-0.8544	-0.3328
0.3635	0.5542	1.9298	...	0.6564	0.6291
0.5890	-0.2903	1.5685	...	-1.6603	-0.5648
0.5153	-0.4606	0.2263	...	1.6934	-1.0449
0.9598	1.2378	-2.0753	...	0.5395	-0.4543
1.9332	2.2013	-0.2789	...	0.2030	-1.4092
-0.0956	-2.0578	-0.9055	...	-2.3387	1.0096
0.5377	2.7694	1.4090	...	-0.3034	0.3252
1.8339	-1.3499	1.4172	...	0.2939	-0.7549
-2.2588	3.0349	0.6715	...	-0.7873	1.3703
0.8622	0.7254	-1.2075	...	0.8884	-1.7115
0.3188	-0.0631	0.7172	...	-1.1471	-0.1022
-1.3077	0.7147	1.6302	...	-1.0689	-0.2414
-0.4336	-0.2050	0.4889	...	-0.8095	0.3192
0.3426	-0.1241	1.0347	...	-2.9443	0.3129
3.5784	1.4897	0.7269	...	1.4384	-0.8649

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} & \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} & \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

$$p = 10, \qquad n = 19$$

beware

$$X = \begin{pmatrix} \color{red}{1} & x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ \color{red}{1} & x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \color{red}{\vdots} & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \color{red}{1} & x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \color{red}{\vdots} & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \color{red}{1} & x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}$$

n lines and $p + 1$ columns

$$\widehat{\mathbf{y}} = X\beta$$

Intercept elimination through preprocessing

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

The normalized input variable is $\widetilde{x_j} = \frac{x_j - \bar{x_j}}{\sigma_j}$ with $\bar{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$

$$\begin{aligned} f(x) &= \beta_0 + \sum_{j=1}^{p} \beta_j(\sigma_j \widetilde{x_j} + \bar{x_j}) \\ &= \widetilde{\beta_0} + \sum_{j=1}^{p} \widetilde{\beta_j}\widetilde{x_j} \end{aligned}$$

with $\widetilde{\beta_0} = \beta_0 + \sum_{j=1}^{p} \beta_j \bar{x_j}$ and $\widetilde{\beta_j} = \beta_j \sigma_j$

The least square estimate of $\widetilde{\beta_0}$ is $\widehat{\tilde{\beta_0}} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
so that, with $\widetilde{y} = y - \bar{y}$ the equivalent intercept free model is

$$\widetilde{y} = \sum_{j=1}^{p} \widetilde{\beta_j}\widetilde{x_j} + \varepsilon$$

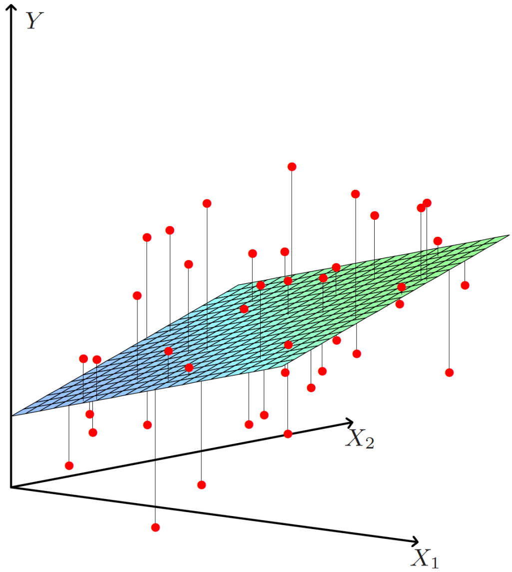


FIGURE 3.1. *Linear least squares fitting with* $X \in \mathbb{R}^2$*. We seek the linear function of* X *that minimizes the sum of squared residuals from* Y*.*

How do we minimize the residual sum-of-squares

The cost

$$
\begin{aligned}
RSS(\beta, data) &= \|\mathbf{y} - X\beta\|^2 \\
&= (\mathbf{y} - X\beta)^\top(\mathbf{y} - X\beta) \\
&= \mathbf{y}^\top\mathbf{y} - 2\beta^\top X^\top\mathbf{y} + \beta^\top X^\top X\beta
\end{aligned}
$$

The gradient

$$\nabla_\beta RSS(\beta, data) = -2X^\top\mathbf{y} + 2X^\top X\beta$$

The Hessian

$$\nabla_\beta^2 RSS(\beta, data) = 2X^\top X$$

The Fermat's rule

$$\widehat{\beta}^{\mathrm{ls}} = \operatorname*{argmin}_{\beta \in \mathbb{R}^{p+1}} RSS(\beta, data) \quad \Leftrightarrow \quad \nabla_\beta RSS(\widehat{\beta}^{\mathrm{ls}}, data) = 0$$

The unique solution

$$\widehat{\beta}^{\mathrm{ls}} = (X^\top X)^{-1}X^\top\mathbf{y}$$

The least square estimator

$(X^\top X)^{-1}$ must exist

Optimality as orthogonality

$$\nabla_\beta RSS(\widehat{\beta}^{\mathsf{ls}}, data) = 0 \quad \Leftrightarrow \quad X^\top(\mathbf{y} - X\widehat{\beta}^{\mathsf{ls}}) = 0$$

$$\widehat{\mathbf{y}} = X\widehat{\beta}^{\mathsf{ls}} = X(X^\top X)^{-1}X^\top \mathbf{y} = H\mathbf{y}$$

with H the hat matrix (also called projector of influence matrix)

$$H = X(X^\top X)^{-1}X^\top$$

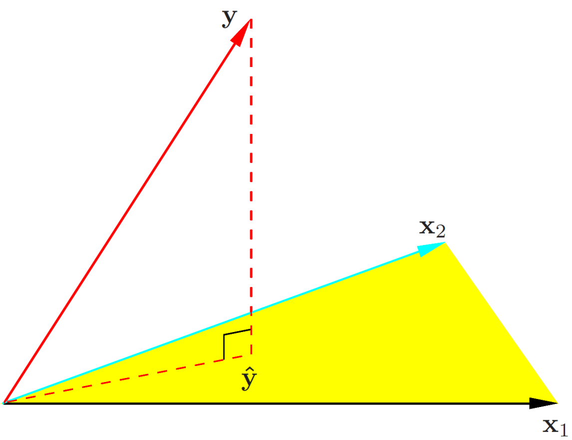


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

Statistical analysis of $\widehat{\beta}^{\text{ls}}$

the x_j are fixed (non random).

