Statistical Learning

Academic year 2021/22
CLAMSES - University of Milano-Bicocca

Aldo Solari
aldo.solari@unimib.it

Prediction, Estimation, and Attribution

Statistical Learning

CLAMSES - University of Milano-Bicocca

Aldo Solari

References

This material reproduces the following

- Efron, B. (2020). Prediction, Estimation, and Attribution. Journal of the American Statistical Association, 115(530), 636-655. With Discussion and Rejoinder.
- Slides
- Recorded presentation for the 62nd ISI World Statistics Congress in Kuala Lumpur [46 mins]

Outline

1. Introduction
2. Surface Plus Noise Models
3. The Pure Prediction Algorithms
4. A Microarray Prediction Problem
5. Advantages and Disadvantages of Prediction
6. The Training/Test Set Paradigm
7. Smoothness
8. A Comparison Checklist
9. TraditionalMethods in the Wide Data Era
10. Two Hopeful Trends

Regression

Gauss (1809), Galton (1877)

- *Prediction: the prediction of new cases*
  e.g. random forests, boosting, support vector machines, neural nets, deep learning
- *Estimation: the estimation of regression surfaces*
  e.g. OLS, logistic regression, GLM (MLE)
- *Attribution: the assignment of significance to individual predictors*
  e.g. Fisher's ANOVA, Neyman-Pearson

How do the pure prediction algorithms relate to traditional regression methods?

That is the central question pursued in what follows.

2. Surface Plus Noise Models

We will assume that the data D available to the statistician has this structure:

$$D = \{(x_i, y_i), i = 1, \ldots, n\}$$

- x_i is a p-dimensional vector of predictors taking its value in a known space $\mathcal{X}$ contained in $\mathbb{R}^p$;
- y_i is a real valued response;
- the n pairs are assumed to be independent of each other.

More concisely we can write

$$D = \{X, y\}$$

where X is the $n \times p$ matrix having x_i^t as the ith row, and $y = (y_1, \ldots, y_n)^t$.

Regression surface

- The regression model is

$$y_i = s(x_i, \beta) + \epsilon_i \quad i = 1, \ldots, n \tag{1}$$

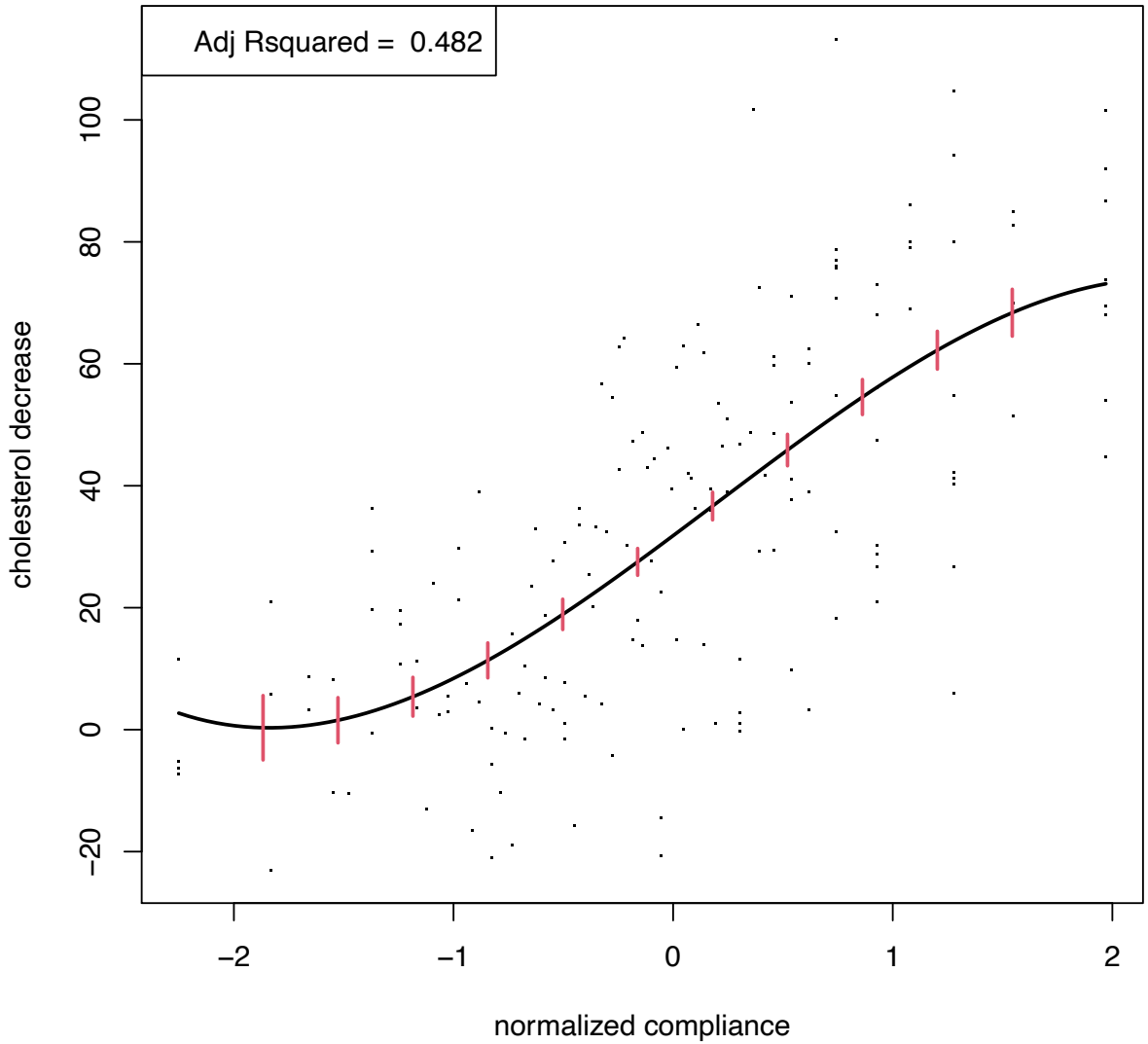
  $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ where $s(x, \beta)$ is some functional form that, for any fixed value of the parameter vector β, gives expectation $\mu = s(x, \beta)$ as a function of $x \in \mathcal{X}$;
- The *regression surface* is

$$\mathcal{S} = \{s(x, \beta), x \in \mathcal{X}\}$$

  Most traditional regression methods depend on some sort of surface plus noise formulation;
- The surface describes the scientific truths we wish to learn, but we can only observe points on the surface obscured by noise;
- The statistician's traditional estimation task is to learn as much as possible about the surface from the data D.

Cholesterol data

- Cholestyramine, a proposed cholesterol lowering drug, was administered to 164 male doctors for an average of seven years each (Efron and Feldman, 1991)
- The response variable (y_i) is a man's decrease in cholesterol level over the course of the experiment.
- The single predictor is compliance (x_i), the fraction of intended dose actually taken. Compliance, the proportion of the intended dose actually taken, ranged from 0% to 100%, −2.25 to 1.97 on the normalized scale. It was hoped to see larger cholesterol decreases for the better compliers.
- https://hastie.su.domains/CASI_files/DATA/cholesterol.html


Adj Rsquared = 0.482
cholesterol decrease
100
80
60
40
20
0
−20
−2
−1
0
1
2
normalized compliance

- A normal regression model was fit, with

$$s(x_i, \beta) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$$

  in other words, a cubic regression model.
- The black curve is the estimated surface

$$\hat{\mathcal{S}} = \{s(x, \hat{\beta}), x \in \mathcal{X}\}$$

  fit by maximum likelihood or, equivalently, by ordinary least squares (OLS).
- The vertical bars indicate one standard error for the estimated values $s(x, \hat{\beta})$, at 11 choices of x, showing how inaccurate $\hat{\mathcal{S}}$ might be as an estimate of the true $\mathcal{S}$
- Only $\hat{\beta}_0$ and $\hat{\beta}_1$ were significantly nonzero. The adjusted R^2 was 0.482, a traditional measure of the model's predictive power.

`birthwt` data

- R package `MASS`
- The birthwt data frame has 189 rows and 10 columns.
- The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

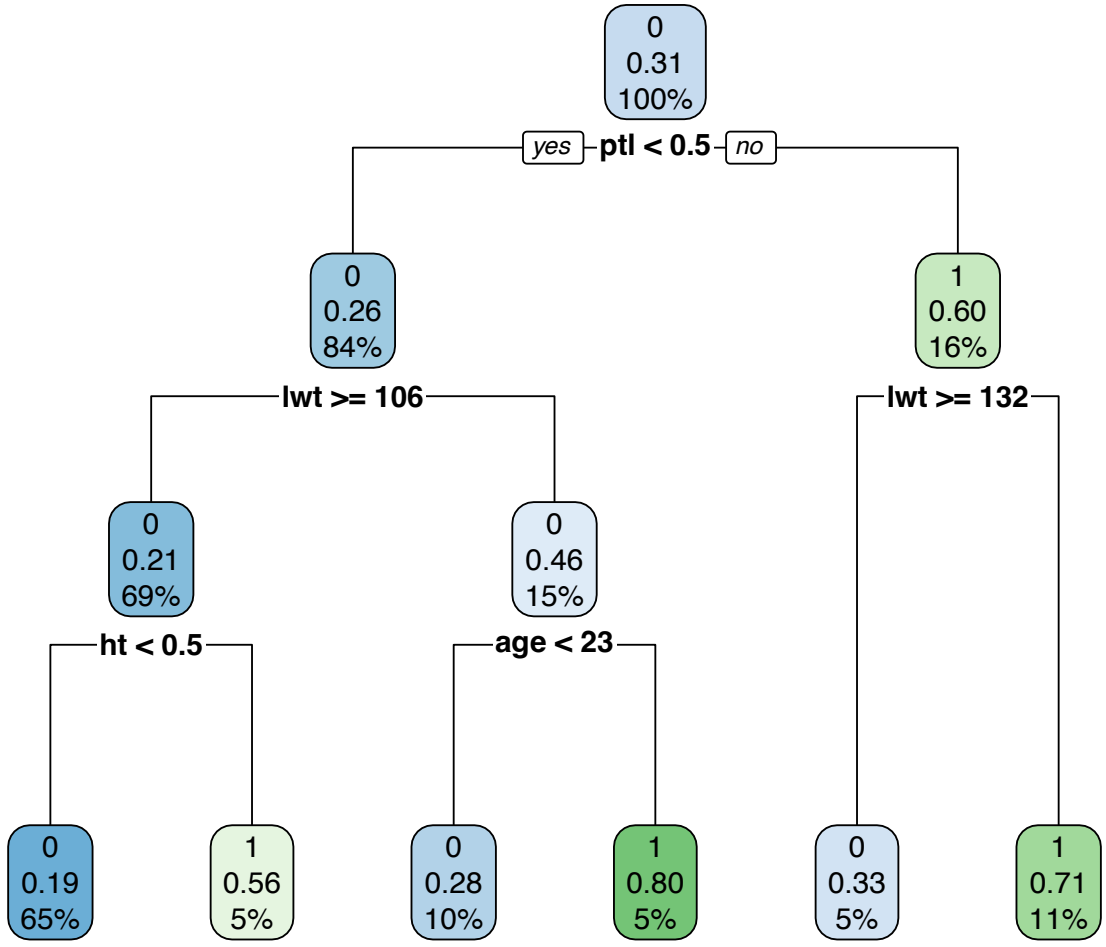
- Another mainstay of traditional methodology is logistic regression.
- The dataset concerns the Risk Factors Associated with Low Infant Birth Weight: $n = 189$ babies, 59 with birth weight less than 2.5 kg and 130 with more than 2.5 kg.
- Eight covariates were measured at entry: mother's age in years, mother's weight in pounds at last menstrual period, body weight, etc., so x_i was 8-dimensional, while y_i equaled 0 or 1
- This is a surface plus noise model, with a linear logistic surface and Bernoulli noise.

	term	estimate	std.error	p.value	
1	(Intercept)	1.07	1.27	0.40	
2	age	-0.04	0.04	0.31	
3	lwt	-0.02	0.01	0.02	*
4	raceblack	1.12	0.54	0.04	*
5	raceother	0.67	0.47	0.16	
6	smoke	0.75	0.43	0.08	
7	ptl	-1.66	0.90	0.07	
8	ht	1.93	0.73	0.01	**
9	ui	0.80	0.48	0.09	
10	ftv1	-0.52	0.49	0.29	
11	ftv2+	0.10	0.46	0.83	
12	ptd	3.41	1.22	0.01	**

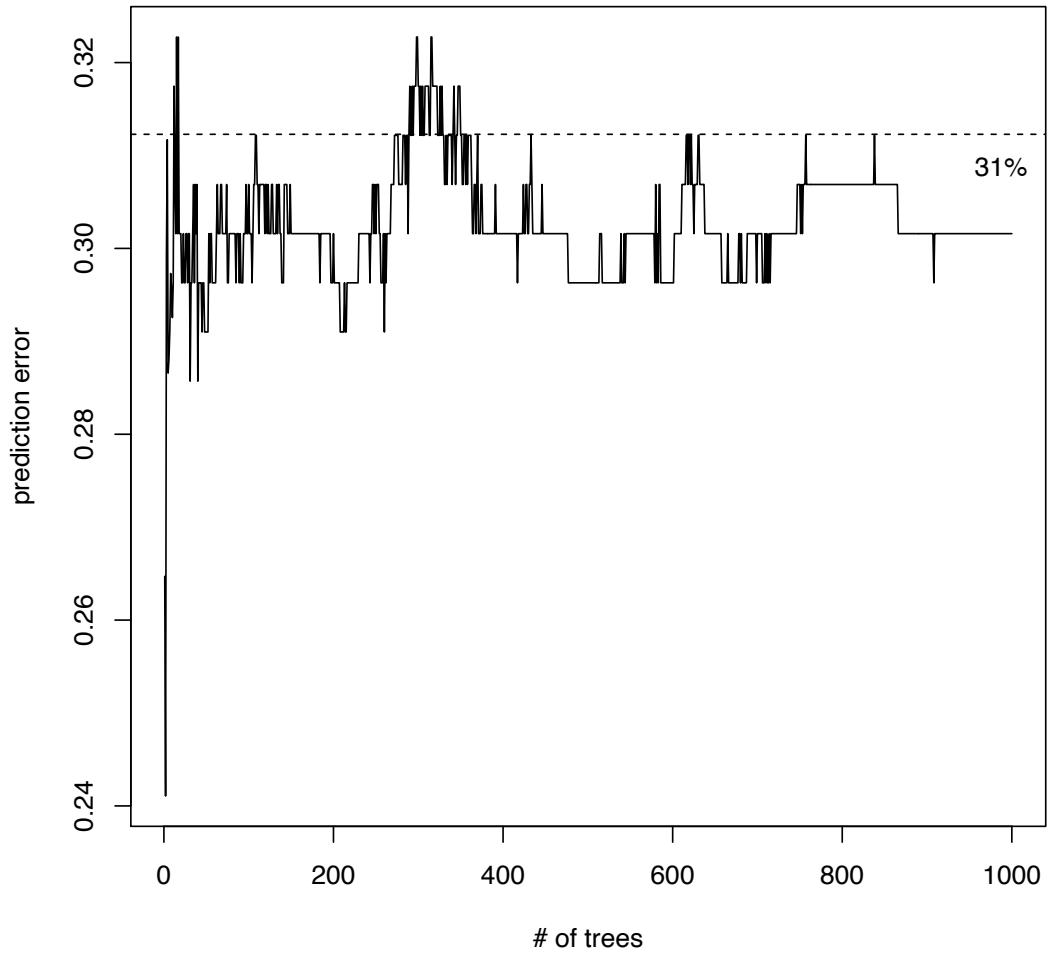
3. The Pure Prediction Algorithms

- Random Forests, Boosting, Deep Learning, etc.
- Data
$$D = \{(x_i, y_i), i = 1, \ldots, n\}$$
- Prediction rule $f(x, D)$
- New $(x, ?)$ gives $\hat{y} = f(x, D)$
- Strategy: Go directly for high predictive accuracy; forget (mostly) about surface + noise

CART


0
0.31
100%
yes
ptl < 0.5
no
0
0.26
84%
1
0.60
16%
lwt >= 106
lwt >= 132
0
0.21
69%
0
0.46
15%
ht < 0.5
age < 23
0
0.19
65%
1
0.56
5%
0
0.28
10%
1
0.80
5%
0
0.33
5%
1
0.71
11%

Random forest



OOB estimates of prediction error as a function of number of bootstrapped trees; dashed line is the 10-fold CV error rate (31%).

Apparent error rate (training error)

$$\widehat{\text{err}} = \#\{f(x_i, D) \neq y_i\}/n$$

	model	error rate
1	rand_forest	0.222
2	logistic_reg	0.243
3	decision_tree	0.228

True error rate

$$\mathbb{E}(f(X, D) \neq Y)$$

where (X, Y) is a random draw from whatever probability distribution gave the (x_i, y_i) pairs in D;

Estimated by 10-fold cross-validated error rate

	model	mean	n	std_err
1	rand_forest	0.312	10	0.04
2	logistic_reg	0.313	10	0.03
3	decision_tree	0.365	10	0.04