$$Y = \beta_0^\star + \sum_{j=1}^{p} \beta_j^\star x_j + \varepsilon = X\beta^\star + \varepsilon$$

Gaussian and uncorrelated error assumption $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

$$\widehat{\beta}^{\text{ls}} = (X^\top X)^{-1} X^\top Y$$

$$\begin{aligned} \mathbb{E}(\widehat{\beta}^{\text{ls}}) &= (X^\top X)^{-1} X^\top \mathbb{E}(Y) \\ &= (X^\top X)^{-1} X^\top X\beta^\star + \mathbb{E}(\varepsilon) = \beta^\star \end{aligned}$$

$$\begin{aligned} V(\widehat{\beta}^{\text{ls}}) &= (X^\top X)^{-1} X^\top V(Y) X (X^\top X)^{-1} \\ &= (X^\top X)^{-1} \sigma^2 \end{aligned}$$

$$\widehat{\beta}^{\text{ls}} \sim \mathcal{N}(\beta^\star, (X^\top X)^{-1}\sigma^2)$$

Variance unbiased estimates

$$\widehat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad \sim \sigma^2 \chi^2_{n-p-1}$$

Z-score: how likely $\beta_j^\star = 0$

$$z_j = \frac{\widehat{\beta}_j^{\mathsf{ls}}}{\widehat{\sigma}\sqrt{v_j}} \sim t_{n-p-1} \qquad \text{with} \qquad v_j \text{ the jth diagonal element of } (X^\top X)^{-1}$$

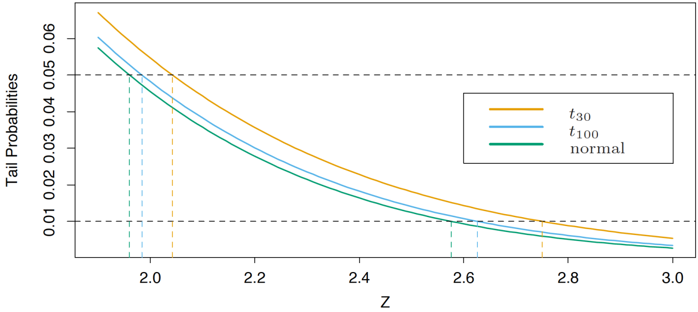


FIGURE 3.3. *The tail probabilities* $\Pr(|Z| > z)$ *for three distributions,* t_{30}, t_{100} *and standard normal. Shown are the appropriate quantiles for testing significance at the* $p = 0.05$ *and* 0.01 *levels. The difference between* t *and the standard normal becomes negligible for* N *bigger than about* 100.

Example: Prostate Cancer

TABLE 3.1. *Correlations of predictors in the prostate cancer data.*

	lcavol	lweight	age	lbph	svi	lcp	gleason
lweight	0.300						
age	0.286	0.317					
lbph	0.063	0.437	0.287				
svi	0.593	0.181	0.129	−0.139			
lcp	0.692	0.157	0.173	−0.089	0.671		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	
pgg45	0.483	0.074	0.276	−0.030	0.481	0.663	0.757

TABLE 3.2. *Linear model fit to the prostate cancer data. The Z score is the coefficient divided by its standard error (3.12). Roughly a Z score larger than two in absolute value is significantly nonzero at the $p = 0.05$ level.*

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	−0.14	0.10	−1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	−0.29	0.15	−1.87
gleason	−0.02	0.15	−0.15
pgg45	0.27	0.15	1.74

Data diagnostic is missing

(x_i, y_i), $i = 1, \dots, n$ → **Estimation** $\widehat{\beta} = (X^\top X)^{-1} X^\top \mathbf{y}$ → $\widehat{\beta}$

$\widehat{\beta}$, $\mathbf{x}$ → **Prediction** $\widehat{y} = \mathbf{x}^\top \widehat{\beta}$ → $\widehat{y}$

Model diagnostic: $\widehat{y} \pm \delta_y$

model is the model well fitted?

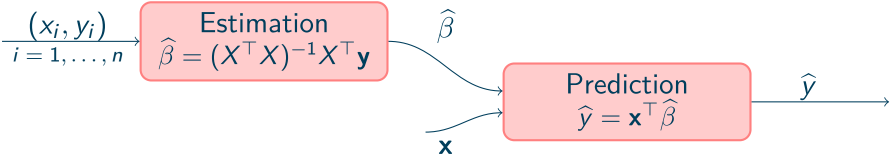
- observation = information + noise
- check model hypothesis

observations is there some wrong data

- wrong x
- wrong y
- wrong (x, y)

variables selection: eliminate useless variables

Data diagnostic is missing



Model diagnostic: $\widehat{y} \pm \delta_y$

model is the model well fitted?

- observation = information + noise
- check model hypothesis

observations is there some wrong data

- wrong x
- wrong y
- wrong (x, y)

variables selection: eliminate useless variables

Outline

Subset selection

Why subset selection?

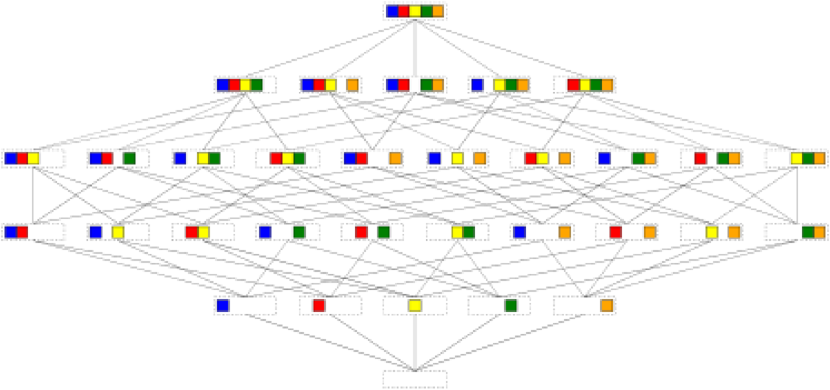
- prediction accuracy $\rightarrow$ remove spurious variables
- interpretation $\rightarrow$ reduce the number of variables

Different solutions

- exact solution: best subset
- greedy approaches

    forward selection (stepwise)
    backward selection (stepwise)
    Forward-Stagewise Regression

The variable lattice



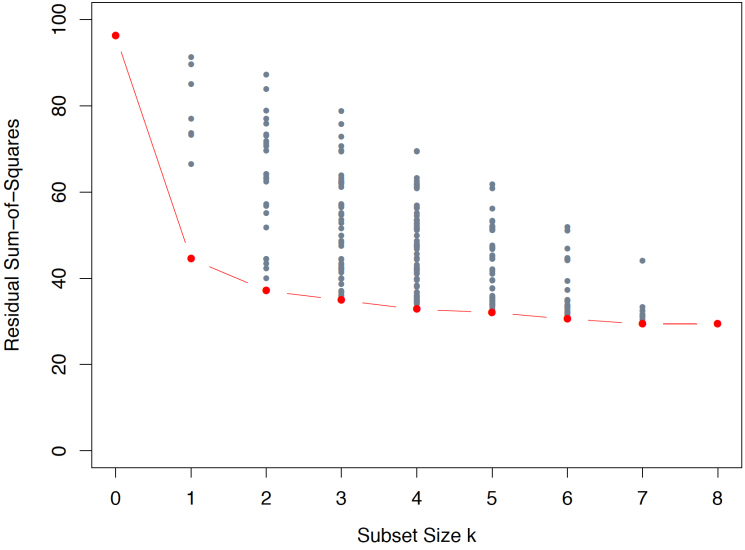


FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

Forward-Stepwise Regression

- Subset selection infeasible with large p: select variables one by one

- The Forward-Stepwise Regression algorithm

initialization: $V_{in} = \emptyset$ and $V_{out} = \{1, \dots, p\}$
for all the variables $j = 1$ to p

$$k^\star = \underset{k \in V_{out}}{\text{argmin}} \min_\beta \sum_{i=1}^n \Big(y_i - \sum_{k' \in V_{in}} x_{ik'}\beta_{k'} + x_{ik}\beta\Big)^2 \qquad \leftarrow \text{1d LS}$$

$V_{in} = V_{in} \cup k^\star$ and $V_{out} = V_{out} \backslash k^\star$

- The nature of the Forward-Stepwise Regression
  - a greedy algorithm
  - producing a nested sequence of models $\quad \rightarrow$ when to stop?
  - sub-optimal

- However, it may be preferable to best subset for
  - Computational reasons ($p(p-1/2$ vs $2^p)$)
  - Statistical reasons (less model, less variance)

Alternatives to Forward-Stepwise Regression

Backward-Stepwise

start with: $V_{in} = \{1, \ldots, p\}$ and $V_{out} = \emptyset$

Forward-Backward-Stepwise

is it better to add or to remove a variable?

Different selection criterion:

Stagewise Regression

Forward-Stagewise Regression

identifies the variable most correlated with the current residual.
X normalized (centered and reduced)
$\beta_0 = \bar{y}$
$\boldsymbol{\beta} = 0$
$\mathbf{r} = \mathbf{y} - \beta_0$
repeat until convergence (for a given stepsize ρ)

$$\widehat{j} = \operatorname*{argmin}_{j=1,\dots,p} |\mathbf{r}^\top \mathbf{x}_j|$$

$$\beta_{\widehat{j}} = \beta_{\widehat{j}} + \rho\, sign(\mathbf{r}^\top \mathbf{x}_{\widehat{j}})$$

$$\mathbf{r} = \mathbf{r} - \mathbf{x}_{\widehat{j}} \rho\, sign(\mathbf{r}^\top \mathbf{x}_{\widehat{j}})$$

Computationally efficient optimization procedure

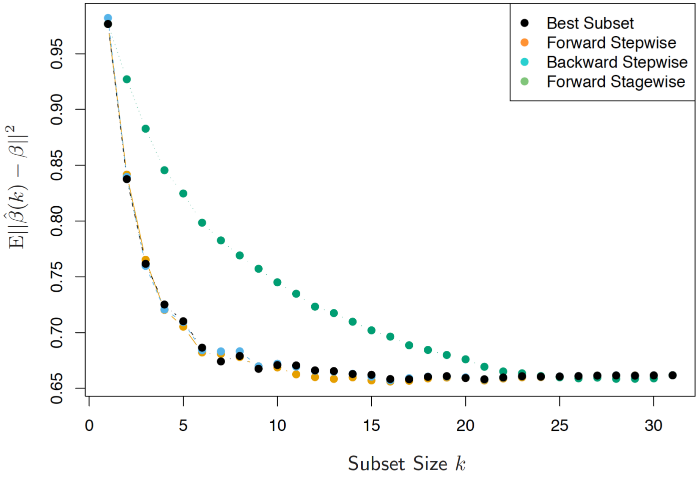


FIGURE 3.6. *Comparison of four subset-selection techniques on a simulated linear regression problem $Y = X^T\beta + \varepsilon$. There are $N = 300$ observations on $p = 31$ standard Gaussian variables, with pairwise correlations all equal to 0.85. For 10 of the variables, the coefficients are drawn at random from a $N(0, 0.4)$ distribution; the rest are zero. The noise $\varepsilon \sim N(0, 6.25)$, resulting in a signal-to-noise ratio of 0.64. Results are averaged over 50 simulations. Shown is the mean-squared error of the estimated coefficient $\hat{\beta}(k)$ at each step from the true β.*