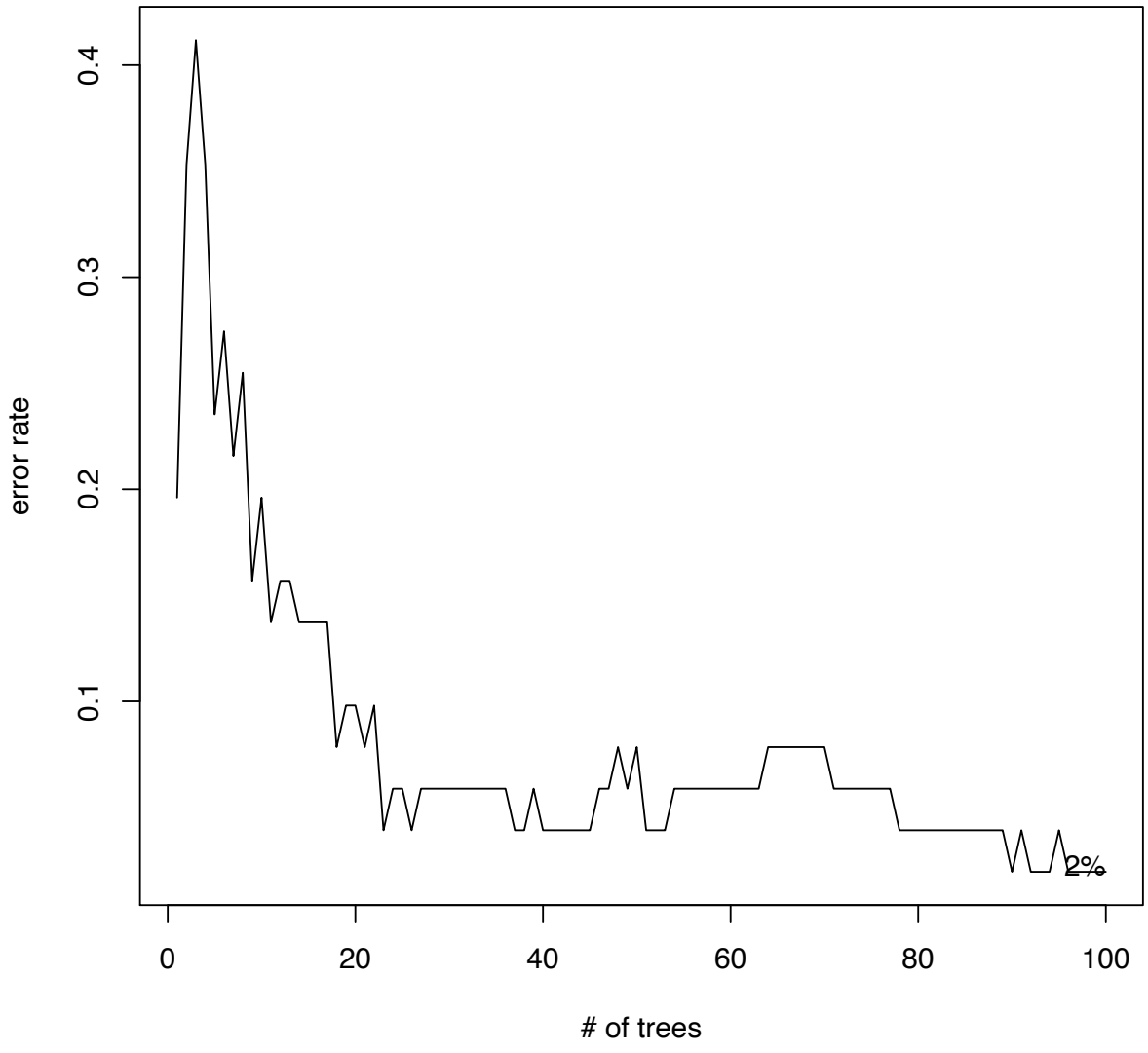
4. A Microarray Prediction Problem

The Prostate Cancer Microarray Study

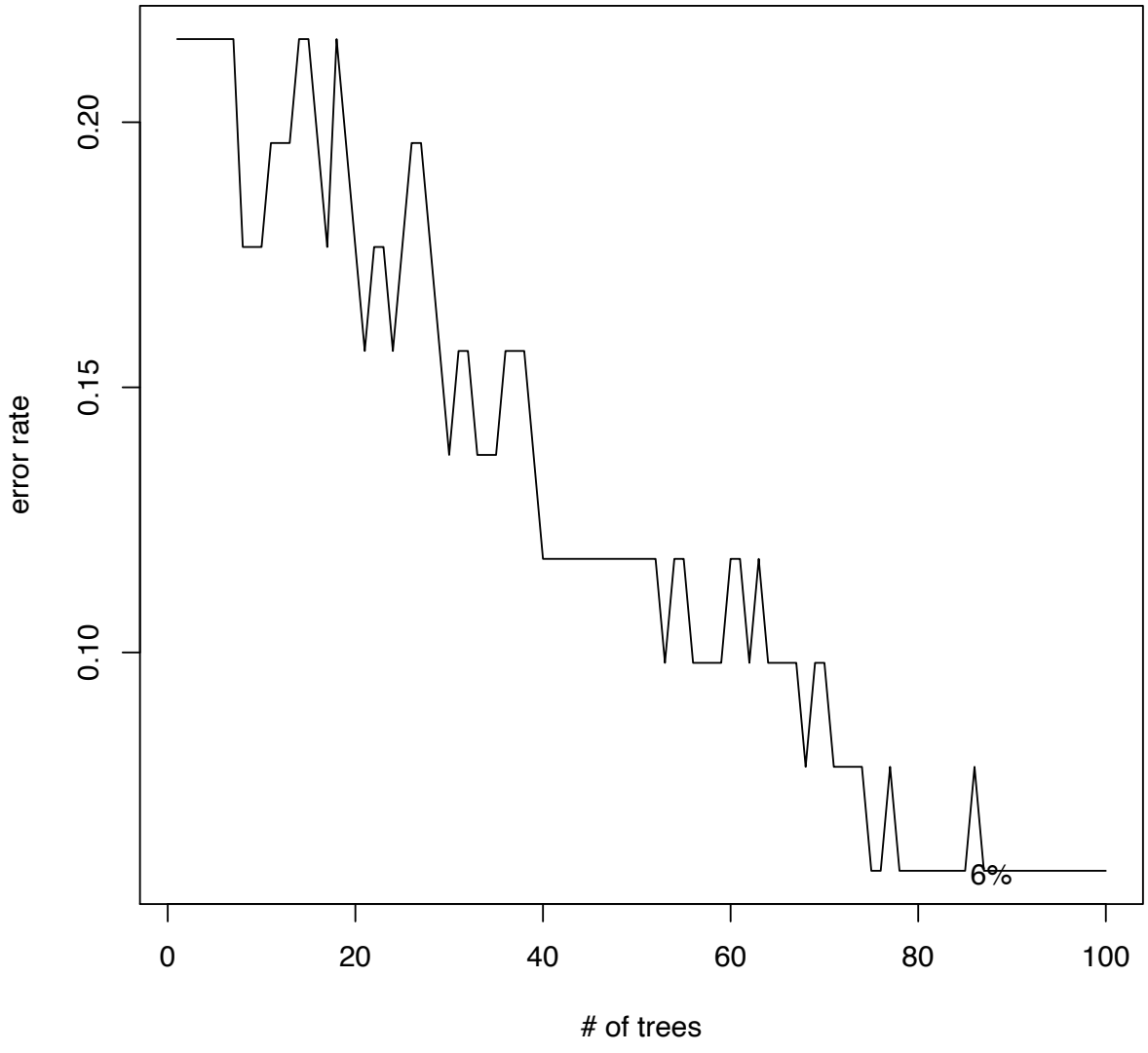
- https://hastie.su.domains/CASI_files/DATA/prostate.html
- $n = 102$ men: 52 prostate cancer, 50 normal controls
- For each man measure activity of $p = 6033$ genes
- Data set D is 102×6033 matrix ("wide")
- Wanted: Prediction rule $f(x, D)$ that inputs new 6033-vector x and outputs $\hat{y}$ correctly predicting cancer/normal

Random forest

- Randomly divide the 102 subjects into:
  - training set of 51 subjects (26 + 25)
  - test set of 51 subjects (26 + 25)
- Run R program `randomForest` on the training set
- Use its rule $f(x_i, D)$ on the test set and see how many errors it makes


error rate
0.1
0.2
0.3
0.4
0
20
40
60
80
100
# of trees
2%

Boosting



5. Advantages and Disadvantages of Prediction

Prediction is Easier than Estimation

- Observe
  - $x_1, \ldots, x_{25} \overset{\text{ind}}{\sim} \mathcal{N}(\mu, 1)$
  - $\bar{x}$ sample mean, $\tilde{x}$ sample median,
- Estimation

$$E\{(\tilde{x} - \mu)^2\}/E\{(\bar{x} - \mu)^2\} = 1.57$$

- Wish to predict new $X \sim \mathcal{N}(\mu, 1)$
- Prediction

$$E\{(\tilde{x} - X)^2\}/E\{(\bar{x} - X)^2\} = 1.02$$

- The reason is that most of the prediction error comes from the variability of X, which neither $\bar{x}$ or $\tilde{x}$ can cure
- Prediction is easier than estimation, at least in the sense of being more forgiving. This allows for the use of inefficient estimators like the `gbm` stumps

Prediction is Easier than Attribution

- Microarray study involving n subjects, $n/2$ healthy controls and $n/2$ sick patients
- Each subject provides a vector of measurements on N genes $\mathbf{x} = (x_1, \ldots, x_N)^t$ with

$$X_j \overset{\text{ind}}{\sim} \mathcal{N}(\pm\delta_j/2c, 1)$$

  where $c = \sqrt{(n/4)}$ "plus" for the sick and "minus" for the healthy; δ_j is the effect size for gene j.
- N_0 genes are null i.e. $\delta_j = 0$
- a small number N_1 of genes are non-null and have $\delta_j = \Delta$
- A new person arrives and produces a microarray of measurements $\mathbf{x} = (x_1, \ldots, x_N)^t$ but without us knowing the person's healthy/sick status; that is, without knowledge of the $\pm$ value
- Question: How small can N_1/N_0 get before prediction becomes impossible?

Prediction is Easier than Attribution

- Asymptotically as $N_0 \to \infty$, accurate prediction is possible if

$$N_1 = O(N_0^{1/2})$$

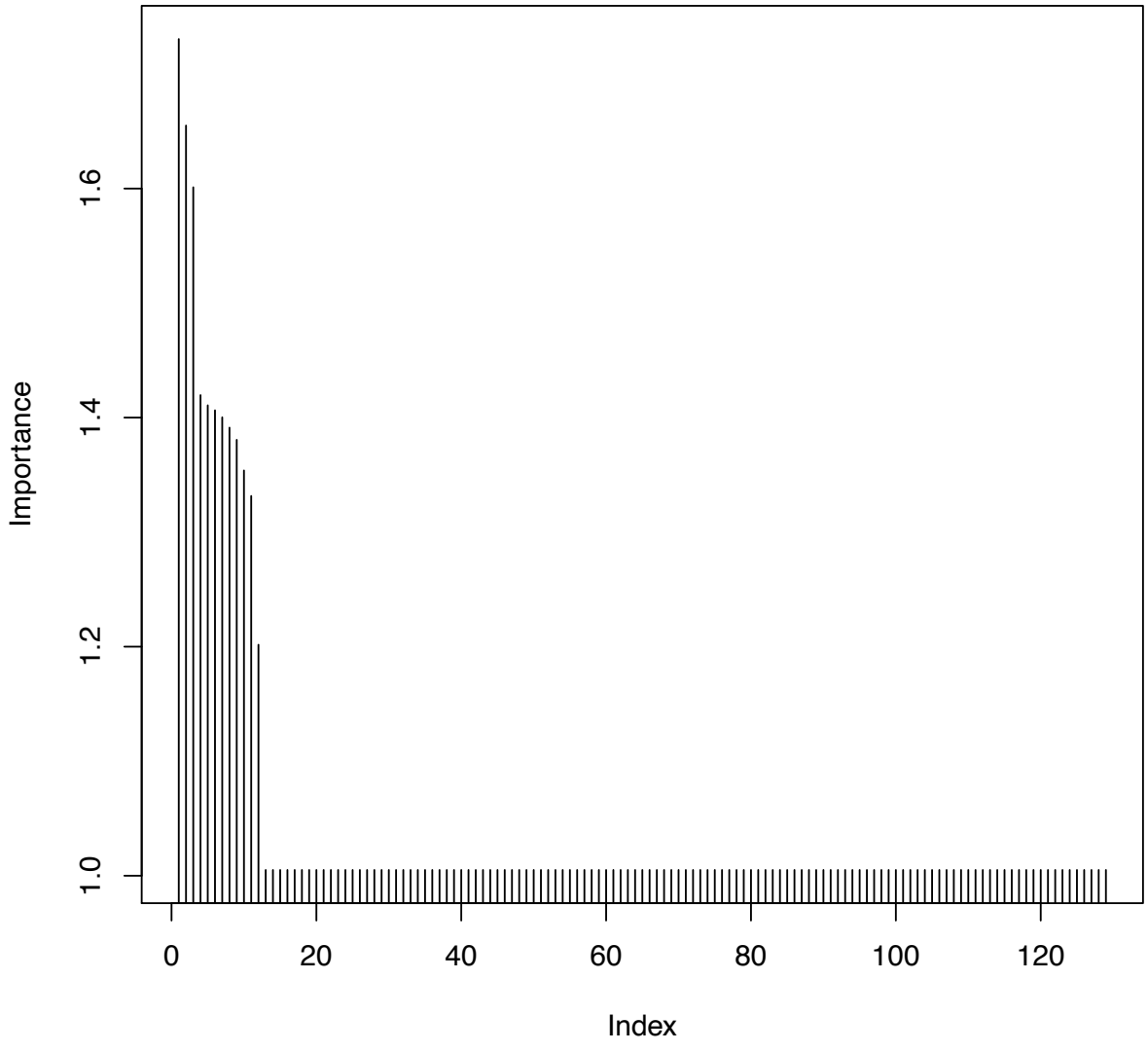
- Effective attribution requires

$$N_1 = O(N_0)$$

- In terms of "needles in haystacks", attribution needs an order of magnitude more needles than prediction.
- The three main regression categories can usefully be arranged in order

$$\text{prediction} \prec \text{estimation} \prec \text{attribution}$$

Variable importance

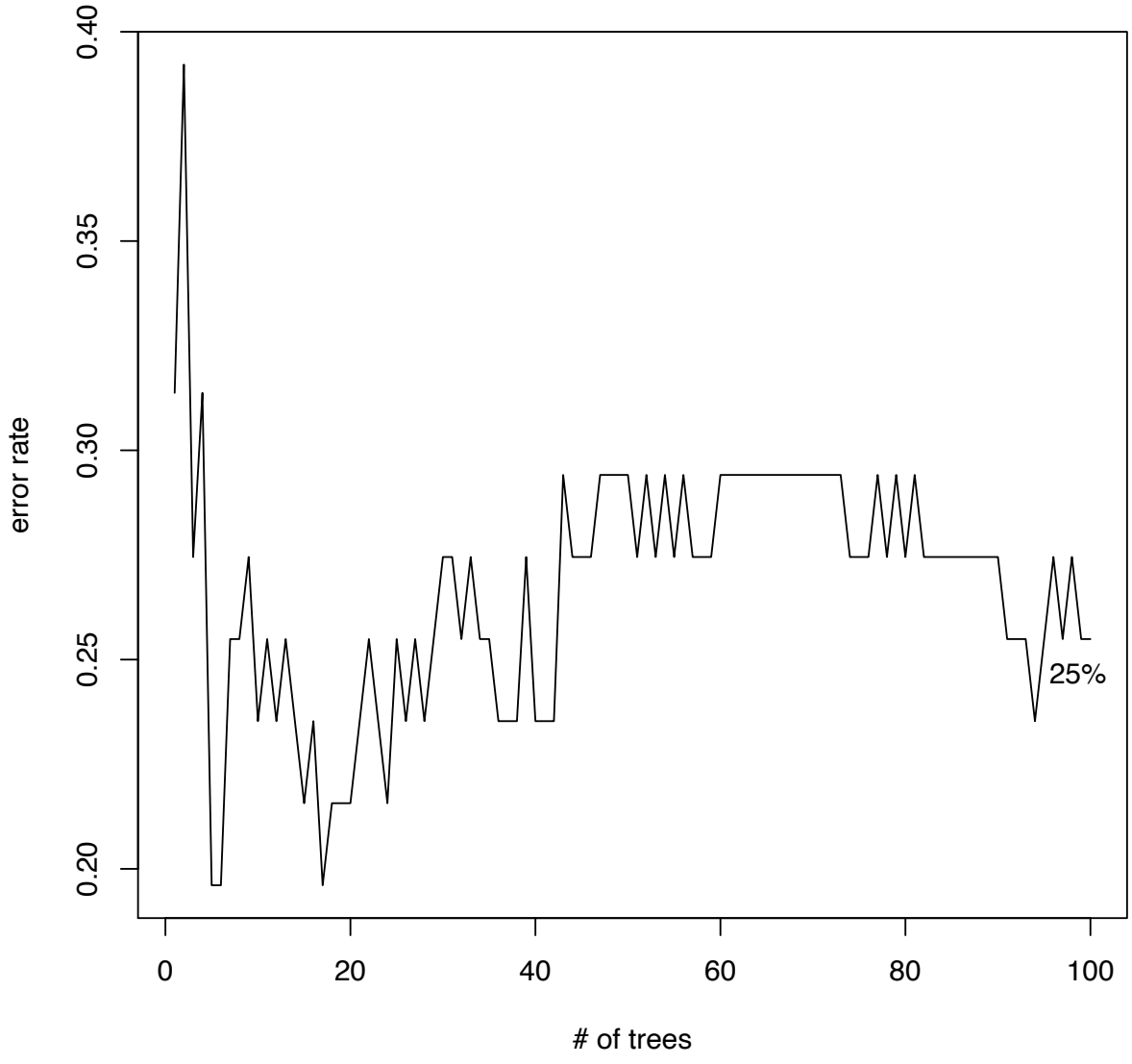


- Importance measure is computed for each of the p predictor variables.
- Of the $p = 6033$ genes, 129 had positive scores, these being the genes that ever were chosen as splitting variables.
- Can we use the importance scores for attribution?
- The answer seems to be no. Removing the most important 100 had similarly minor effects on the number of test set prediction errors
- Evidently there are a great many genes weakly correlated with prostate cancer, which can be combined in different combinations to give near-perfect predictions.

6. The Training/Test Set Paradigm

Were the Test Sets Really a Good Test?

- Prediction can be highly context-dependent and fragile
- Before Randomly divided subjects into training and test
- Next:
  - 51 earliest subjects for training (25 control + 26 cancer with lowest ID numbers)
  - 51 latest subjects for test
- Study subjects might have been collected in the order listed, with some small methodological differences creeping in as time progressed (concept drift)


error rate
0.20
0.25
0.30
0.35
0.40
0
20
40
60
80
100
# of trees
25%

Hypothetical microarray study

- $n = 400$ subjects participate in the study, arriving one per day in alternation between Treatment and Control
- Each subject is measured on a microarray of $p = 200$ genes
- The 400×200 data matrix X has independent normal entries

$$X_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{ij}, 1)$$

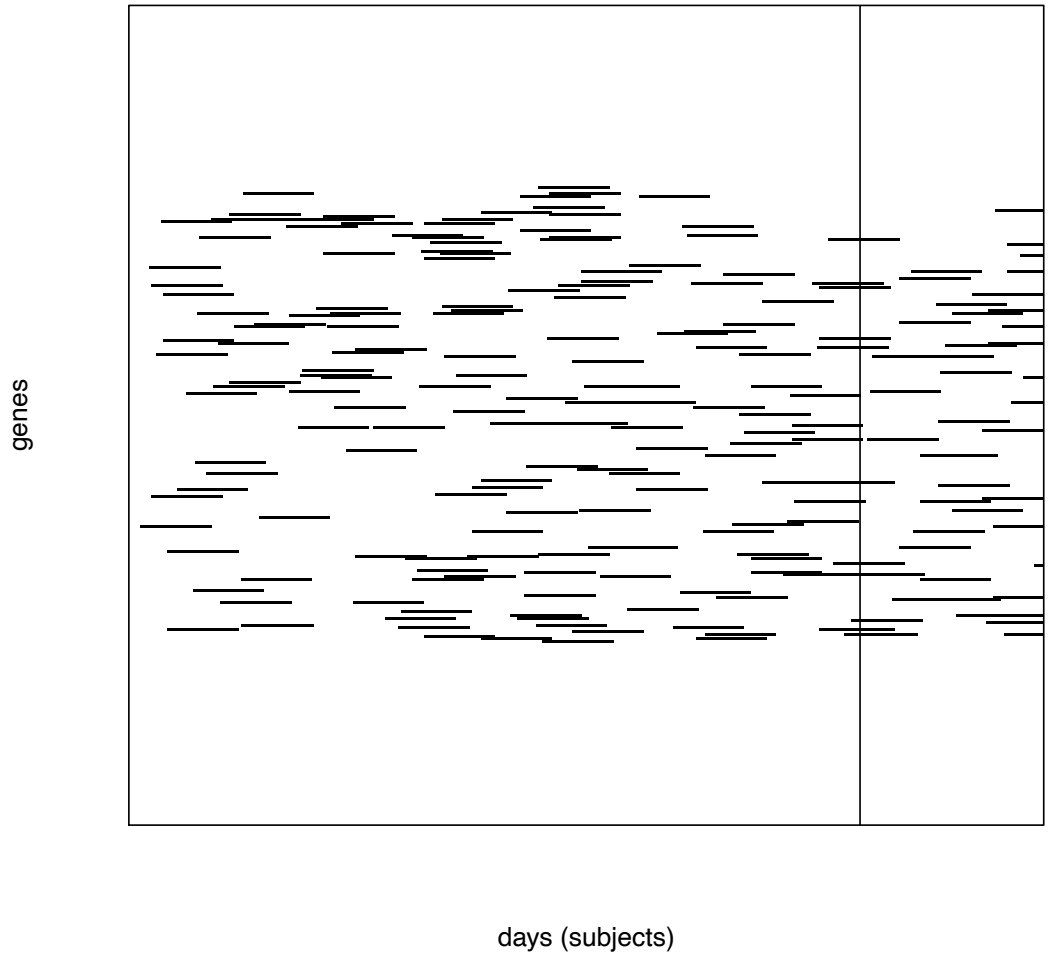
- Most of the μ_{ij} are null, $\mu_{ij} = 0$, but occasionally a gene will have an active episode of 30 days during which

$$\mu_{ij} = 2 \quad \text{for Treatment} \quad \mu_{ij} = -2 \quad \text{for Control}$$