Outline

Outline

The Shrinkage

Find better estimate - continuous

Shrinkage estimate

$$\widehat{\beta}_j^{\mathsf{S}} = (1-\alpha)\widehat{\beta}_j^{\mathsf{ls}} \qquad 0 \leq \alpha \leq 1$$

Justification from the James-Stein estimator (1960)

$$\widehat{\beta}_j^{\mathsf{JS}} = \left(1 - \frac{(p-2)\sigma^2}{\|\widehat{\beta}_j^{\mathsf{ls}}\|^2}\right)_+ \widehat{\beta}_j^{\mathsf{ls}}$$

the estimator risk and its bias variance decomposition

$$\begin{aligned} \mathbb{E}(\|\beta^\star - \widehat{\beta}\|^2) &= \mathbb{E}(\|\beta^\star - \mathbb{E}(\widehat{\beta}) + \mathbb{E}(\widehat{\beta}) - \widehat{\beta}\|^2) \\ &= \|\beta^\star - \mathbb{E}(\widehat{\beta})\|^2 + \mathbb{E}(\|\widehat{\beta} - \mathbb{E}(\widehat{\beta})\|^2) \end{aligned}$$

improve the estimation

- add some bias
- reduce the variance

The ridge regression: a L^2 penalty

(Hoerl and Kennard, 1970).
given $0 \leq \lambda$, minimize

$$RSSR_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

center X and $\beta_0 = \bar{y}$

$$RSSR(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2.$$

$$\widehat{\beta}^{\mathrm{R}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}}\, RSSR_\lambda(\boldsymbol{\beta})$$

$$\begin{aligned} \lambda = 0 \quad & \widehat{\beta}^{\mathrm{R}}(0) = \widehat{\beta}^{\mathrm{ls}} \\ \lambda \to \infty \quad & \widehat{\beta}^{\mathrm{R}}(\lambda) \to 0 \end{aligned}$$

for a well chosen λ, $\widehat{\beta}^{\mathrm{R}}(0)$ improves on $\widehat{\beta}^{\mathrm{ls}}$

The ridge regression: a bias and variance case study

the estimator risk and its bias variance decomposition

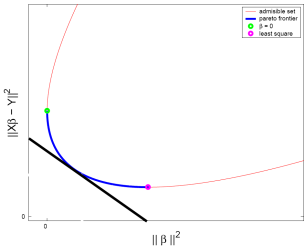
$$\begin{array}{ccccc} \mathbb{E}(\|\beta^\star - \widehat{\beta}\|^2) & = & \|\beta^\star - \mathbb{E}(\widehat{\beta})\|^2 & + & \mathbb{E}(\|\widehat{\beta} - \mathbb{E}(\widehat{\beta})\|^2) \\ Risk_\lambda & = & B_\lambda^2 & + & V_\lambda \end{array}$$

3 equivalent formulations

$$\min_{\beta} \; \|\mathbf{y} - X\beta\|^2 + \lambda \, \|\beta\|^2$$

$$\begin{cases} \min_{\beta} & \|\mathbf{y} - X\beta\|^2 \\ \text{s.t.} & \|\beta\|^2 \leq t \end{cases}$$

$$\begin{cases} \min_{\beta} & \|\beta\|^2 \\ \text{s.t.} & \|\mathbf{y} - X\beta\|^2 \leq k \end{cases}$$



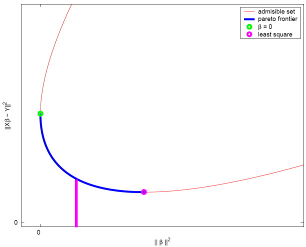
thanks to convexity

3 equivalent formulations

$$\min_{\beta} \; \|\mathbf{y} - X\beta\|^2 + \lambda \, \|\beta\|^2$$

$$\begin{cases} \min\limits_{\beta} \; \|\mathbf{y} - X\beta\|^2 \\ \text{s.t.} \;\; \|\beta\|^2 \le t \end{cases}$$

$$\begin{cases} \min\limits_{\beta} \; \|\beta\|^2 \\ \text{s.t.} \;\; \|\mathbf{y} - X\beta\|^2 \le k \end{cases}$$



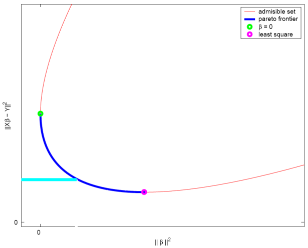
thanks to convexity

3 equivalent formulations

$$\min_{\beta} \ \|\mathbf{y} - X\beta\|^2 + \lambda \ \|\beta\|^2$$

$$\begin{cases} \min\limits_{\beta} \ \|\mathbf{y} - X\beta\|^2 \\ \text{s.t.} \ \ \|\beta\|^2 \leq t \end{cases}$$

$$\begin{cases} \min\limits_{\beta} \ \|\beta\|^2 \\ \text{s.t.} \ \ \|\mathbf{y} - X\beta\|^2 \leq k \end{cases}$$



thanks to convexity

How to minimize the ridge loss

The cost

$$
\begin{aligned}
RSSR(\boldsymbol{\beta}) &= \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \\
&= \mathbf{y}^\top\mathbf{y} - 2\beta^\top X^\top\mathbf{y} + \beta^\top X^\top X\beta + \lambda\beta^\top\beta \\
&= \mathbf{y}^\top\mathbf{y} - 2\beta^\top X^\top\mathbf{y} + \beta^\top(X^\top X + \lambda I)\beta
\end{aligned}
$$

The gradient

$$\nabla_\beta RSS(\beta, data) = -2X^\top\mathbf{y} + 2(X^\top X + \lambda I)\beta$$

The Hessian

$$\nabla^2_\beta RSS(\beta, data) = 2(X^\top X + \lambda I)$$

The Fermat's rule

$$\widehat{\beta}^{\mathrm{R}} = \operatorname*{argmin}_{\beta\in\mathbb{R}^{p+1}} RSSR(\boldsymbol{\beta}) \quad \Leftrightarrow \quad \nabla_\beta RSSR(\widehat{\beta}^{\mathrm{R}}) = 0$$

The unique solution

$$\widehat{\beta}^{\mathrm{R}} = (X^\top X + \lambda I)^{-1}X^\top\mathbf{y}$$

A ridge regression interpretation: the orthogonal case

orthogonal design
$X^\top X = I$

$$\widehat{\beta}^{\mathrm{R}} = (I + \lambda I)^{-1} X^\top \mathbf{y}$$

That is component wise the following shrinkage:

$$\widehat{\beta}^{\mathrm{R}}_j = \left(1 - \frac{\lambda}{1+\lambda}\right) \widehat{\beta}^{\mathrm{ls}}_j$$

with the associated bias

The ridge regression and the SVD

$$X = UDV^\top$$

$$\begin{array}{rcl} X\widehat{\beta}_j^{\text{ls}} & = & X(X^\top X)^{-1}X^\top \mathbf{y} \\ & = & UDV^\top(VD^2V^\top)^{-1}VDU^\top \mathbf{y} \\ & = & UU^\top \mathbf{y} \end{array}$$

$$\begin{array}{rcl} X\widehat{\beta}_j^{\text{R}} & = & X(X^\top X + \lambda I)^{-1}X^\top \mathbf{y} \\ & = & UDV^\top(V(D^2 + \lambda I)V^\top)^{-1}VDU^\top \mathbf{y} \\ & = & \sum_{j=1}^p \mathbf{u}_j \frac{d_j^2}{d_j^2+\lambda} \mathbf{u}_j^\top \mathbf{y} \end{array}$$

Shrinking the eigenvalues

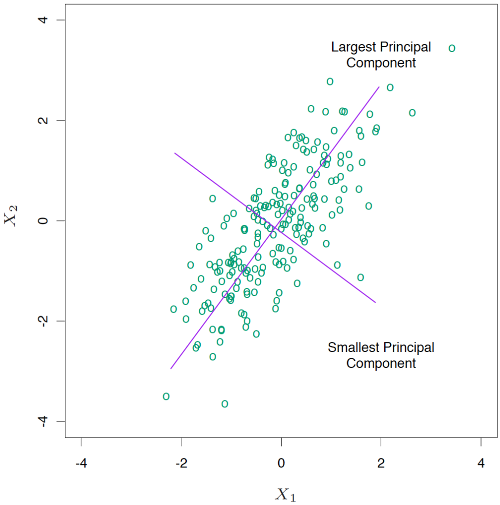


FIGURE 3.9. *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low–variance components more than the high-variance components.*

Degrees of freedom

Considering the least square estimate (no shrinkage, no penalty) the hat matrix H is a projector ($H^2 = H$)

$$\widehat{\mathbf{y}} = X\widehat{\beta}_{\mathrm{ls}} = H\mathbf{y}$$

so that $\widehat{\mathbf{y}}$ lies in a subspace of dimension $p = Trace(H)$

For the ridge regression, by analogy

$$\widehat{\mathbf{y}} = X\widehat{\beta}^{\mathrm{R}}_{\lambda} = H_{\lambda}\mathbf{y}$$

with

$$H_{\lambda} = X(X^{\top}X + \lambda I)^{-1}X^{\top}$$

the effective degrees of fredom (number of parameters)

$$df = Trace(H_{\lambda})$$

that is, if d_j are the singular values of X

$$df(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

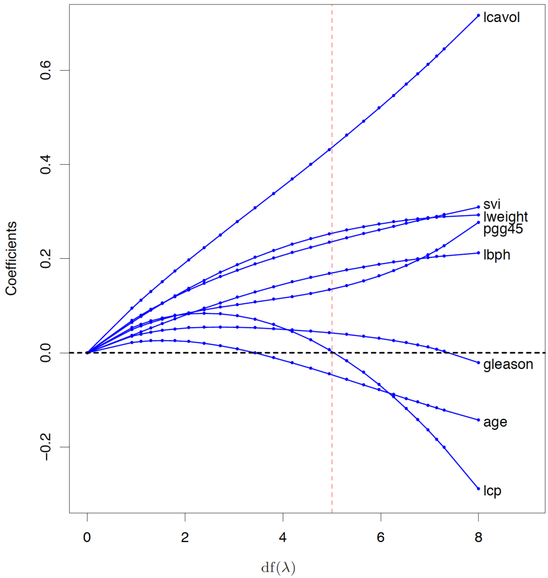


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.*

Outline

Least absolute shrinkage and selection operator

(Tibshirani, 1996)

given $0 \leq \lambda$, minimize

$$RSSR_{\lambda}(\beta) = \sum_{i=1}^{n} \big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\big)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