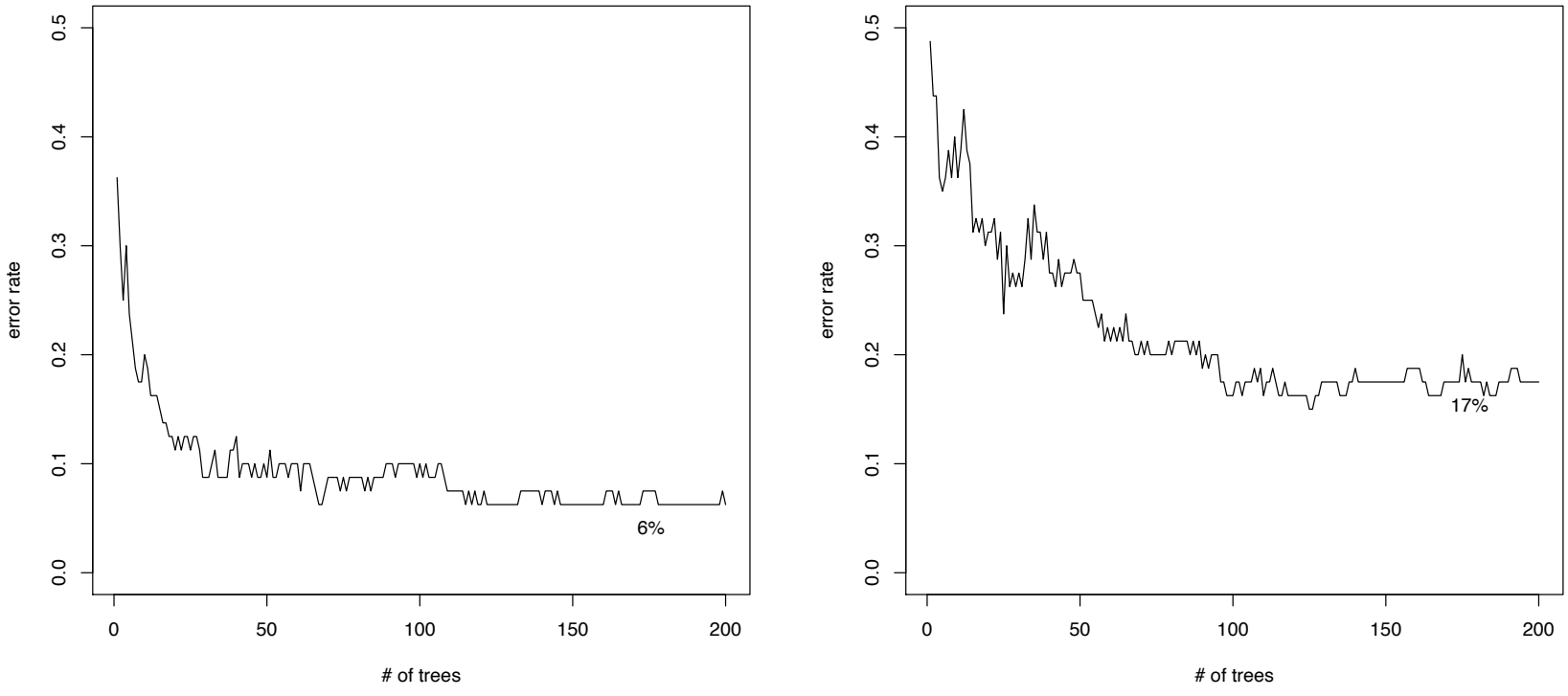
  for the entire episode, or

$$\mu_{ij} = -2 \quad \text{for Treatment} \quad \mu_{ij} = 2 \quad \text{for Control}$$

  for the entire episode.
- Each gene has expected number of episodes equal 1



Black line segments indicate active episodes in the hypothetical microarray study. (Matrix transposed for typographical convenience.)



randomForest prediction applied to the hypothetical microarray study microarray study. Left panel: Test set of size 80, selected randomly from 400 days; Right panel: Test set days 321-400

- From any one day's measurements it is possible to predict Treatment or Control from the active episode responses on nearby days
- This works for the random training/test division, where most of the test days will be intermixed with training days.
- Not so for the early/late division, where most of the test days are far removed from training set episodes.
- To put it another way, prediction is easier for interpolation than extrapolation

Replicability

Year (study) 1: $n = 812$, $p = 11$ (selected from an initial list of 81)

Table 1. Logistic regression analysis of neonate data.

	Estimate	SE	*p*-value
Intercept	−1.549	0.457	0.001***
gest	−0.474	0.163	0.004**
ap	−0.583	0.110	0.000***
bwei	−0.488	0.163	0.003**
resp	0.784	0.140	0.000***
cpap	0.271	0.122	0.027*
ment	1.105	0.271	0.000***
rate	−0.089	0.176	0.612
hr	0.013	0.108	0.905
head	0.103	0.111	0.355
gen	−0.001	0.109	0.994
temp	0.015	0.124	0.905

NOTE: Significant two-sided *p*-values indicated for 6 of 11 predictors; estimated logistic regression made 18% prediction errors.

Year (study) 2: $n = 246$, $p = 11$

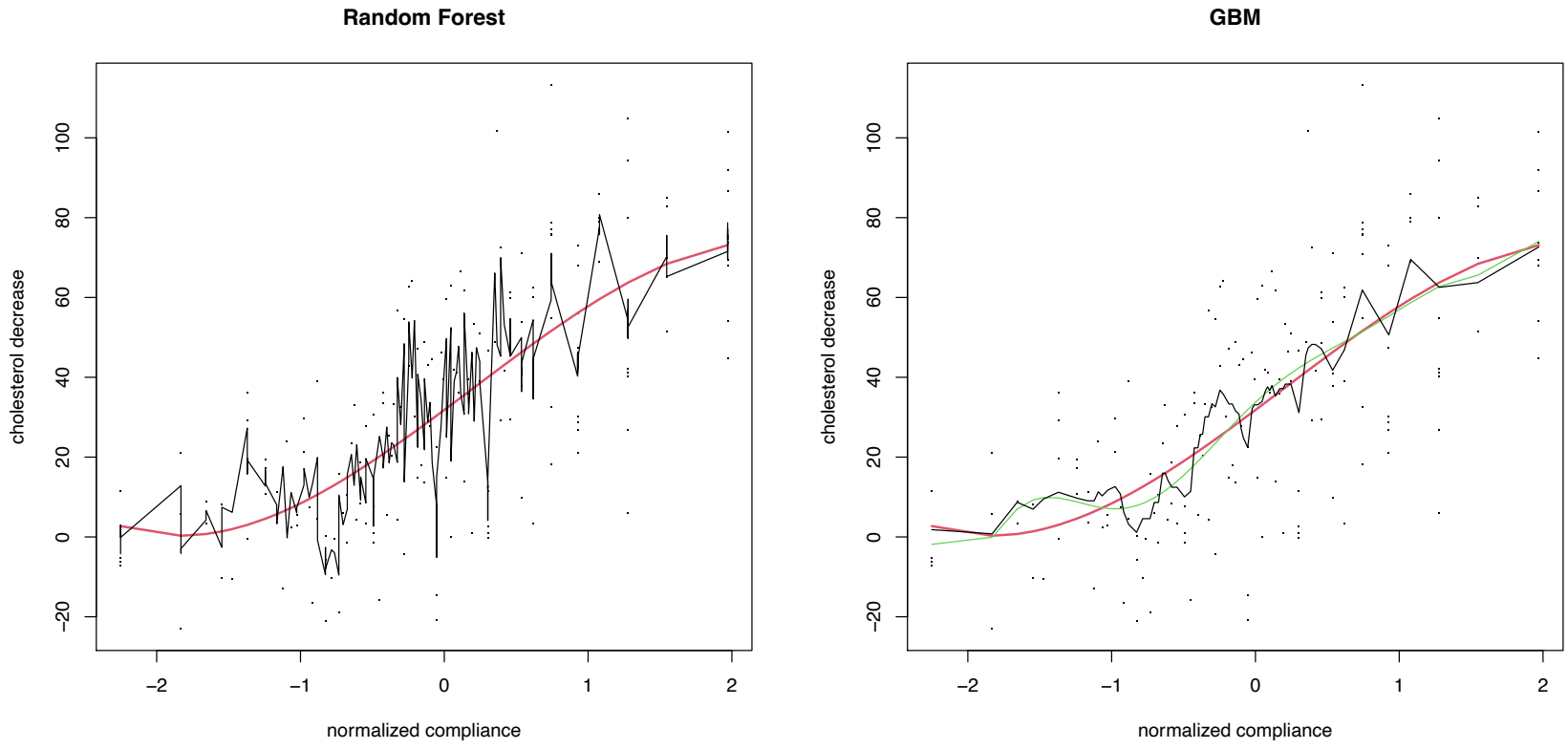
Table 4. Comparing logistic regression coefficients for neonate data for year 1 (as in Table 1) and year 2; correlation coefficient 0.79.

	gest	ap	bwei	resp	cpap	ment	rate	hr	head	gen	temp
Year 1	−0.47	−0.58	−0.49	0.78	0.27	1.10	−0.09	0.01	0.1	0.00	0.02
Year 2	−0.65	−0.27	−0.19	1.13	0.15	0.41	−0.47	−0.02	−0.2	−0.04	0.16

7. Smoothness

- Estimation and Attribution: seek long-lasting scientific truths
  - physics
  - astronomy
  - medicine
  - economics?
- Prediction algorithms: truths and ephemeral relationships
  - credit scores
  - movie recommendations
  - image recognition
- Estimation and Attribution: theoretical optimality (MLE, Neyman-Pearson)
- Prediction: training-test performance
- Nature: rough or smooth?

- The parametric models of traditional statistical methodology enforce the smooth-world paradigm
- Looking back at the Cholesterol data, we might not agree with the exact shape of the cholostyramine cubic regression curve but the smoothness of the response seems unarguable
- The choice of cubic was made on the basis of a Cp comparison of polynomial regressions degrees 1 through 8, with cubic best.
- Smoothness of response is not built into the pure prediction algorithms.
- Random forest and algorithm `gbm` take X to be the 164×8 matrix `poly(c,8)` - an 8th degree polynomial basis



`randomForest` and `gbm` fits to the Cholesterol data. Heavy red curve is cubic OLS; dashed green curve in right panel is 8th degree OLS fit.

8. A Comparison Checklist

	Traditional regressions methods	Pure prediction algorithms
1.	Surface plus noise models (continuous, smooth)	Direct prediction (possibly discrete, jagged)
2.	Scientific truth (long-term)	Empirical prediction accuracy (possibly short-term)
3.	Parametric modeling (causality)	Nonparametric (black box)
4.	Parsimonious modeling (researchers choose covariates)	Anti-parsimony (algorithm chooses predictors)
5.	X $n \times p$ with $p \ll n$ (homogeneous data)	$p \gg n$, both possibly enormous (mixed data)
6.	Theory of optimal inference (mle, Neyman–Pearson)	Training/test paradigm (Common Task Framework)

9. Traditional Methods in the Wide Data Era

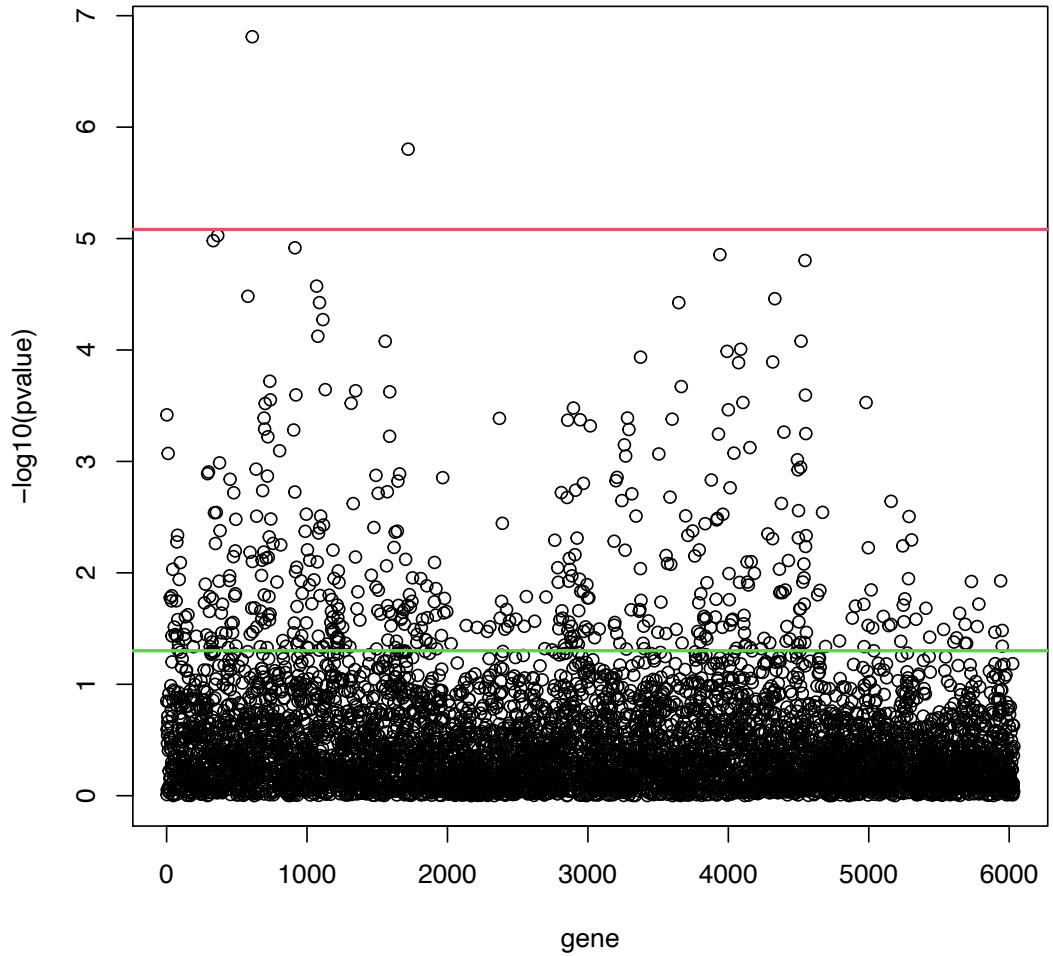
Estimation and Attribution in the Wide-Data Era

- Large p (the number of features) affects Estimation
  - MLE can be badly biased for individual parameters
  - "surface" if, say, $p = 6033$?
- Attribution still of interest. Compute p-value p_i for the null hypothesis H_i: no difference in gene expression between cancer and control at the ith gene
- The Bonferroni threshold for 0.05 significance is

$$p_i \leq 0.05/6033$$

$$\begin{aligned} \Pr(\text{Type I error} > 0) &= \Pr\left(\bigcup_{i \in I_0} \{p_i \leq \alpha/p\}\right) \\ &\leq \sum_{i \in I_0} \mathrm{P}(p_i \leq \alpha/p) \leq |I_0| \frac{\alpha}{p} \leq \alpha \end{aligned}$$

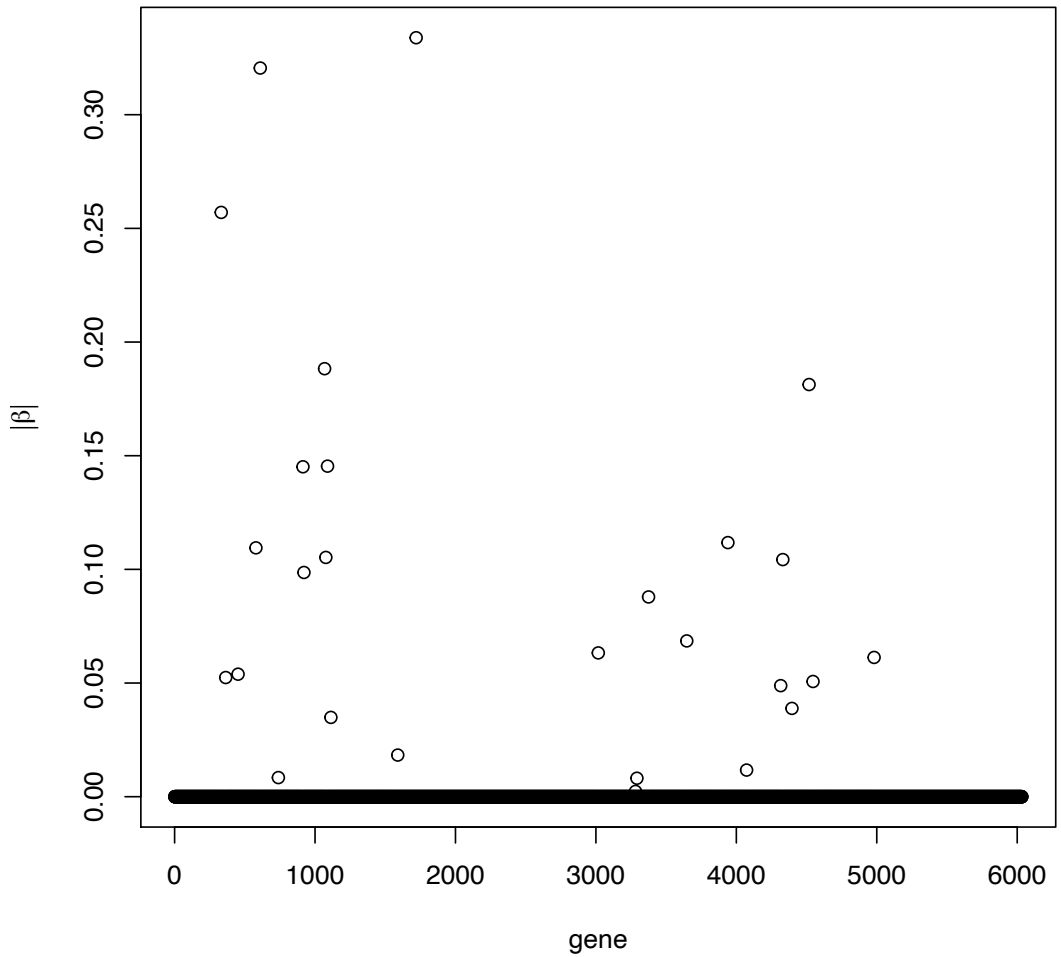
- Instead of performing a traditional attribution analysis with $p = 6033$ predictors, a microarray analysis performs 6033 analyses with $p = 1$


−log10(pvalue)
0
1
2
3
4
5
6
7
0
1000
2000
3000
4000
5000
6000
gene

- Sparsity offers another approach to wide-data estimation and attribution: we assume that most of the p predictor variables have no effect and concentrate effort on finding the few important ones.
- The lasso provides a key methodology. Estimate β, the p-vector of regression coefficients, by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^t\beta) + \lambda\|\beta\|_1$$

  where $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$
- Here λ is a fixed tuning parameter: $\lambda = 0$ corresponds to the OLS solution for β (if $p \leq n$) while $\lambda = \infty$ makes $\hat{\beta} = 0$. For large values of λ only a few of the coordinates $\hat{\beta}_j$ will be nonzero.
- The lasso produced biased estimates of β, with the coordinate values $\hat{\beta}_j$ shrunk toward zero.


|β|
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
1000
2000
3000
4000
5000
6000
gene

10. Two Hopeful Trends

- Making prediction algorithms better for scientific use
  - smoother
  - more interpretable
- Making traditional estimation/attribution methods better for large-scale (n, p) problems
  - more flexible
  - better scaled
- We do have optimality theory for estimation (MLE) and attribution (Neyman-Pearson), but we do not have an optimality theory for prediction.