center X and $\mathbf{y}$ and set $\beta_0 = \bar{y}$

$$RSSL(\beta) = \|\mathbf{y} - X\beta\|^2 + \lambda \|\beta\|_1.$$

Most often, it is advisable to standardize the data

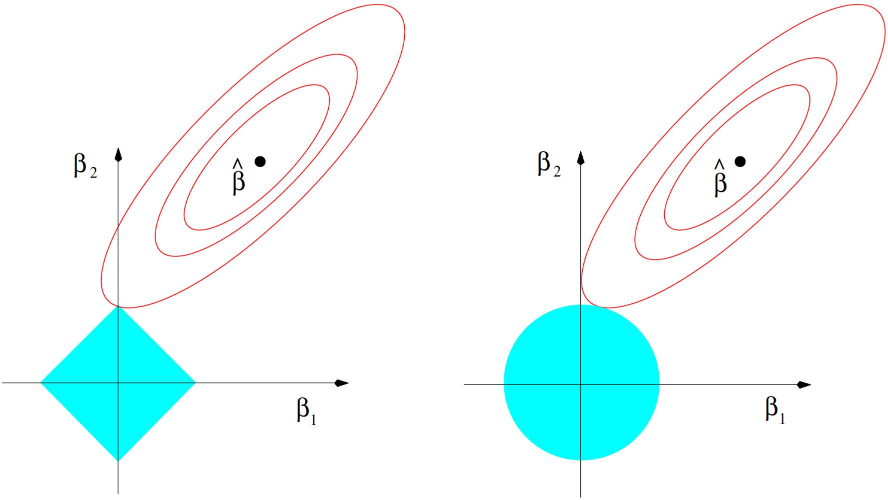


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

3 equivalent formulations

$$\min_{\beta} \; \|\mathbf{y} - X\beta\|^2 + \lambda \, \|\beta\|_1$$

$$\begin{cases} \min\limits_{\beta} \; \|\mathbf{y} - X\beta\|^2 \\ \text{s.t.} \;\; \|\beta\|_1 \leq t \end{cases}$$

$$\begin{cases} \min\limits_{\beta} \; \|\beta\|_1 \\ \text{s.t.} \;\; \|\mathbf{y} - X\beta\|^2 \leq k \end{cases}$$

The Lasso as a pQP

The Lasso is a pQP

$$\begin{cases} \min_{\boldsymbol{\beta}} \ \|\mathbf{y} - X\boldsymbol{\beta}\|^2 \\ \text{s.t.} \ \|\boldsymbol{\beta}\|_1 \leq t \end{cases}$$

introducing $\alpha_j = |\beta_j|$

$$\begin{cases} \min_{\boldsymbol{\beta},\boldsymbol{\alpha}} & \|\mathbf{y} - X\boldsymbol{\beta}\|^2 \\ \text{s.t.} & \sum_{j=1}^{p} \alpha_j \leq t \\ & -\alpha_j \leq \beta_j \leq \alpha_j \qquad j = 1, \ldots, p \\ & 0 \leq \alpha_j \end{cases}$$

The Lasso as a standard QP

The Lasso

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\boldsymbol{\alpha}} & \|\mathbf{y} - X\boldsymbol{\beta}\|^2 \\ \text{s.t.} & \sum\limits_{j=1}^{p} \alpha_j \le t & . \\ & -\alpha_j \le \beta_j \le \alpha_j \qquad j = 1, \dots, p \\ & 0 \le \alpha_j & . \end{cases}$$

A standard QP

$$\begin{cases} \min\limits_{z} & \frac{1}{2} z^\top Q z - c^\top z \\ \text{s.t.} & Az \le b \\ & lb \le z \le ub \end{cases}$$

$z = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$, $Q = (X^\top X, 0_p; (0_p, 0_p))$, $c = (X^\top y, 0_p)^\top$,
$A = ((0_p, 1_p); I_p, -I_p); -I_p - I_p)$, $b = t(1, 0_p, 0_p)^\top$,
$lb = (-\infty, 0_p)^\top$, $ub = (\infty, \infty)^\top$,

The KKT

The Lasso

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\boldsymbol{\alpha}} & \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|^2 \\ \text{s.t.} & \sum\limits_{j=1}^{p} \alpha_j \leq t \\ & -\alpha_j \leq \beta_j \leq \alpha_j \qquad j = 1, \dots, p \\ & 0 \leq \alpha_j \end{cases}$$

The associated lagrangian

$$\begin{aligned} \mathcal{L}(\alpha; \beta,) = & \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|^2 - \lambda(\sum_{j=1}^{p} \alpha_j - t) \\ & + \sum_{j=1}^{p} \mu_j^-(-\alpha_j - \beta_j) + \sum_{j=1}^{p} \mu_j^+(-\alpha_j + \beta_j) - \sum_{j=1}^{p} \gamma_j \alpha_j \end{aligned} \tag{1}$$

The KKT

First order optimality conditions

$$\widehat{\beta}^{\mathrm{L}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}}\, RSSL(\beta) \quad \Leftrightarrow \quad \widehat{\beta}^{\mathrm{L}} \text{ has to fulfill the KKT}$$

Karush, Kuhn and Tucker (KKT) conditions

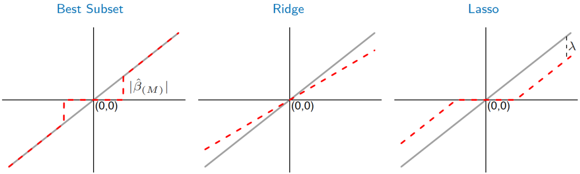
stationarity	$X^\top(X\beta - \mathbf{y}) - \mu^- + \mu^+ = 0$	
	$\lambda \mathbf{e} - \mu^- + \mu^+ - \gamma = 0$	
primal admissibility	$\sum_{j=1}^{p} \alpha_j \leq t$	
	$-\alpha_j \leq \beta_j \leq \alpha_j$	$j = 1, \ldots, p$
	$0 \leq \alpha_j$	$j = 1, \ldots, p$
dual admissibility	$\lambda, \mu_j^-, \mu_j^+, \gamma_j \geq 0$	$j = 1, \ldots, p$
complementarity	$\lambda(\sum_{j=1}^{p} \alpha_j - t) = 0$	
	$\mu_j^-(-\alpha_j - \beta_j) = 0$	$j = 1, \ldots, p$
	$\mu_j^+(-\alpha_j + \beta_j) = 0$	$j = 1, \ldots, p$
	$\gamma_j \alpha_j = 0$	$j = 1, \ldots, p$

The orthogonal case

$$\widehat{\beta}_j^{\mathsf{L}} = \begin{cases} \widehat{\beta}_j^{\mathsf{ls}} - \lambda sign(\widehat{\beta}_j^{\mathsf{ls}}) & 0 \leq \lambda \leq \widehat{\beta}_j^{\mathsf{ls}} \\ 0 & \text{else} \end{cases}$$

TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of $\mathbf{X}$. M and λ are constants chosen by the corresponding techniques;* sign *denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The 45° line in gray shows the unrestricted estimate for reference.*