James-Stein estimation

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Samworth (2012). Stein's paradox. eureka, 62:38–41
- Candés (2022) Lecture notes (Stats 300C - Theory of Statistics)

- A very surprising result arises in a remarkably simple estimation problem.
- Let $X_1, \ldots, X_p$ be independent random variables, with $X_i \sim N(\mu_i, 1)$ for $i = 1, \ldots, p$. Writing $X = (X_1, \ldots, X_p)^t$, suppose we want to find a good estimator $\hat{\mu} = \hat{\mu}(X)$ of $\mu = (\mu_1, \ldots, \mu_p)^t$
- Squared error loss function:

$$L(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|^2 = \sum_{i=1}^{p} (\hat{\mu}_i - \mu_i)^2$$

  where $\|\cdot\|$ denotes the Euclidean norm
- Risk function: $R(\hat{\mu}, \mu) = \mathbb{E}[L(\hat{\mu}, \mu)]$

Inadmissible estimators

- If $\hat{\mu}$ and $\tilde{\mu}$ are both estimators of μ, we say that $\hat{\mu}$ strictly dominates $\tilde{\mu}$ if $R(\hat{\mu}, \mu) \leq R(\tilde{\mu}, \mu)$ for all μ, with strict inequality for some value of μ. In this case we say that $\tilde{\mu}$ is *inadmissible*.
- If $\hat{\mu}$ is not strictly dominated by any estimator of μ, it is said to be admissible. Note that admissible estimators are not necessarily sensible: for $p = 1$, the estimator $\hat{\mu} = 37$ (which ignores the data!) is *admissible*.
- On the other hand decision theory dictates that inadmissible estimators can be discarded
- $\hat{\mu} = X$ is a very obvious estimator of μ: it is the maximum likelihood estimator and the uniform minimum variance unbiased estimator with

$$R(\hat{\mu}, \mu) = p \quad \forall\ \mu \in \mathbb{R}^p$$

since $\|X - \mu\|^2 \sim \chi^2_p$

James-Stein estimator

- It has been proved that $\hat{\mu} = X$ is admissible for $p = 1, 2$
- James and Stein (1961) showed that the estimator

$$\hat{\mu}_{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right) X$$

strictly dominates $\hat{\mu} = X$ for $p \geq 3$:

$$R(\hat{\mu}_{JS}, \mu) = p - (p-2)^2 \mathbb{E}\left(\frac{1}{\|X\|^2}\right) < p \quad \forall\ \mu \in \mathbb{R}^p$$

$\|X\|^2 = \sum_{i=1}^p X_i^2$ follows a noncentral χ^2 distribution with p degrees of freedom and noncentrality parameter $\|\mu\|^2$. Using a result about noncentral χ^2 variables, we can write

$$\|X\|^2 \sim \chi^2_{p+2K}$$

where $K \sim \text{Poisson}(\|\mu\|^2/2)$.

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{\|X\|^2}\right) &= \mathbb{E}\left(\frac{1}{\chi^2_{p+2K}}\right) = \mathbb{E}\left\{\mathbb{E}\left(\frac{1}{\chi^2_{p+2K}}\right)|K\right\} \\
&= \mathbb{E}\left\{\frac{1}{(p-2)+2K}\right\} \geq \frac{1}{(p-2)+\|\mu\|^2}
\end{aligned}$$

with equality if $\mu = 0$, where we used $\mathbb{E}(1/\chi^2_p) = 1/(p-2)$ for $p > 2$ and Jensen's inequality. Then

$$R(\hat{\mu}_{JS}, \mu) \leq p - \frac{p-2}{1+\|\mu\|^2/(p-2)}$$

Oracle linear estimator

- A linear estimator of the form

$$\tilde{\mu} = bX = (bX_1, \ldots, bX_p)^t$$

  with $0 \leq b \leq 1$ shrinks X towards the origin
- The risk of a linear estimator is

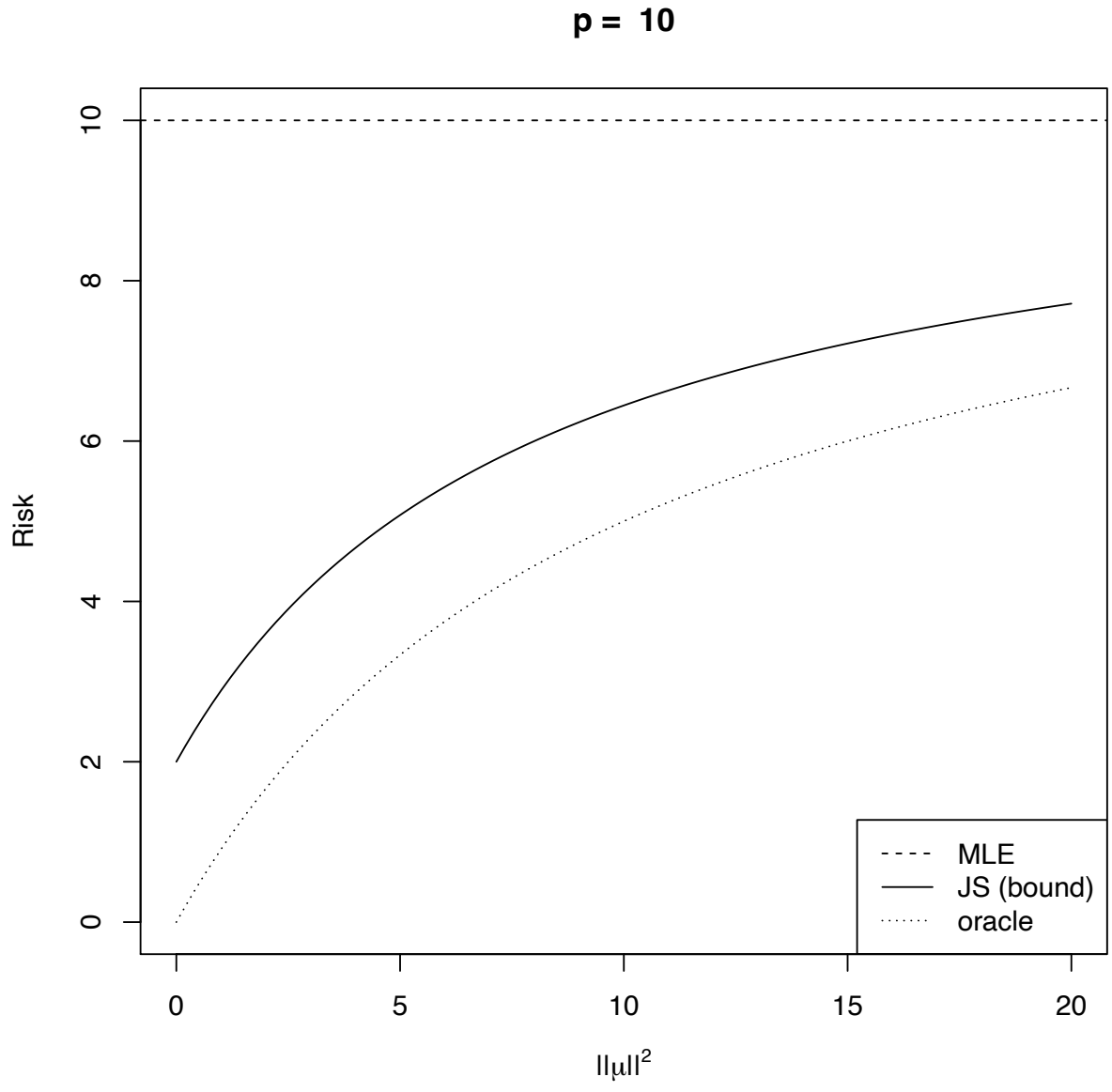
$$R(\tilde{\mu}, \mu) = (1-b)^2 \|\mu\|^2 + b^2 p$$

  minimized by

$$b^* = \frac{\|\mu\|^2}{p + \|\mu\|^2}$$

- The risk of the oracle linear estimator $\tilde{\mu}^* = b^* X$ is

$$R(\tilde{\mu}^*, \mu) = p\|\mu\|^2 / (p + \|\mu\|^2)$$


p = 10
Risk
0
2
4
6
8
10
0
5
10
15
20
||μ||²
MLE
JS (bound)
oracle

- Geometrically, the James-Stein estimator shrinks each component of X towards the origin, and the biggest improvement comes when μ is close to zero
- For $\mu = 0$ we have $R(\hat{\mu}_{JS}, 0) = 2$ for all $p \geq 2$
- As $\|\mu\|^2 \to \infty$, $R(\hat{\mu}_{JS}, \mu) \to p$

Stein's heuristic argument (1956)

- Stein argued that a good estimate should obey $\hat{\mu}_i \approx \mu_i$ for every i. Thus we should also have $\hat{\mu}_i^2 \approx \mu_i^2$, which further implies $\sum_i \hat{\mu}_i^2 \approx \sum_i \mu_i^2$
- Consider the estimator $\hat{\mu} = X$. For this estimator we have

$$\mathbb{E}\|X\|^2 = \mathbb{E}\sum_i X_i^2 = \mathbb{E}\sum_i (\mu_i + Z_i)^2 = \|\mu\|^2 + p$$

  where $Z_i \sim N(0, 1)$
- This suggests that for large p, $\|X\|^2$ is likely to be considerably larger than $\|\mu\|^2$, and hence we may be able to obtain a better estimator by shrinking the estimator $\hat{\mu} = X$ towards o.

Positive James-Stein estimator

- If the shrinkage in $\hat{\mu}_{JS}$ is too large, it is possible that the estimator switches to the other sign when $\|X\|^2 < p - 2$
- By precluding the possibility of a sign reversal, the positive JS estimator

$$\hat{\mu}_{JS}^{+} = \left(1 - \frac{p-2}{\|X\|^2}\right)_{+} X$$

  where $(a)_+ = \max(a, 0)$ denotes the positive part
- $\hat{\mu}_{JS}^{+}$ further improves upon the $\hat{\mu}_{JS}$ estimate, i.e., $R(\hat{\mu}_{JS}^{+}, \mu) < R(\hat{\mu}_{JS}, \mu)$ for all μ
- However, this estimator is not admissible either.

Shrinking toward an arbitrary point

- In terms of choosing a point to shrink towards, though, there is nothing special about the origin, and we could equally well shrink towards any pre-chosen $m \in \mathbb{R}^p$ using the estimator

$$\hat{\mu}^m_{JS} = m + \left(1 - \frac{p-2}{\|X-m\|^2}\right)(X-m)$$

- In this case, we have $R(\hat{\mu}^m_{JS}, \mu - m) = R(\hat{\mu}_{JS}, \mu)$, so $\hat{\mu}^m_{JS}$ still strictly dominates $\hat{\mu} = X$

Correlated data

- Assume that $X \sim N_p(\mu, \Sigma)$ where Σ is a known covariance matrix
- A a generalization of James-Stein estimator

$$\hat{\mu}_{JS}^{\Sigma} = \left(1 - \frac{c(\tilde{p} - 2)}{X^t \Sigma^{-1} X}\right) X$$

  with $0 < c < 2$ and $\tilde{p} = \text{tr}(\Sigma)/\lambda_{\text{max}}(\Sigma)$ is the effective dimension of the problem, where $\lambda_{\text{max}}(\Sigma)$ is the maximum eigenvalue of Σ
- If $\tilde{p} > 2$, then the generalization of the JS estimator $\hat{\mu}_{JS}^{\Sigma}$ dominates the MLE $\hat{\mu} = X$

Linear model

- We can apply the previous result to the case of linear regression $y \sim N_n(X\beta, \sigma^2 I_n)$, where the MLE is the OLS estimator $\hat{\beta} = (X^t X)^{-1} X^t y \sim N_p(\beta, \sigma^2 (X^t X)^{-1})$, so with $\mu = X\beta$ and $\hat{\mu} = X\hat{\beta}$ we have $R(\hat{\mu}, \mu) = \sigma^2 p$
- James-Stein estimator becomes

$$\hat{\beta}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\hat{\beta}^t X^t X \hat{\beta}}\right) \hat{\beta}$$

- Letting $\hat{\mu}_{JS} = X\hat{\beta}_{JS}$ and $\mu = X\beta$, the James–Stein Theorem guarantees that

$$R(\hat{\mu}_{JS}, \mu) \leq \sigma^2 p$$

no matter what β is, as long as $p \geq 3$

- It is natural to ask how crucial the normality and squared error loss assumptions are to the Stein phenomenon
- The normality assumption is not critical at all;
- The original result can also be generalised to different loss functions, but there is an important caveat here: the Stein phenomenon only holds when we are interested in simultaneous estimation of all components of μ. If our loss function were $L(\hat{\mu}, \mu) = (\hat{\mu}_1 - \mu_1)^2$ then we could not improve on $\hat{\mu} = X$

$p = 5$, $\mu = (\sqrt{p/2}, \sqrt{p/2}, 0, 0, 0)^t$, $\|\mu\|^2 = p$

10^4 repetitions

	Risk	Risk1	Risk2	Risk3	Risk4	Risk5
MLE	5.00	1.01	1.01	1.00	0.98	0.99
JS	3.65	1.08	1.07	0.50	0.49	0.50

$R(\hat{\mu}_{JS}, \mu) \leq p - (p-2)/(1 + p/(p-2)) = 3.875$

An Empirical Bayes interpretation

Bayesian setup

- Consider the Bayesian setup

$$\mu_i \sim N(0, \tau^2) \quad X|\mu \sim N(\mu, I_p) \tag{1}$$

- Given the data X, the posterior of μ is

$$\mu|X \sim N(\lambda X, \lambda I_p)$$

  where $\lambda = \tau^2/(1+\tau^2)$

- The Bayes estimator is simply the mean of the posterior

$$\hat{\mu}_B = \lambda X = \left(1 - \frac{1}{1+\tau^2}\right) X$$

- Assuming (1), the Bayes risk is $R(\hat{\mu}_B, \mu) = \lambda p$

Connection to James-Stein

- We cannot directly compute the shrinkage factor $\lambda = \tau^2/(1+\tau^2)$, but perhaps we can estimate it using the data
- Since $X_i = \mu_i + Z_i \sim N(0, 1+\tau^2)$, where $Z_i \sim N(0,1)$. This implies $\|X\|^2 \sim (1+\tau^2)\chi^2_p$
- Combining this result with $\mathbb{E}[(p-2)/\chi^2_p] = 1$, we arrive at an unbiased estimate for λ

$$\hat{\lambda} = \left(1 - \frac{(p-2)}{\|X\|^2}\right)$$

- Assuming (1), the Bayes risk is $R(\hat{\mu}_{JS}, \mu) = \left(1 + \frac{2}{p\tau^2}\right) R(\hat{\mu}_B, \mu)$

$p = 5, \tau^2 = 2, \mu_i \sim N(0, \tau^2)$

10^4 repetitions

	Bayes Risk	B.Risk1	B.Risk2	B.Risk3	B.Risk4	B.Risk5
MLE	5.01	1.01	1.01	1.00	0.99	1.00
BAYES	3.34	0.67	0.68	0.67	0.67	0.66
JS	4.02	0.81	0.82	0.80	0.80	0.79

$R(\hat{\mu}, \mu) = 5, R(\hat{\mu}_B, \mu) = 3.33, R(\hat{\mu}_{JS}, \mu) = 4,$

Shrinking Toward the Group Mean

- In practice, instead of arbitrarily picking some point, it might instead make sense to chose $m = \bar{X}$ as so as to adapt to the true center of μ_i
- Consider the Bayesian setup

$$\mu_i \sim N(m, \tau^2) \quad X|\mu \sim N(\mu, I_p) \tag{2}$$

  with m and τ^2 unknown
- The marginal distribution of our data is

$$X_i \overset{i.i.d.}{\sim} N(m, 1 + \tau^2)$$

  and the posterior mean is

$$\mu|X \sim N(m + \lambda(X - m), \lambda I_p)$$

- $\hat{\mu}_B = m + \lambda(X - m)$ but m is unknown. Taking the empirical Bayes approach, we can use the unbiased estimator $\bar{X}$ in its place
- Similarly, we can use the sample variance $S = \sum_i (X_i - \bar{X})^2 \sim (1 + \tau^2)\chi^2_{p-1}$ to estimate λ. Now we have $\mathbb{E}[(p-3)/\chi^2_{p-1}] = 1$
- This gives us the estimator

$$\hat{\mu}^{\bar{X}}_{JS} = \bar{X} + \left(1 - \frac{p-3}{S}\right)(X - \bar{X})$$

If $p > 3$, this estimator dominates the MLE everywhere

A baseball data example

Player	MLE	TRUTH
1	0.34	0.30
2	0.33	0.35
3	0.32	0.22
4	0.31	0.28
5	0.29	0.26
6	0.29	0.27
7	0.28	0.30
8	0.26	0.27
9	0.24	0.23
10	0.23	0.26
11	0.23	0.26
12	0.22	0.21
13	0.22	0.26
14	0.22	0.27
15	0.21	0.32
16	0.21	0.23
17	0.20	0.28
18	0.14	0.20

The column labelled MLE is the batting average for 18 players in the 1970 season, using the first 90 at bats.

The column labelled TRUTH is the batting average for the remainder of the 1970 season.

- Each player Batting average =(# hits / # at bats) value is a binomial proportion

$$Y_i \sim \text{Binomial}(n, \pi_i)/n$$

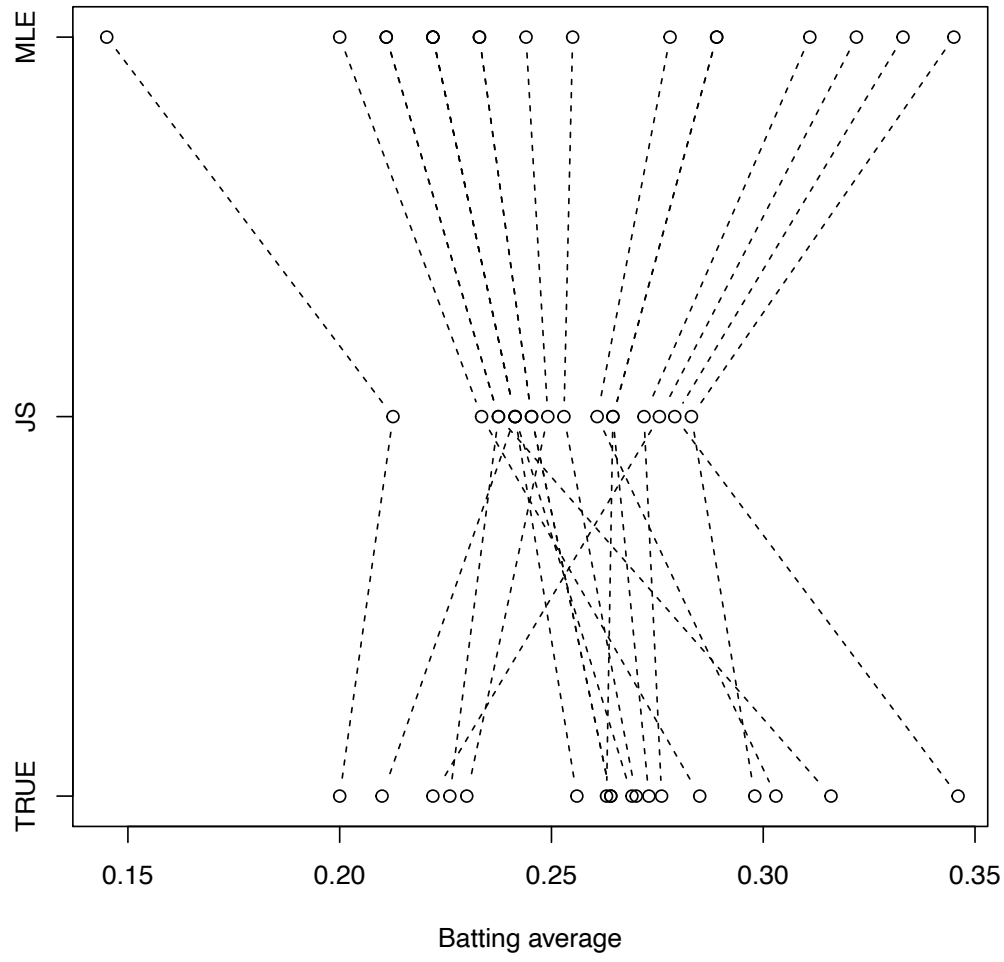
  where π_i is the true average and $n = 90$
- Since batting averages are binomial, we can use the normal approximation

$$Y_i \approx N\left(\pi_i, \frac{\pi_i(1-\pi_i)}{n}\right)$$

  but the variance depends on the mean
- One solution is to make a variance stabilizing transformation

$$X_i = 2\sqrt{n+0.5}\,\arcsin\left(\sqrt{\frac{nY_i + 3/8}{n+3/4}}\right) \approx N(\mu_i, 1)$$

  where $\mu_i = 2\sqrt{n+0.5}\,\arcsin(\sqrt{\frac{n\pi_i+3/8}{90+3/4}})$
- Inverted back $y_i^{JS} = \frac{1}{n}\left[(n+0.75)(\sin(\frac{\hat{\mu}_i^{JS}}{2\sqrt{n+0.5}}))^2 - 0.375\right]$



$$\sum_i (y_i - y_i^{\text{TRUE}})^2 = 0.0425 \quad \sum_i (y_i^{JS} - y_i^{\text{TRUE}})^2 = 0.0205$$

Ridge regression

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Hastie, T. (2020). Ridge regularization: an essential concept in data science. Technometrics, 62(4), 426-433.
- van Wieringen (2015). Lecture notes on ridge regression. arXiv preprint arXiv:1509.09169.