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\mathrm{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



Lasso implementation as a QP with CVX

$$\begin{cases} \min_{\beta \in \mathbb{R}^p} & \|\mathbf{y} - X\beta\|^2 \\ \text{s.t.} & \|\beta\|_1 \leq t \end{cases}$$

Given X, $\mathbf{y}$ and $0 \leq t$

with R

```
b<-Variable(p)
o<-Minimize(sum((y-X%*%b)^2))
c<-list(sum(abs(b)) <=  t)
problem <- Problem(o, c)
result <- solve(problem)
```

with python

```
import cvxpy as cp

b = cp.Variable(p)
o = cp.Minimize(cp.sum_squares(X@b-y))
c = [cp.norm(b, 1) <= t]
problem = cp.Problem(o, c)
problem.solve()
print("Cost: ", problem.value)
print("Estimation: ", b.value)
```

It works as well with other formulations

http://cvxr.com/cvx/download/

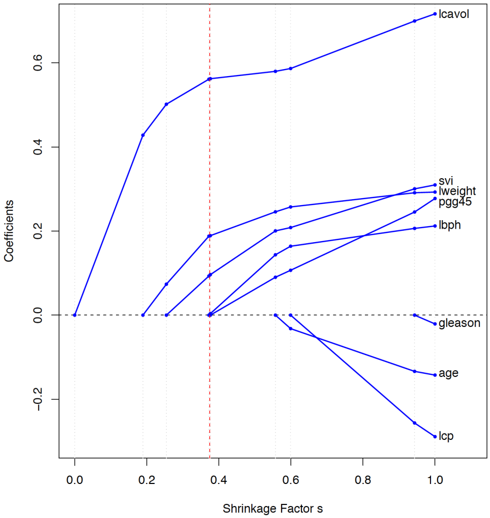
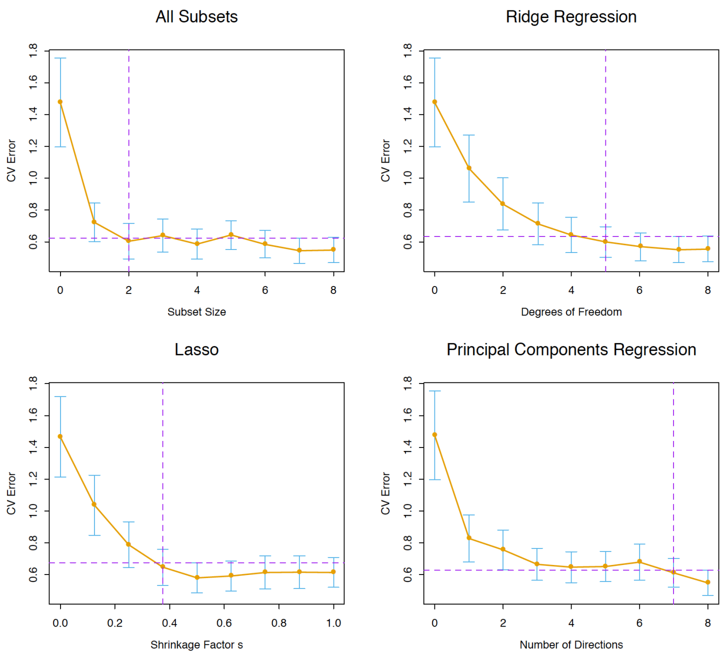


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*



All Subsets
Ridge Regression
Lasso
Principal Components Regression
CV Error
Subset Size
Degrees of Freedom
Shrinkage Factor s
Number of Directions
0.6
0.8
1.0
1.2
1.4
1.6
1.8
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0

Outline

Related regression methods

Ridge + Lasso = Elastic Net

$$\min_{\beta} \ \|\mathbf{y} - X\beta\|^2 + \lambda \left(\alpha\|\beta\|_2^2 + (1-\alpha)\ \|\beta\|_1\right)$$

The Dantzig Selector

$$\begin{cases} \min\limits_{\beta \in \mathbb{R}^p} \ \|\beta\|_1 \\ \text{s.t. } \|X^\top(\mathbf{y} - X\beta)\|_\infty \leq t \end{cases}$$

The adaptive Lasso

$$\min_{\beta} \ \|\mathbf{y} - X\beta\|^2 + \lambda \ \sum_{j=1}^{p} w_j|\beta_j|$$

The Grouped Lasso
The Fused Lasso
Non convex penalties: MCP and SCAD
...just to name a few

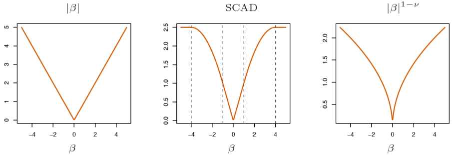


FIGURE 3.20. *The lasso and two alternative non-convex penalties designed to penalize large coefficients less. For SCAD we use $\lambda = 1$ and $a = 4$, and $\nu = \frac{1}{2}$ in the last panel.*

the L^q penalty

Generalize the Ridge (q=2) and the Lasso (q=1)

$$\min_{\beta} \|\mathbf{y} - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

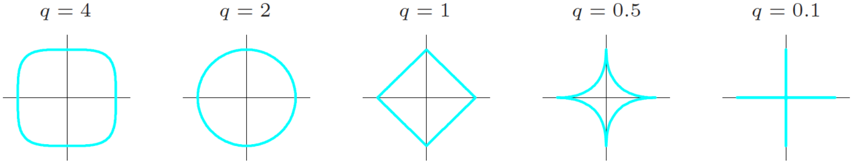
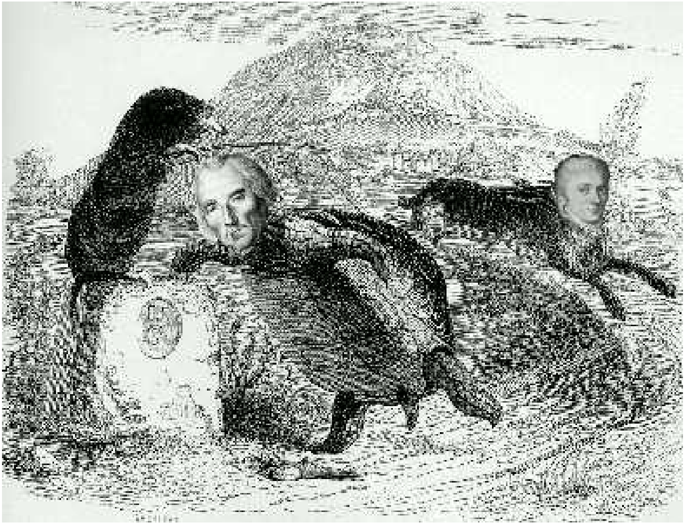