Condition number

- In the linear model, the estimate of β is obtained by solving the normal equations
$$X^\mathsf{T} X\beta = X^\mathsf{T} y$$
- The difficulty of solving this system of linear equations can be described by the *condition number*
$$\kappa(X^\mathsf{T} X) = \frac{d_{\max}}{d_{\min}}$$
the ratio between the largest and smallest singular values of $X^\mathsf{T} X$
- If the condition number is very large, then the matrix is said to be *ill-conditioned* (see Section 2.6 of CASL)

Toy linear model with $n = p = 2$. We set X and β as

$$X = \begin{bmatrix} 10^9 & -1 \\ -1 & 10^{-5} \end{bmatrix} \quad \beta = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

And if we define $y = X\beta$, this gives

$$y = \begin{bmatrix} 10^9 & -1 \\ -1 & 10^{-5} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 10^9 - 1 \\ -0.99999 \end{bmatrix}$$

The reciprocal of condition number, i.e. $1/\kappa(X^\mathsf{T} X) = 9.998e - 29$, is smaller than (my) machine precision, i.e. $2.220446e - 16$

```
X <- matrix(c(10^9, -1, -1, 10^(-5)), 2, 2)
beta <- c(1,1)
y <- X %*% beta

solve( crossprod(X), crossprod(X, y) )

Error in solve.default(crossprod(X)) :
system is computationally singular:
reciprocal condition number = 9.998e-29

.Machine$double.eps
2.220446e-16
```

Ridge regression solution

- Ridge provides a remedy for an *ill-conditioned* X^tX matrix
- If our $n \times p$ design matrix X has column rank less than p (or nearly so in terms of its condition number), then the usual least-squares regression equation is in trouble:

$$\hat{\beta} = (X^tX)^{-1}X^ty$$

- What we do is add a *ridge* on the diagonal - $X^tX + \lambda I_p$ with $\lambda > 0$ - which takes the problem away:

$$\hat{\beta}_\lambda = (X^tX + \lambda I_p)^{-1}X^ty$$

- This is the ridge regression solution proposed by Hoerl and Kennard (1970)

- Ridge regression modifies the normal equations to

$$(X^\mathsf{T} X + \lambda I_p)\beta = X^\mathsf{T} y$$

  and the condition number of $(X^\mathsf{T} X + \lambda I_p)$ is

$$\kappa(X^\mathsf{T} X + \lambda I_p) = \frac{d_{\max} + \lambda}{d_{\min} + \lambda}$$

- Notice that even if $d_{\min} = 0$, the condition number will be finite if $\lambda > 0$
- This technique is known as Tikhonov regularization, after the Russian mathematician Andrey Tikhonov

Penalized (Lagrange) form

- The optimization problem that ridge is solving

$$\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|^2 \tag{1}$$

  where $\|\cdot\|$ is the ℓ_2 Euclidean norm
- The ridge remedy comes with consequences. The ridge estimate is biased toward zero. It also has smaller variance than the OLS estimate.
- Selecting λ amounts to a bias-variance trade-off

Cement data

$n = 13, p = 4$

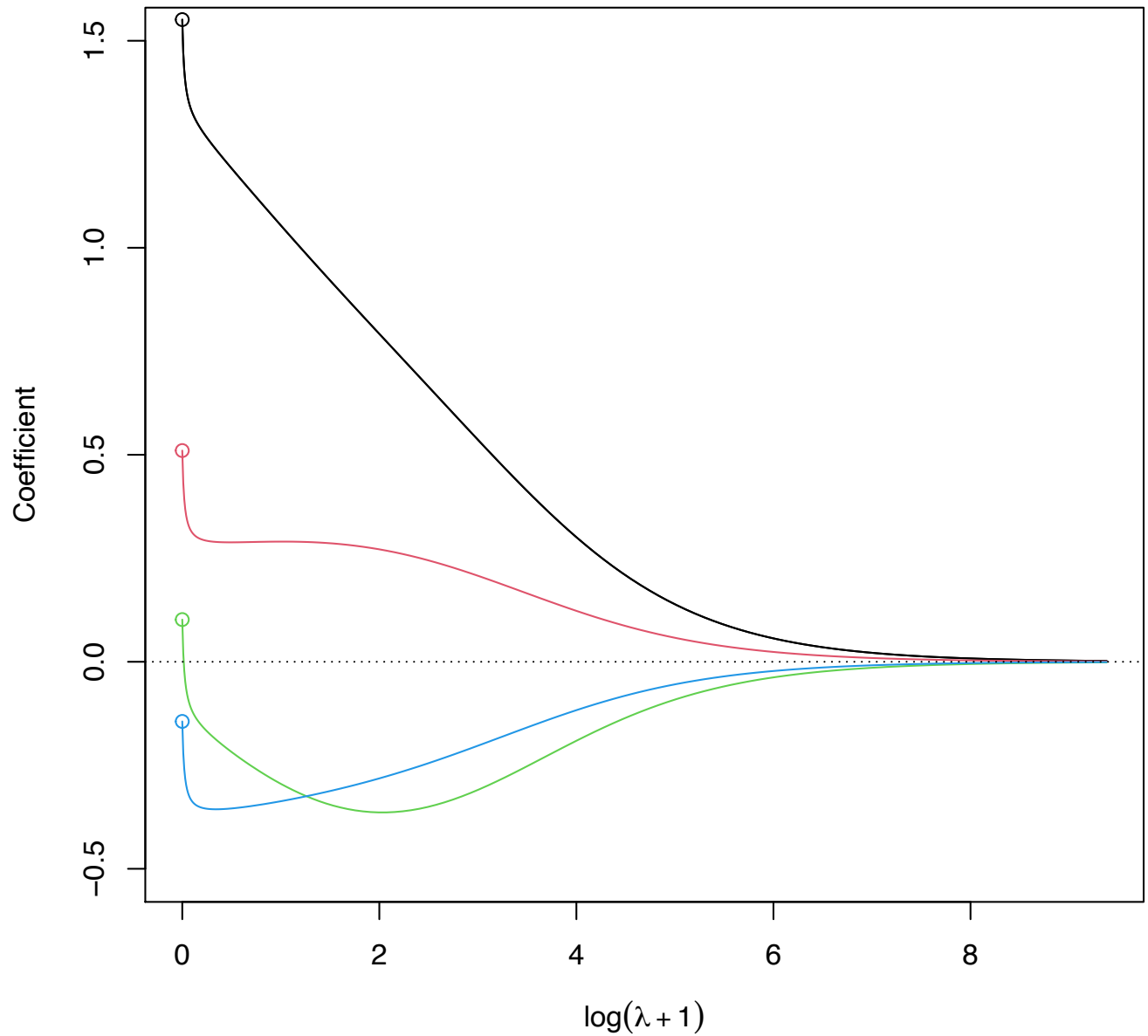
$$R = \begin{bmatrix} 1 & 0.23 & -0.82 & -0.25 \\ 0.23 & 1 & -0.14 & -0.97 \\ -0.82 & -0.14 & 1 & 0.03 \\ -0.25 & -0.97 & 0.03 & 1 \end{bmatrix}$$

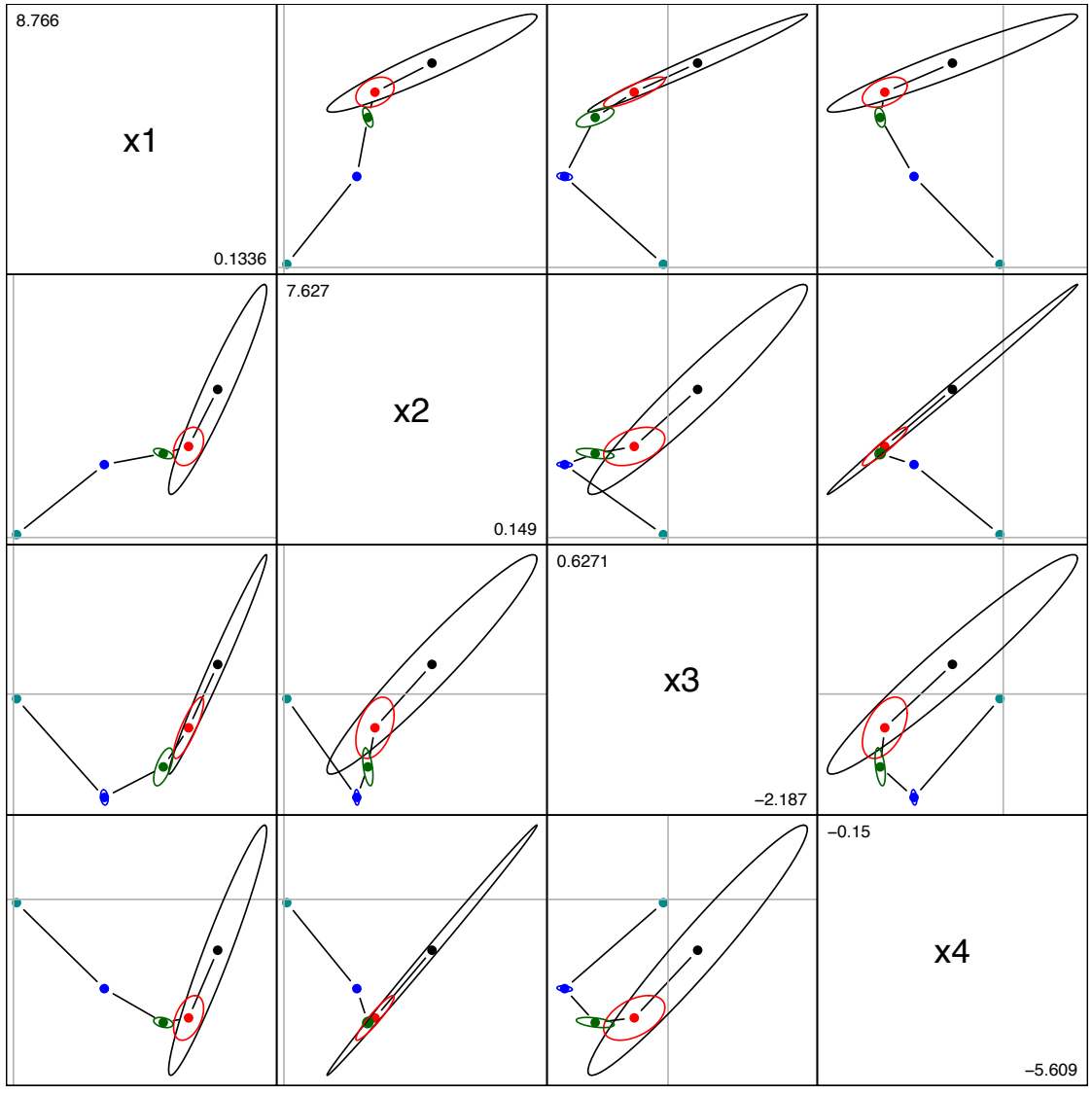
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62.41	70.07	0.89	0.40
x1	1.55	0.74	2.08	0.07
x2	0.51	0.72	0.70	0.50
x3	0.10	0.75	0.14	0.90
x4	-0.14	0.71	-0.20	0.84

R-squared: 0.9824

	x1	x2	x3	x4
VIF	38.50	254.42	46.87	282.51

Piepel, Redgate (1998) A Mixture Experiment Analysis of the Hald Cement Data, The American Statistician, 52:23–30


Coefficient
1.5
1.0
0.5
0.0
−0.5
0
2
4
6
8
log(λ + 1)



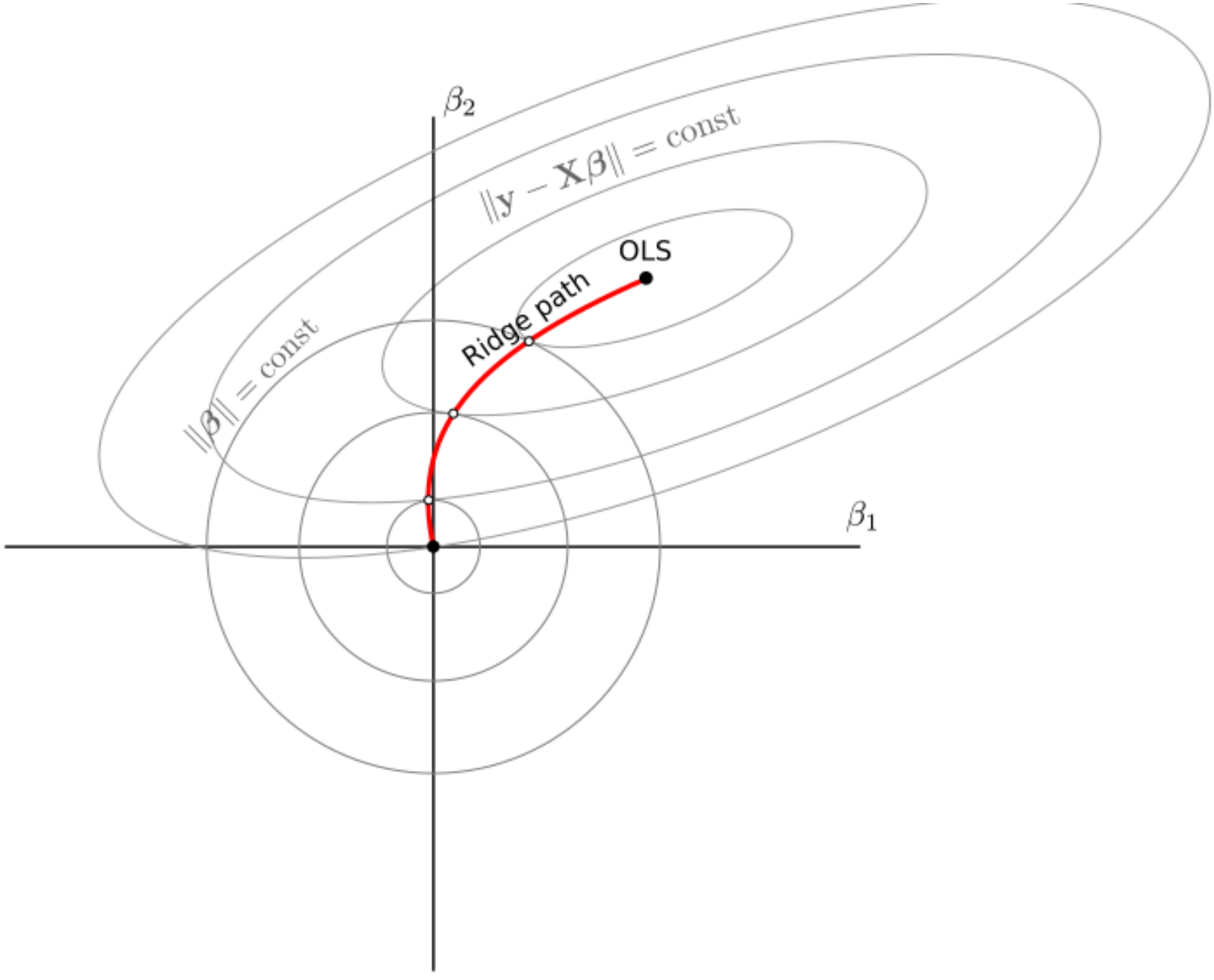
$$\lambda = 0, 0.1, 1, 10, 1000$$

Constrained form

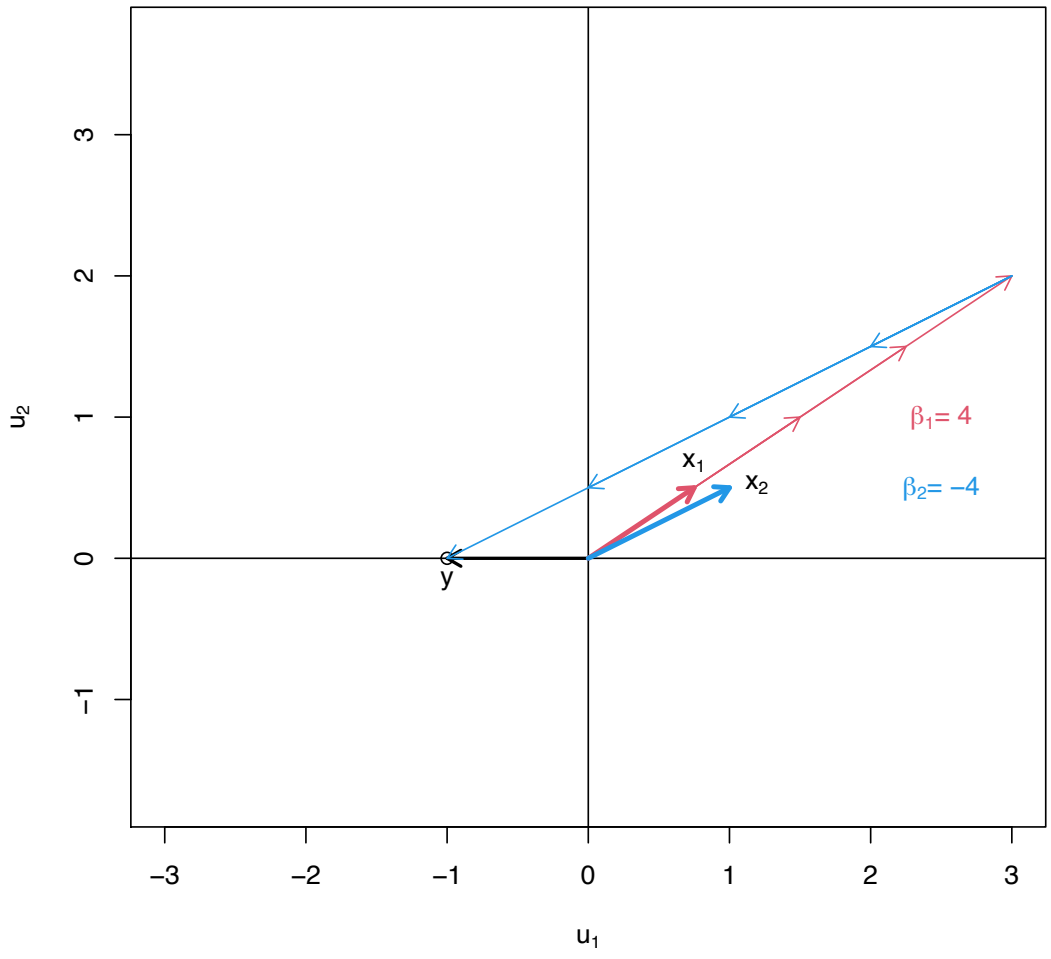
- We can also express the ridge problem as

$$\min_{\beta} \|y - X\beta\|^2 \quad \text{subject to } \|\beta\| \leq c \tag{2}$$

- The two problems are of course equivalent: every solution $\hat{\beta}_\lambda$ in (1) is a solution to (2) with $c = \|\hat{\beta}_\lambda\|$


β_2
$\|\mathbf{y} - \mathbf{X}\beta\| = \text{const}$
OLS
Ridge path
$\|\beta\| = \text{const}$
β_1

Overfitting



Large estimates of β are often an indication of overfitting

Bayesian view

- Assume

$$y_i|\beta, X = x_i \sim x_i^t\beta + \epsilon_i$$

  with ϵ_i i.i.d. $N(0, \sigma_\epsilon^2)$. Here we think of β as random as well, and having a prior distribution

$$\beta \sim N(0, \sigma_\beta^2 I_p)$$

- Then the negative log posterior distribution is proportional to (1), with

$$\lambda = \frac{\sigma_\epsilon^2}{\sigma_\beta^2}$$

  and the posterior mean is the ridge estimator
- The smaller the prior variance parameter σ_β^2, the more the posterior mean is shrunk toward zero, the prior mean for β

Important details

- When including an intercept term, we usually leave this coefficient unpenalized, solving

$$\min_{\alpha,\beta} \|y - 1\alpha - X\beta\|^2 + \lambda\|\beta\|^2$$

- Ridge regression is not invariant under scale transformations of the variables, so it is standard practice to centre each column of X (hence making them orthogonal to the intercept term) and then scale them to have Euclidean norm $\sqrt{n}$
- It is straightforward to show that after this standardisation of X, $\hat{\alpha} = \bar{y}$, so we can also centre y and then remove α from our objective function
- Different `R` packages have different defaults, e.g. `glmnet` also standardizes y

- Let $\tilde{y} = (y - 1\bar{y})$ and $\tilde{X} = (X - 1\bar{x}^t)\text{diag}(1/s)$ be the centered y and standardized X, respectively, with
  - $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$,
  - $\bar{x} = (1/n)X'1$,
  - $s = (s_1, \ldots, s_p)^t$ and $s_j^2 = (1/n)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$
- Compute the scaled coefficients

$$\tilde{\beta}_\lambda = (\tilde{X}^t\tilde{X} + \lambda I_p)\tilde{X}^t\tilde{y}$$

- Transform back to unscaled coefficients

$$\hat{\beta}_\lambda = \text{diag}(1/s)\tilde{\beta}_\lambda \quad \hat{\alpha} = \bar{y} - \bar{x}^t\hat{\beta}_\lambda$$

Ridge computations and the SVD

Tuning parameter

- In many wide-data and other ridge applications, we need to treat λ as a tuning parameter, and select a good value for the problem at hand.
- For this task we have a number of approaches available for selecting λ from a series of candidate values:
  - With a validation dataset separate from the training data, we can evaluate the prediction performance at each value of λ
  - Cross-validation does this effciently using just the training data, and leave-one-out (LOO) CV is especially efficient

SVD

- Whatever the approach, they all require computing a number of solutions $\hat{\beta}_\lambda$ at different values of λ: the *ridge regularization path*
- We can achieve great efficiency via the (full form) Singular Value Decomposition (SVD)

$$X = UDV^t$$

where U $n \times n$ orthogonal, V $p \times p$ orthogonal and D $n \times p$ diagonal, with diagonal entries $d_1 \geq \ldots \geq d_m \geq 0$, where $m = \min(n, p)$

- From the SVD we get

$$\begin{aligned}\hat{\beta}_\lambda &= (VD^tU^tUDV^t + \lambda VV^t)^{-1}VD^tU^ty \\ &= V(D^tD + \lambda I_p)^{-1}D^tU^ty \\ &= \sum_{d_j>0} v_j \frac{d_j}{d_j^2+\lambda}\langle u_j, y\rangle\end{aligned} \tag{3}$$

  where v_j (u_j) is the jth column of V (U), and $\langle a, b\rangle = a^tb$
- Once we have the SVD of X, we have the ridge solution for all values of λ
- When $n > p$ the ridge solution with $\lambda = 0$ is simply the OLS solution for β
- When $p > n$, there are infinitely many least squares solutions for β, all leading to a zero-residual solution. From (3) with $\lambda = 0$ we get a unique solution, the one with minimum Euclidean norm

– Fitted values

$$\begin{aligned}\hat{y}_\lambda &= U\mathrm{diag}\Big(\frac{d_1^2}{d_1^2+\lambda},\ldots,\frac{d_p^2}{d_p^2+\lambda}\Big)U^t y \\ &= \sum_{d_j>0} u_j \frac{d_j^2}{d_j^2+\lambda}\langle u_j, y\rangle\end{aligned}$$

Principal components regression

- Ridge

$$\hat{\beta}_\lambda = V\text{diag}\Big(\frac{d_1}{d_1^2+\lambda}, \ldots, \frac{d_p}{d_p^2+\lambda}\Big) U^t y$$

- Principal components regression with q components

$$\hat{\beta}_q = V\text{diag}\Big(\frac{1}{d_1}, \ldots, \frac{1}{d_q}, 0, \ldots, 0\Big) U^t y$$

- Both operate on the singular values, but where principal component regression thresholds the singular values, ridge regression shrinks them

Ridge and the bias-variance trade-off

Bias

- Assume that the data arise from a linear model $y \sim N(X\beta, \sigma^2 I_n)$, then $\hat{\beta}_\lambda$ will be a biased estimate of β. Throughout this section X is assumed fixed, $n > p$ and X has full column rank
- The ridge estimator can be expressed as

$$\hat{\beta}_\lambda = (X^t X + \lambda I_p)^{-1} X^t X \hat{\beta}$$

- We can get an explicit expression for the bias

$$\begin{aligned} \text{Bias}(\hat{\beta}_\lambda) &= \mathbb{E}(\hat{\beta}_\lambda) - \beta \\ &= V \text{diag}\left(\frac{\lambda}{d_1^2 + \lambda}, \ldots, \frac{\lambda}{d_p^2 + \lambda} \right) V^t \beta \\ &= \sum_{j=1}^{p} v_j \frac{\lambda}{d_j^2 + \lambda} \langle v_j, \beta \rangle \end{aligned}$$

Variance

- Similarly there is a nice expression for the covariance matrix

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_\lambda) &= \sigma^2 V \mathrm{diag}\left(\frac{d_1^2}{(d_1^2+\lambda)^2}, \ldots, \frac{d_p^2}{(d_p^2+\lambda)^2}\right) V^t \\ &= \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^2} v_j v_j^t \end{aligned}$$

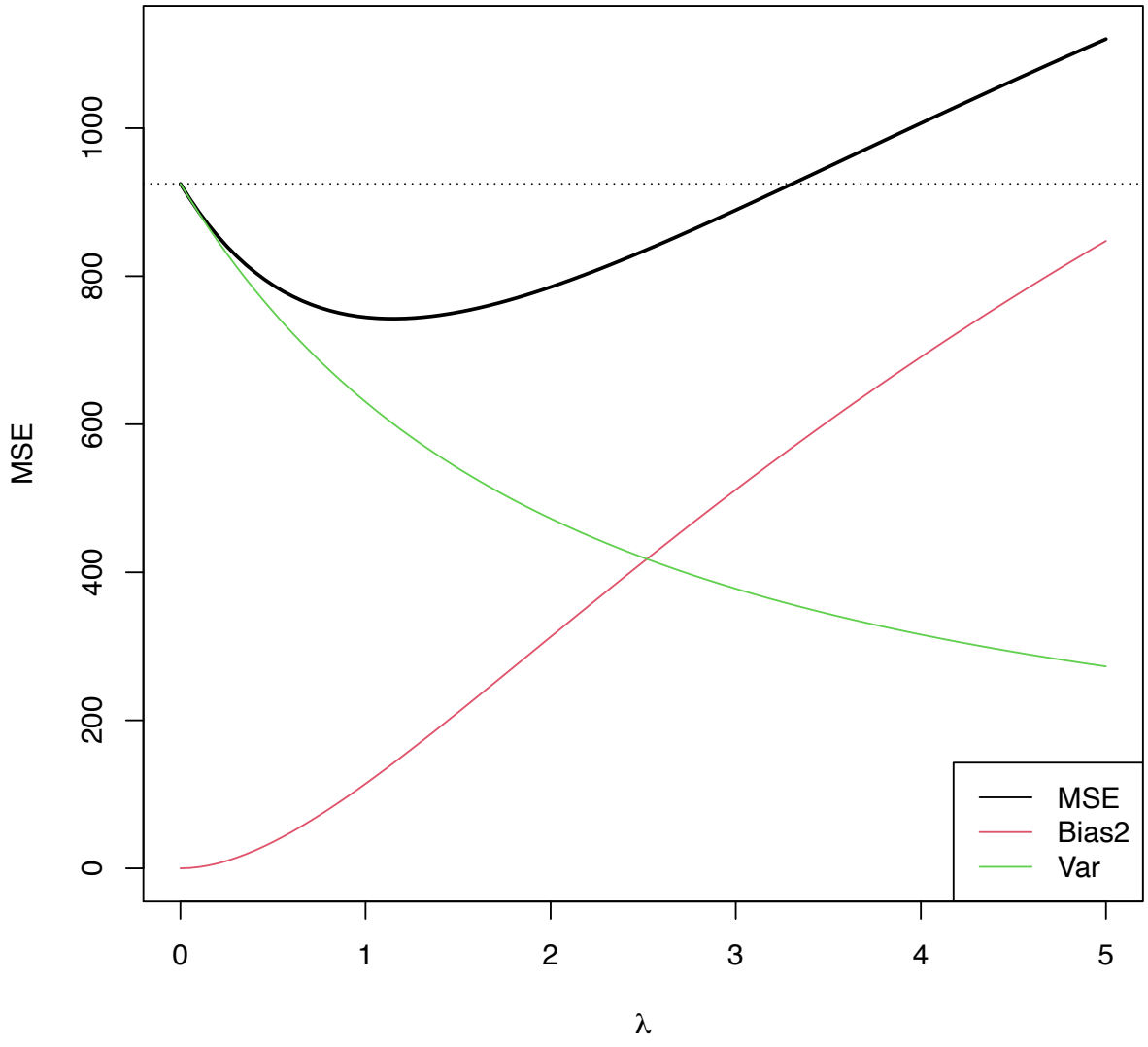
- With $\lambda = 0$, this is $\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^t X)^{-1} \succeq \mathrm{Var}(\hat{\beta}_\lambda)$ for $\lambda > 0$

Mean Squared Error

- MSE of the ridge regression estimator

$$\begin{aligned} \text{MSE}(\hat{\beta}_\lambda) &= \mathbb{E}[(\hat{\beta}_\lambda - \beta)^t(\hat{\beta}_\lambda - \beta)] \\ &= \text{tr}[\text{Var}(\hat{\beta}_\lambda)] + \text{Bias}(\hat{\beta}_\lambda)^t\text{Bias}(\hat{\beta}_\lambda) \end{aligned}$$

- *Theorem (Theobald, 1974)*
  There exists $\lambda > 0$ such that $\text{MSE}(\hat{\beta}_\lambda) < \text{MSE}(\hat{\beta})$.


MSE
0
200
400
600
800
1000
0
1
2
3
4
5
λ
MSE
Bias2
Var

Expected prediction error

- When we make predictions $\hat{y}_i = x_i^t \hat{\beta}_\lambda$ at x_i

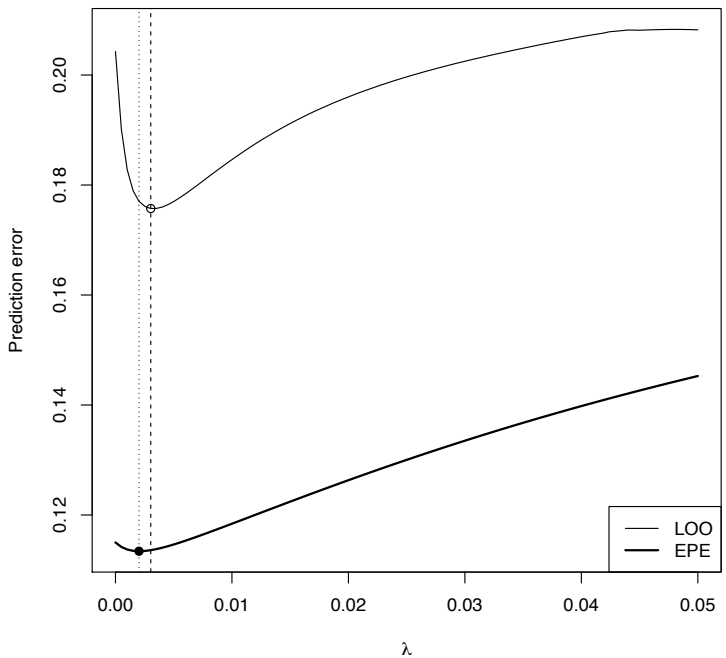
$$\begin{aligned} \text{MSE}(\hat{y}_i) &= \mathbb{E}[(x_i^t \hat{\beta}_\lambda - x_i^t \beta)^2] \\ &= x_i^t \text{Var}(\hat{\beta}_\lambda) x_i + [x_i^t \text{Bias}(\hat{\beta}_\lambda)]^2 \end{aligned}$$

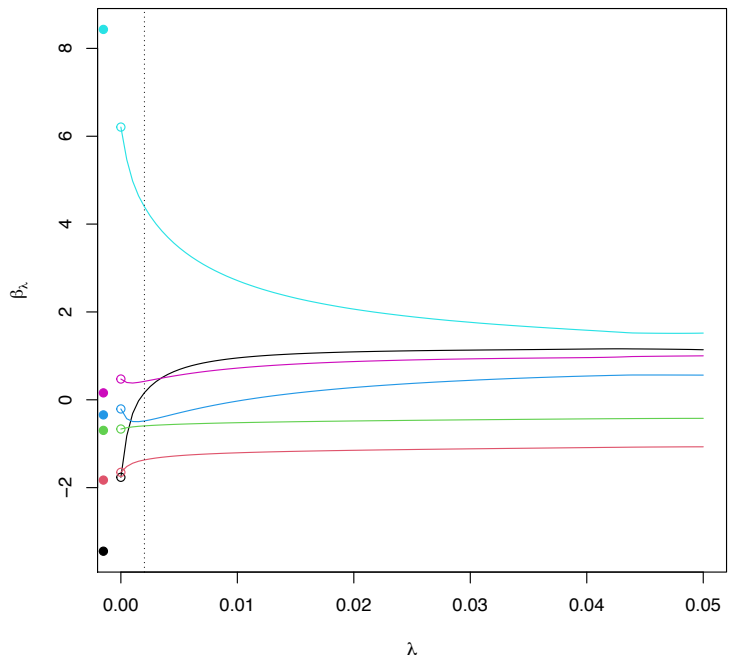
- Expected prediction error

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i^{\text{new}})^2\Big] = \frac{1}{n}\sum_{i=1}^{n}\text{MSE}(\hat{y}_i) + \sigma^2$$

Longley data

$n = 16, p = 6$





Orthonormal design matrix

- Consider an orthonormal design matrix X, i.e. $X^tX = I_p = (X^tX)^{-1}$, e.g.

$$X = \frac{1}{2}\begin{bmatrix} -1 & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix}$$

- $\hat{\beta}_\lambda = \frac{1}{(1+\lambda)}\hat{\beta}$
- $\text{Var}(\hat{\beta}_\lambda) = \frac{\sigma^2}{(1+\lambda)^2} I_p$
- $\text{MSE}(\hat{\beta}_\lambda) = \frac{p\sigma^2}{(1+\lambda)^2} + \frac{\lambda^2 \|\beta\|^2}{(1+\lambda)^2}$ with minimum at $\lambda = \frac{p\sigma^2}{\|\beta\|^2}$

Ridge and leave-one-out cross validation

LOO

- For n-fold (LOO) CV, we have another beautiful result for ridge and other linear operators

$$\text{LOO}_\lambda = \frac{1}{n}\sum_{i=1}^n (y_i - x_i^t \hat{\beta}_\lambda^{(-i)})^2 = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - x_i^t \hat{\beta}_\lambda}{1 - R_{ii}^\lambda}\right)^2$$

  where $\hat{\beta}_\lambda^{(-i)}$ is the ridge estimate computed using the $(n-1)$ observations with the pair (x_i, y_i) and

$$R^\lambda = X(X^tX + \lambda I)^{-1}X^t$$

- The equation says we can compute all the LOO residuals for ridge from the original residuals, each scaled up by $1/(q - R_{ii}^\lambda)$
- We can obtain R^λ efficiently for all λ via

$$R^\lambda = U\text{diag}\left(\frac{d_1^2}{d_1^2 + \lambda}, \ldots, \frac{d_p^2}{d_p^2 + \lambda}\right)U^t$$

- For each pair (x_i, y_i) left out, we solve

$$\min_{\beta} \sum_{l \neq i} (y_l - x_l^t \beta) + \lambda \|\beta\|^2$$

  with solution $\hat{\beta}_\lambda^{(-i)}$.
- Let $y_i^* = x_i^t \hat{\beta}_\lambda^{(-i)}$. If we insert the pair (x_i, y_i^*) back into the size $n-1$ dataset, it will not change the solution
- Back at a full n dataset, and using the linearity of the ridge operator, we have

$$y_i^* = \sum_{l \neq i} R_{il}^\lambda y_l + R_{ii}^\lambda y_i^* = \sum_{l=1}^{n} R_{il}^\lambda y_l - R_{ii}^\lambda y_i + R_{ii}^\lambda y_i^* = \hat{y}_i - R_{ii}^\lambda y_i + R_{ii}^\lambda y_i^*$$

  from which we see that $(y_i - y_i^*) = (y_i - \hat{y}_i)/(1 - R_{ii}^\lambda)$

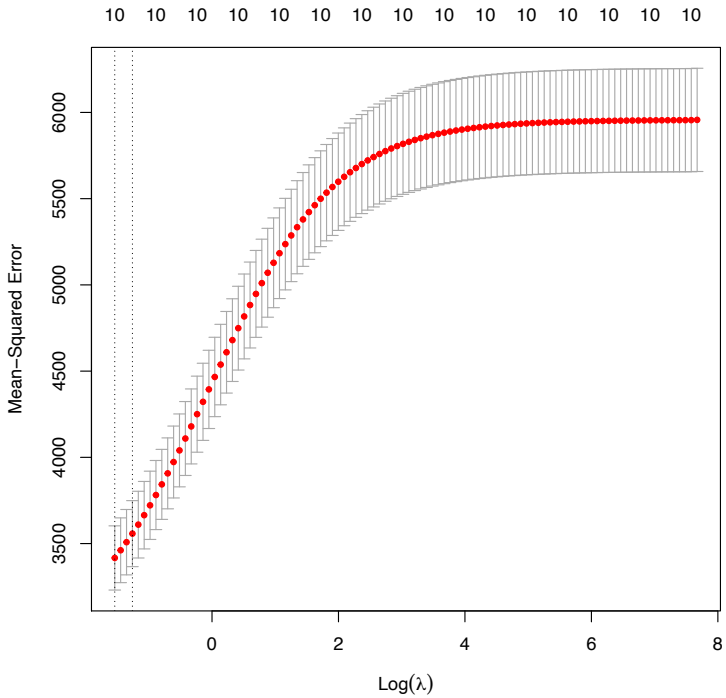
GCV

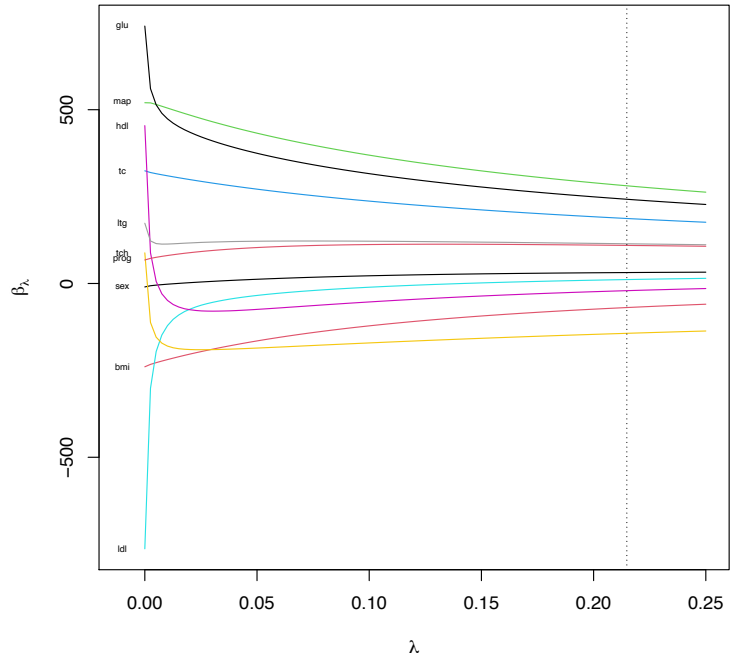
- The identity $\mathrm{tr}(R^\lambda) = \sum_{i=1}^n R_{ii}^\lambda$ suggests $R_{ii}^\lambda \approx \frac{1}{n}\mathrm{tr}(R^\lambda)$
- Generalized cross validation

$$\mathrm{GCV}_\lambda = \frac{1}{n}\sum_{i=1}^n \frac{(y_i - x_i^t\hat{\beta}_\lambda)^2}{(1 - \frac{1}{n}\mathrm{tr}(R^\lambda))^2}$$