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

and the best subset for $q = 0$

Outline

The Gaussian Hare and the Laplacian Tortoise



Portnoy , Koenker, Statistical Science, 1997

The linear piecewise regularization path of the Lasso

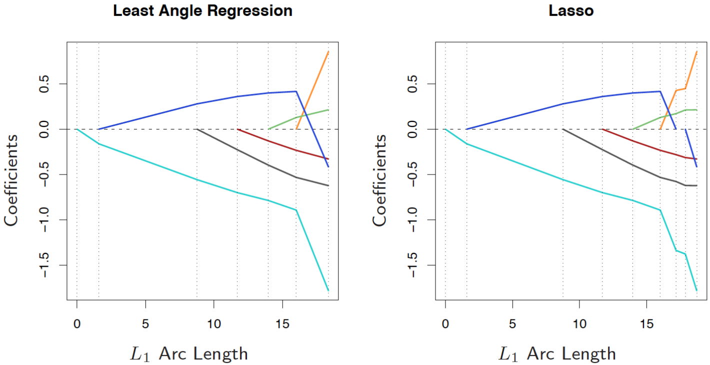


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.*

Outline

Principal component regression: PCR

Kendall (1957), Hotelling (1957)

let $u_1, u_2, \dots, u_p$ the p singular vectors of matrix X associated with singular values $d_1 \geq d_2 \geq \dots \geq d_p$ so that $X = UDV^\top$

perform a least square on the principal components:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - UDV^\top\boldsymbol{\beta}\|^2 = \min_{\gamma} \|\mathbf{y} - UD\gamma\|^2$$

with $\gamma = V^\top\beta$. Now consider only $k < p$ components the OLS

$$\hat{\gamma}_k = \operatorname*{argmin}_{\gamma} \|\mathbf{y} - U_k D_k \gamma\|^2$$

transform this vector back

$$\hat{\boldsymbol{\beta}}_k = V_k \hat{\gamma}_k$$

It is NOT scale invariant

the solution is a sequence (a path) $\beta_k, k = 1, \dots, p$

Partial least square: PLS

H. Wold (1975) puis S. Wold (1983)

Initially a procedure (an algorithm, NIPALS)

Algorithm 3.3 *Partial Least Squares.*

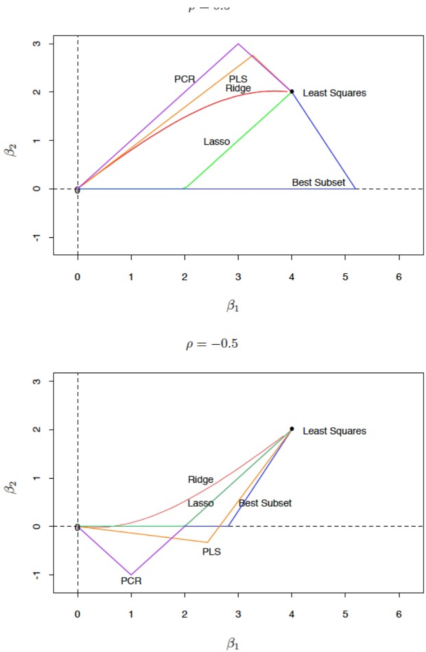
1. Standardize each $\mathbf{x}_j$ to have mean zero and variance one. Set $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$, and $\mathbf{x}_j^{(0)} = \mathbf{x}_j,\ j = 1, \dots, p$.
2. For $m = 1, 2, \dots, p$
   (a) $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\varphi}_{mj} \mathbf{x}_j^{(m-1)}$, where $\hat{\varphi}_{mj} = \langle \mathbf{x}_j^{(m-1)}, \mathbf{y} \rangle$.
   (b) $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$.
   (c) $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$.
   (d) Orthogonalize each $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m$: $\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - [\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle] \mathbf{z}_m,\ j = 1, 2, \dots, p$.
3. Output the sequence of fitted vectors $\{\hat{\mathbf{y}}^{(m)}\}_1^p$. Since the $\{\mathbf{z}_\ell\}_1^m$ are linear in the original $\mathbf{x}_j$, so is $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

Better regression direction by taking into account the response variable y

PCA direction	PLS direction
$\min_{\mathbf{v}} \ \|X\mathbf{v}\|^2$, $\ \|\mathbf{v}\| = 1,$	$\min_{\mathbf{v}} \ cov^2(\mathbf{y}, X\mathbf{v}) = (\mathbf{y}^\top X\mathbf{v})^2$, $\ \|\mathbf{v}\| = 1,$

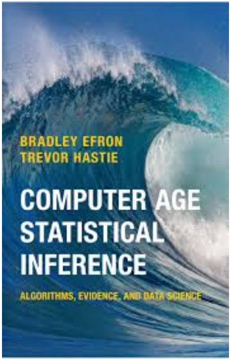
Very popular in Chemometrics (analytical chemistry)

Illustration: Regularization paths in 2d



Outline

Conclusions



- key issue: Statistics and optimization
- beware: hyperparameter tuning
- focus: details (preprocessing, normalization, intercept. . .)
- my choice: I use the adaptive lasso after variable screening
- my research: non convex and L_0 penalty (MIP optimization)