Diabetes data

$n = 442, p = 10$





Ridge and the kernel trick

- The fitted values from ridge regression are

$$\hat{y}_\lambda = X(X^tX + \lambda I_p)^{-1}X^ty \tag{4}$$

- An alternative way of writing this is suggested by the following

$$\begin{aligned} X^t(XX^t + \lambda I_n) &= (X^tX + \lambda I_p)X^t \\ (X^tX + \lambda I_p)^{-1}X^t &= X^t(XX^t + \lambda I_n)^{-1} \\ X(X^tX + \lambda I_p)^{-1}X^ty &= XX^t(XX^t + \lambda I_n)^{-1}y \end{aligned}$$

  giving

$$\hat{y}_\lambda = K(K + \lambda I_n)^{-1}y \tag{5}$$

  where $K = XX^t = \{x_i^tx_j\}_{ij}$ is the $n \times n$ *gram* matrix of pairwise inner products, where x_i^t and x_j^t are the ith and jth row of X

- Complexity can be expressed in terms of floating point operations (flops) required to find the solution. (4) requires $O(np^2 + p^3)$ operations, (5) $O(pn^2 + n^3)$ operations

- Suppose we want to add all pairwise interactions

$$
\begin{aligned}
&x_{i1}, x_{i2}, \ldots, x_{ip} \\
&x_{i1}x_{i1}, x_{i1}x_{i2}, \ldots, x_{i1}x_{ip} \\
&\vdots \\
&x_{ip}x_{i1}, x_{ip}x_{i2}, \ldots, x_{ip}x_{ip}
\end{aligned}
$$

  giving $O(p^2)$ columns in the design matrix. Since (5) now requires $O(p^2n^2 + n^3)$ operations, for large p it can be computationally prohibitive
- However, K can be computed directly with

$$K_{ij} = (\frac{1}{2} + x_i^t x_j)^2 - \frac{1}{4} = \sum_k x_{ik}x_{jk} + \sum_{k,l} x_{ik}x_{il}x_{jk}x_{jl}$$

  this amounts to an inner product between vectors of the form

$$(x_{i1}, \ldots, x_{ip}, x_{i1}x_{i1}, \ldots, x_{i1}x_{ip}, x_{i2}x_{i1}, \ldots, x_{i2}x_{ip}, \ldots, x_{ip}x_{ip})$$

  and it requires $O(pn^2)$ operations

Smoothing splines

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

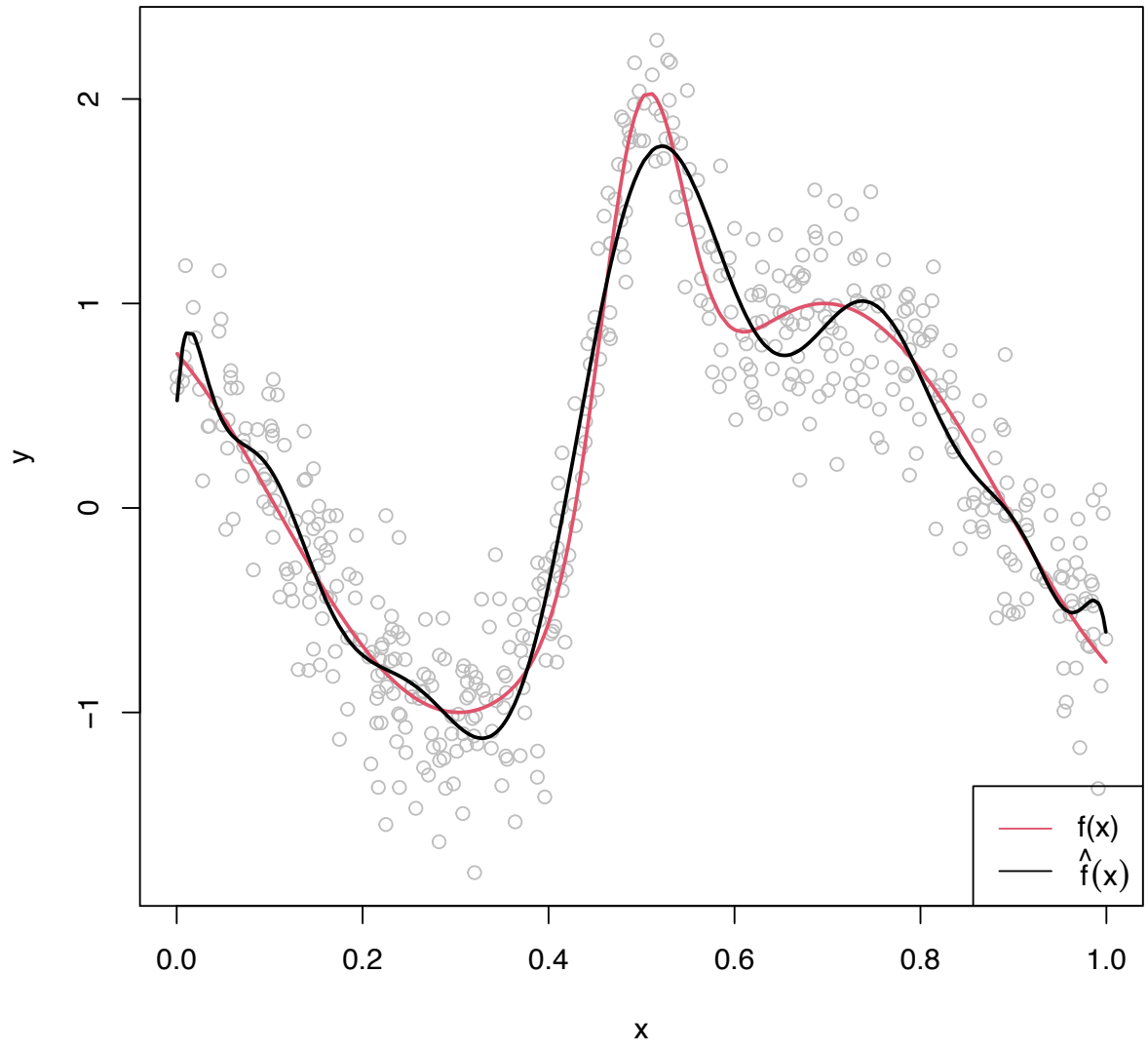
- Bowman, Evers. Lecture Notes on Nonparametric Smoothing. Section 3
- Eilers, Marx (1996). Flexible smoothing with B-splines and penalties. Statistical science, 11(2), 89–121.

Nonlinearity

- In many situations, the relationship between x and y is not linear
- Suppose $y = f(x) + \varepsilon$ with

$$f(x) = \sin(2(4x - 2)) + 2e^{-(16^2)(x-.5)^2}$$

- Even with a polynomial of degree 15, the fit is fairly poor in many areas, and 'wiggles' in some places where there doesn't appear to be a need to



Polynomial regression of degree 15

Piecewise polynomial

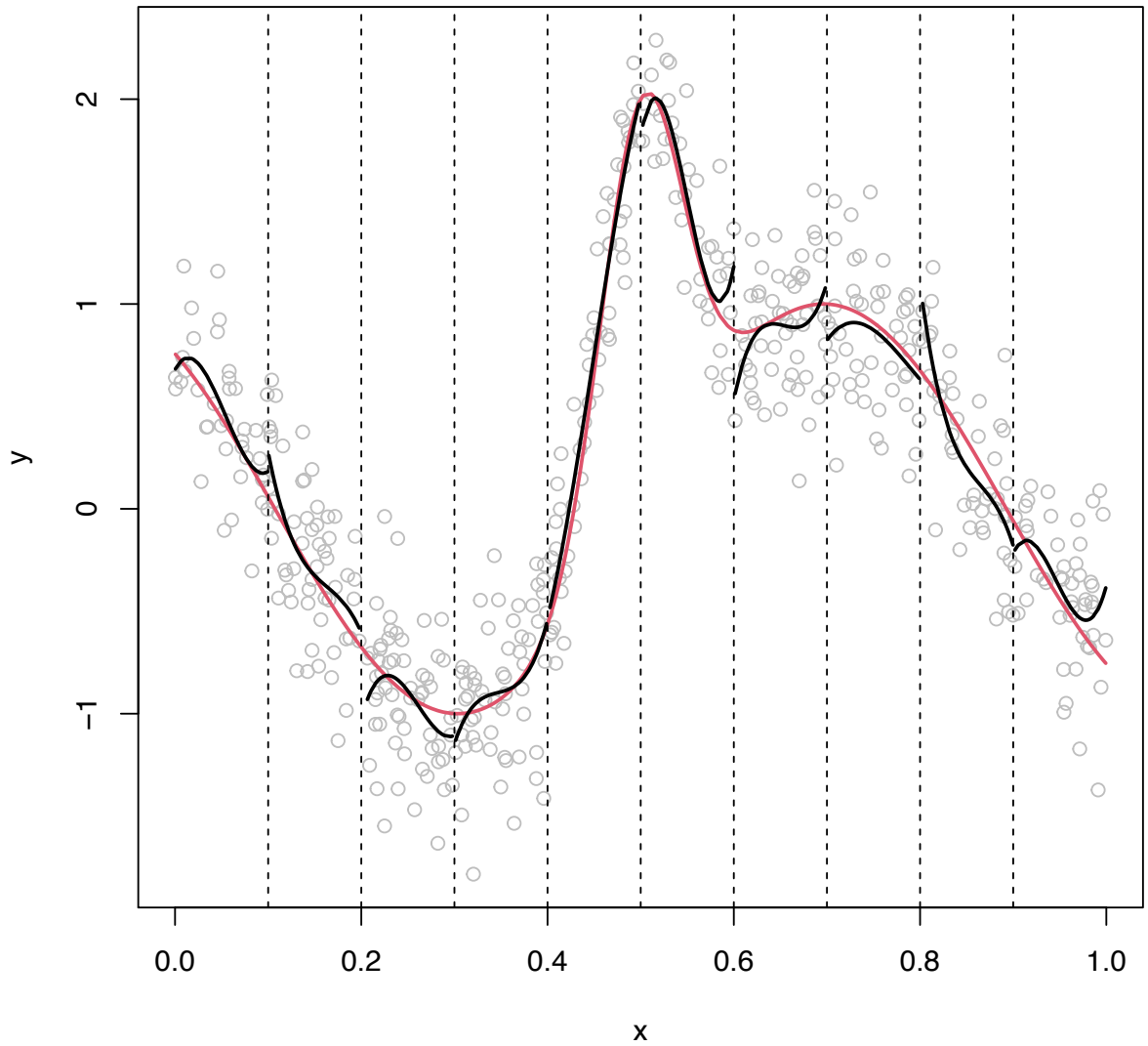
- Divide the data into chunks at various points (*knots*), and fit a polynomial model within that subset of data
- Specify K *internal knots* on the range of x:

$$\min(x) < \xi_1 < \ldots < \xi_K < \max(x)$$

  which define $K + 1$ intervals
- Fit a polynomial model of degree M on each of the $K + 1$ intervals

$$(-\infty, \xi_1], (\xi_1, \xi_2], \ldots, (\xi_{K-1}, \xi_K], (\xi_K, +\infty)$$

- A shortcoming is that at each knot the predicted values will generally not be continuous



Piecewise cubic regression

Basis expansion

$$f(x) = \sum_{j=1}^{p} \beta_j B_j(x)$$

where $B_j(\cdot)$ are known functions called *basis functions*. For example

- 3rd degree polynomial:

$$B_1(x) = 1, B_2(x) = x, B_3(x) = x^2, B_4(x) = x^3$$

- Step function with K knots:

$$B_1(X) = \mathbb{1}\{x < \xi_1\}, B_2(x) = \mathbb{1}\{\xi_1 \leq x < \xi_2\}, \ldots,$$
$$B_K(x) = \mathbb{1}\{\xi_{K-1} \leq x < \xi_K\}, B_{K+1}(x) = \mathbb{1}\{x \geq \xi_K\}$$

Regression splines

Regression splines

A *spline* of degree M with knots $\xi_1, \ldots, \xi_K$

- is a polynomial of degree M on each of the intervals

$$(-\infty, \xi_1], [\xi_1, \xi_2], [\xi_2, \xi_3], \ldots, [\xi_{K-1}, \xi_K], [\xi_K, \infty)$$

- it has continuous derivatives of orders $0, \ldots, M-1$ at each knot

$$f(\xi_k^-) = f(\xi_k^+), \quad \ldots, \quad f^{(M-1)}(\xi_k^-) = f^{(M-1)}(\xi_k^+), \quad k = 1, \ldots, K$$

where ξ_k^+ and ξ_k^- indicate indicate the left and right limits of the function at ξ_k

Truncated power basis

- A spline of degree M with knots $\xi_1, \ldots, \xi_K$ can be defined by the *truncated power basis*

$$\begin{aligned} B_1(x) &= 1 \\ B_{j+1}(x) &= x^j, \quad j = 1, \ldots, M \\ B_{M+k+1}(x) &= (x - \xi_k)_+^M, \quad k = 1, \ldots, K \end{aligned}$$

  where $(\cdot)_+$ defines the positive portion of its argument, i.e.

$$(x - \xi_k)_+^M = \begin{cases} (x - \xi_k)^M & x \geq \xi_k \\ 0 & \text{otherwise} \end{cases}$$

- There are $K + 1$ polynomials of order M and K sets of M constraints; the truncated power basis has $(K + 1)(M + 1) - KM$, or $1 + M + K$ free parameters.

- We compute the solution by explicitly constructing a $n \times (1 + M + K)$ design matrix B with

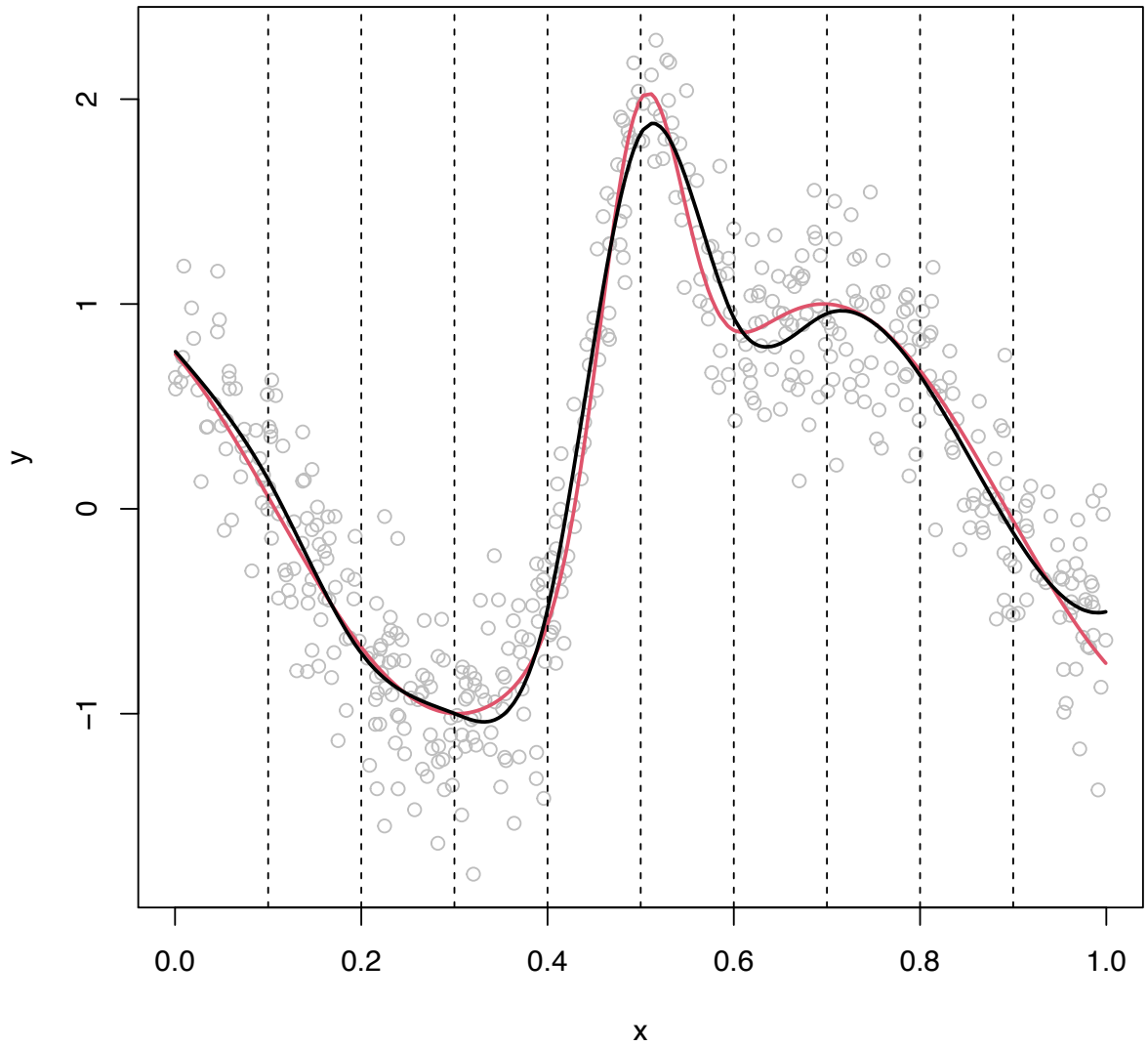
$$B_{i,j} = B_j(x_i) \qquad i = 1, \ldots, n, \quad j = 1, \ldots, 1 + M + K$$

- The regression spline can be written as

$$\hat{f}(x) = \sum_{j=1}^{1+M+K} \hat{\beta}_j B_j(x)$$

where $\hat{\beta}$ is given by

$$\hat{\beta} = (B^t B)^{-1} B^t y$$

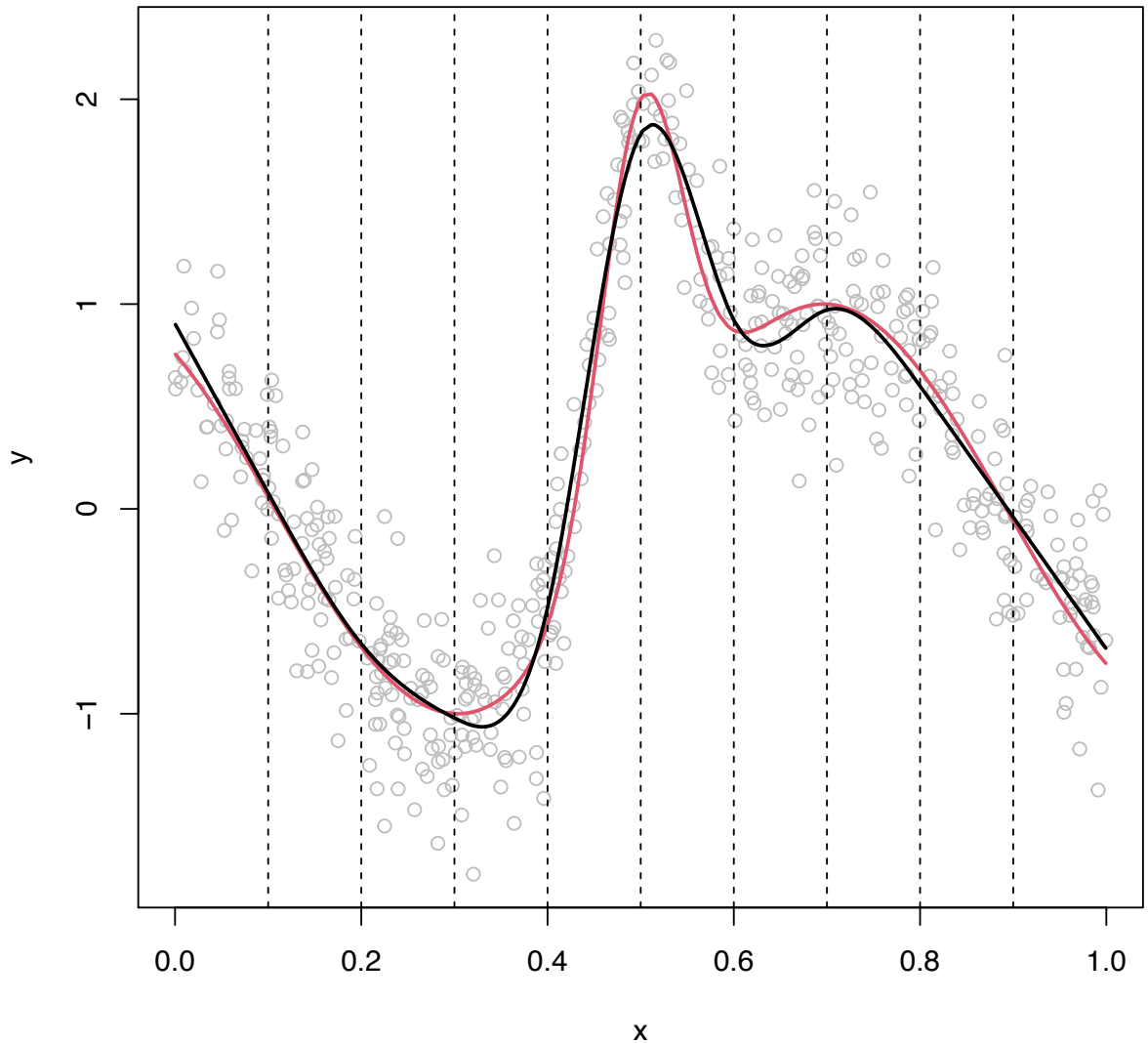


Regression cubic spline

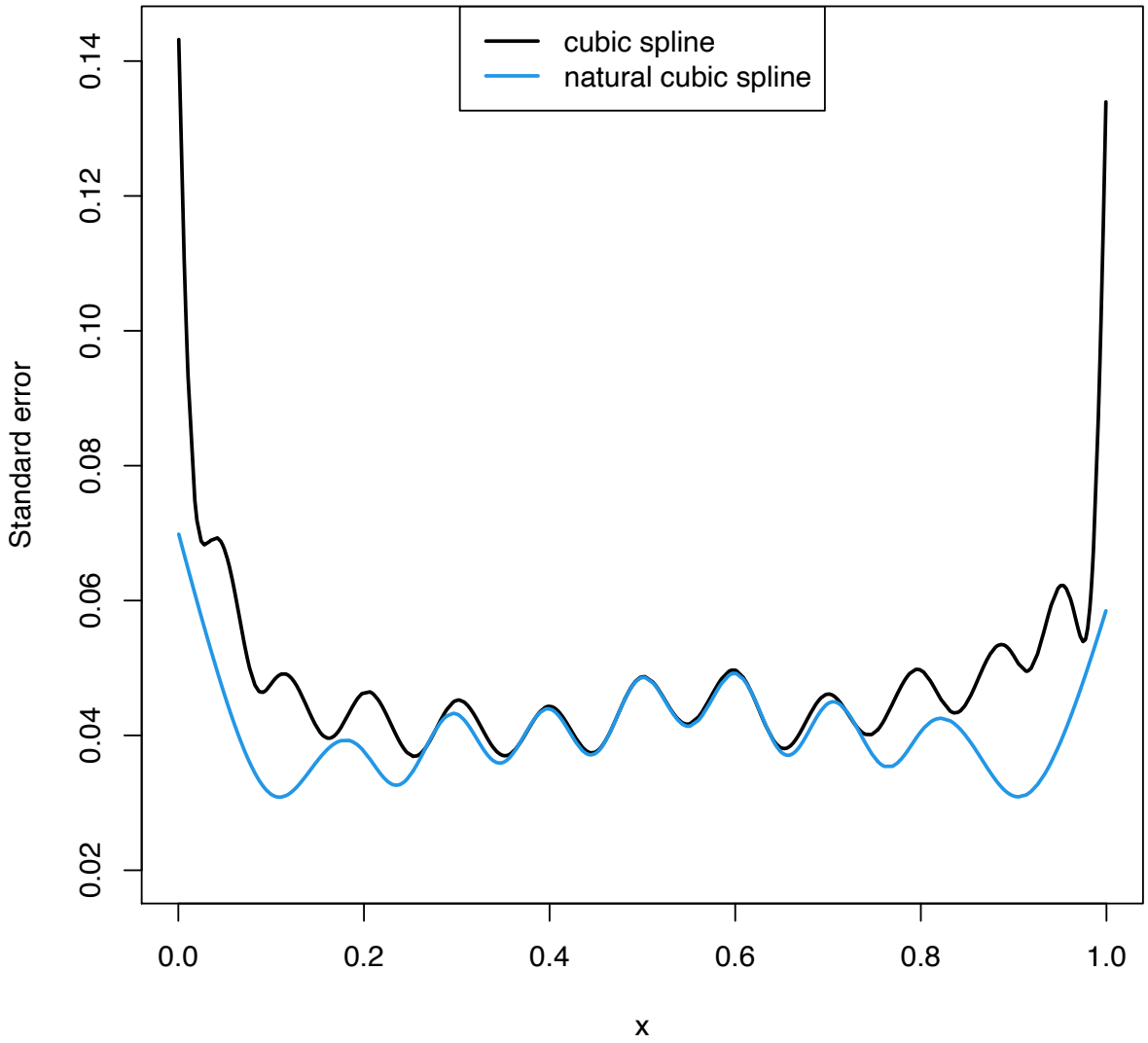
Natural cubic splines

- The most commonly used are cubic splines. Splines exhibit erratic behaviour for values less than ξ_1 and larger than ξ_K
- A natural cubic spline adds additional constraints, namely that the function is linear beyond the boundary knots.
- This frees up 4 degrees of freedom: a natural cubic spline with K knots is represented by K basis functions:

$$
\begin{aligned}
B_1(x) &= 1 \\
B_2(x) &= x \\
B_{j+2}(x) &= \frac{(x-\xi_j)_+^3 - (x-\xi_K)_+^3}{\xi_K - \xi_j} - \frac{(x-\xi_{K-1})_+^3 - (x-\xi_K)_+^3}{\xi_K - \xi_{K-1}} \\
& \qquad j = 1, \ldots, K-2
\end{aligned}
$$



Natural cubic spline



Standard error of the fit

- A set of n points (x_i, y_i) can be exactly interpolated using a natural cubic spline with the $x_1 < \ldots < x_n$ as knots. The interpolating natural cubic spline is unique.
- Amongst all functions on $[a, b]$ which are twice continuously differentiable and which interpolate the set of points (x_i, y_i), a natural cubic spline with knots at the x_i yields the smallest roughness penalty

$$\int_a^b (f''(x))^2 dx$$

- $f''(x)$ is the second derivative of f with respect to x - it would be zero if f were linear, so this measures the curvature of f at x.

Smoothing splines

Smoothing spline

- Smoothing splines circumvent the problem of knot selection by performing regularized regression over the natural spline basis, placing knots at all inputs $x_1, \ldots, x_n$
- With inputs $x_1 < \ldots < x_n$ contained in an interval $[a, b]$, the minimiser of

$$\hat{f} = \arg\min_{f \in C_2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx$$

  amongst all twice continuously differentiable functions on $[a, b]$ is given by a a natural cubic spline with knots in the unique x_i

- The previous result tells us that we can choose natural cubic spline basis $B_1, \ldots, B_n$ with knots $\xi_1 = x_1, \ldots, \xi_n = x_n$ and solve

$$\hat{\beta}_\lambda = \arg\min_\beta \sum_{i=1}^{n} (y_i - \sum_{j=1}^{n} \beta_j B_j(x_i))^2 + \lambda \int_a^b \left(\sum_{j=1}^{n} \beta_j B_j''(x) \right)^2 dx$$

  to obtain the smoothing spline estimate $\hat{f}(x) = \sum_{i=1}^{n} \hat{\beta}_j B_j(x)$
- Rewriting

$$\hat{\beta}_\lambda = \arg\min_\beta \|y - B\beta\|^2 + \lambda \beta^t \Omega \beta$$

  where $B_{ij} = B_j(x_i)$ and $\Omega_{jk} = \int B_j''(x) B_k''(x) dx$, shows the smoothing spline problem to be a type of generalized ridge regression problem with solution

$$\hat{\beta}_\lambda = (B^t B + \lambda \Omega)^{-1} B^t y$$

- Fitted values in Reinsch form

$$\begin{aligned}\hat{y} &= B(B^tB + \lambda\Omega)^{-1}B^ty \\ &= (I_n + \lambda K)^{-1}y\end{aligned}$$

  where $K = (B^t)^{-1}\Omega B^{-1}$ does not depend on λ, and $S = (I_n + \lambda K)^{-1}$ is the $n \times n$ *smoothing matrix*

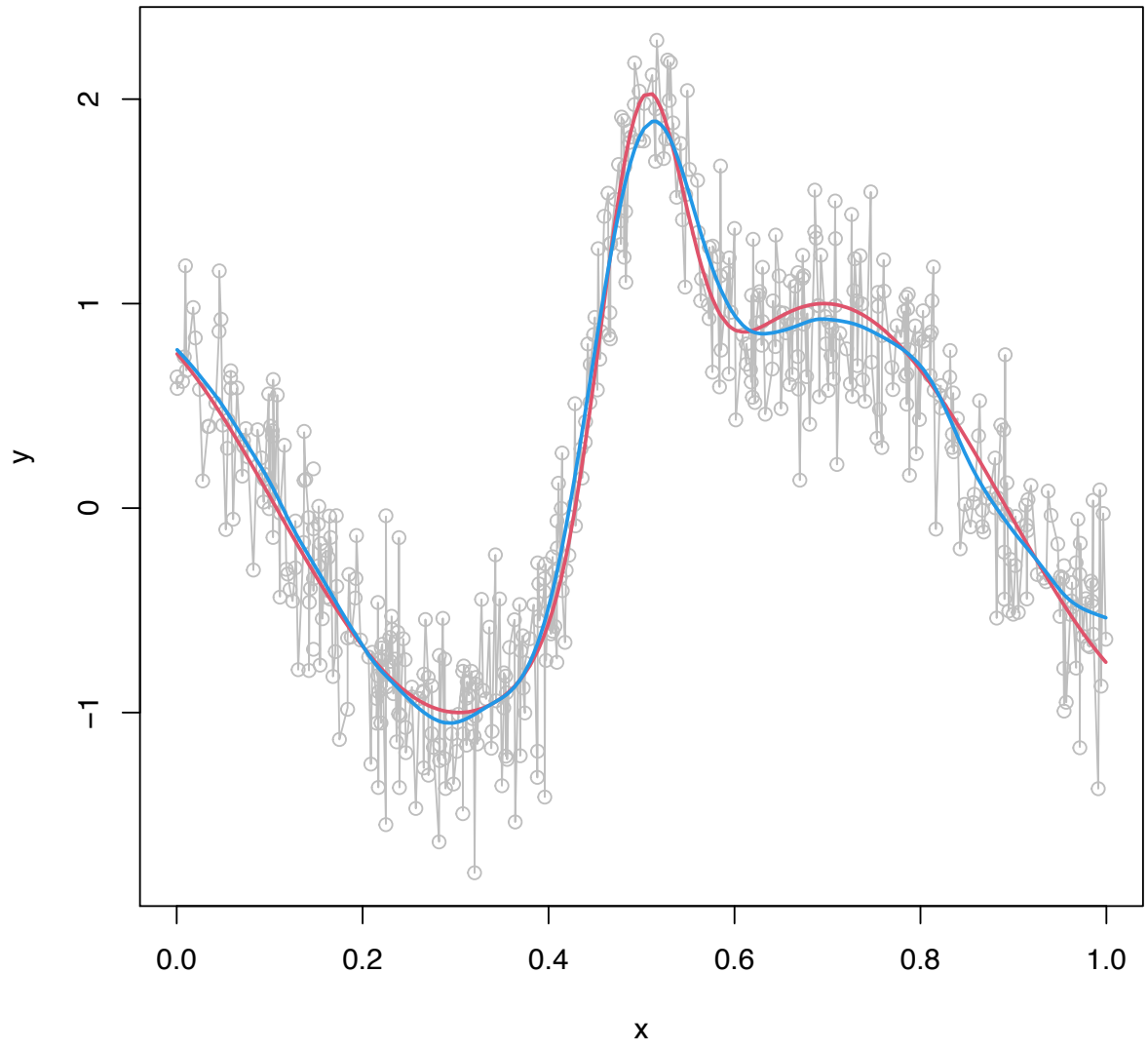
- Leave-one-out cross validation

$$\text{LOO} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - S_{ii}}\right)^2$$

- Generalized cross validation

$$\text{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - \text{tr}(S)/n}\right)^2$$

  where $\text{tr}(S)$ is the effective degrees of freedom



`smooth.spline` result with $\lambda = 0$ and 6.9e-15 by LOO

Reinsch original solution

- The original Reinsch (1967) algorithm solves the constrained optimization problem

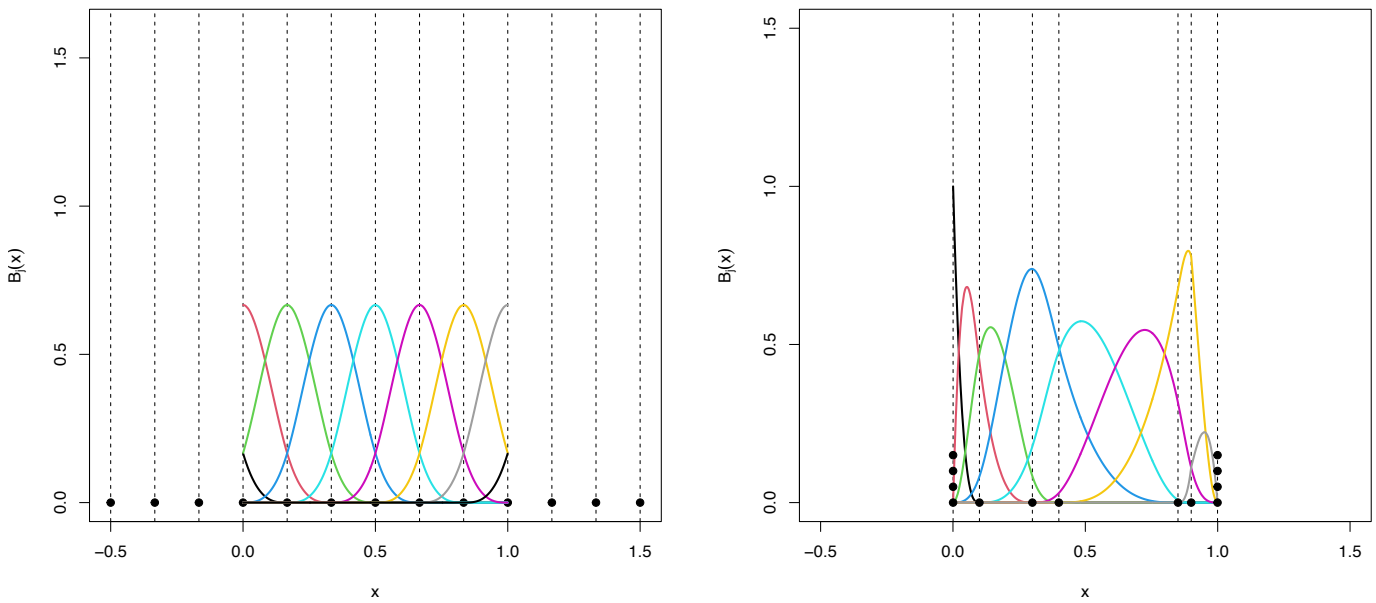
$$\hat{f} = \arg\min_{f \in C_2} \int_a^b (f''(x))^2 dx \text{ such that } \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq c$$

- The previous formulation with a Lagrange parameter on the integral smoothing term instead of the least squares term is equivalent
- See `casl_smspline` implementation in Section 2.6 of CASL

P-splines

B-spline basis

- The truncated power basis suffers from computational issues. The *B*-spline basis is a re-parametrization of the truncated power basis spanning an equivalent space
- The appearance of *B*-splines depends on their knot spacing, e.g.
  - uniform *B*-splines on equidistant knots;
  - non-uniform *B*-splines on unevenly spaced knots and repeated boundary;



Left plot: uniform cubic B-splines with equidistant knots

Right plot: non-uniform cubic B-splines with unevenly spaced knots and duplicated boundary knots

B-spline basis

- B-splines can be computed as differences of truncated power functions
- The general formula for equally-spaced knots is

$$B_j(x) = \frac{(-1)^{M+1}\Delta^{M+1} f_j(x, M)}{h^M M!}$$

satisfying

$$\sum_j B_j(x) = 1$$

where $f_j(x, M) = (x - \xi_j)_+^M$, h is is the distance between knots and Δ^O is the Oth order difference with
$\Delta f_j(x, M) = f_j(x, M) - f_{j-1}(x, M)$,
$\Delta^2 f_j(x, M) = \Delta(\Delta f_j(x, M)) = f_j(x, M) - 2f_{j-1}(x, M) + f_{j-2}(x, M)$

P-splines

- There is an intermediate solution between regression and smoothing splines, proposed more recently by Eilers and Marx (1996)
- P-splines use a basis of (quadratic or cubic) B-splines, B, computed on x and using equally-spaced knots. Minimize

$$\|y - B\beta\|^2 + \lambda\|D\beta\|^2$$

  where $D = \Delta^O$ is the matrix of Oth order differences, with $\Delta\beta_j = \beta_j - \beta_{j-1}$, $\Delta^2\beta_j = \Delta(\Delta\beta_j) = \beta_j - 2\beta_{j-1} + \beta_{j-2}$ and so on for higher O. Mostly $O = 2$ or $O = 3$ is used.
- Minimization leads to the system of equations

$$(B^tB + \lambda D^tD)\hat{\beta} = B^ty$$

Cross-validation

- We have that $\hat{y} = B(B^tB + \lambda D^tD)^{-1}B^ty = Sy$
- $$\text{LOO} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - S_{ii}}\right)^2$$
- $$\text{GCV} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{(1 - \text{tr}(S)/n)^2}$$
- We can compute the trace of R without actually computing its diagonal, using

  $$\text{tr}(S) = \text{tr}((B^tB + P)^{-1}B^tB) = \text{tr}(I_n - (B^tB + P)^{-1}P)$$

  where $P = \lambda D^tD$

Classical vs high-dimensional theory

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Sur, Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. Proceedings of the National Academy of Sciences, 116, 14516–14525

Classical theory

- It concerns the behaviour when the *sample size* $n \to \infty$
- Suppose $y_1, \ldots, y_n \overset{i.i.d.}{\sim} y \in \mathbb{R}^p$ with mean $\mu = \mathbb{E}(y)$ and finite variance $\Sigma = \mathbb{V}\mathrm{ar}(y)$
- *Law of large numbers*: the sample mean $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$ converges in probability to μ
- *Central limit theorem*: the rescaled deviation $\sqrt{n}(\hat{\mu}_n - \mu)$ converges in distribution to a centered Gaussian with covariance matrix Σ
- *Consistency of maximum likelihood estimation*
- Etc.

Suppose that we are given $n = 1000$ samples from a statistical model in $p = 500$ dimensions

Will theory that requires $n \to \infty$ with the dimension p remaining fixed provide useful predictions?

High-dimensional data

- The data sets arising in many parts of modern science have a "high-dimensional flavor", with p on the same order as, or possibly larger than n

$$p \gg n$$

- Classical "large n, fixed p" theory fails to provide useful predictions
- Classical methods can break down dramatically in high-dimensional regimes

Neyman-Scott problem (Bartlett, 1937)

- Let (x_i, y_i) be independent $N(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$
- The MLE for μ_i is $\hat{\mu}_i = (x_i + y_i)/2$
- The MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} s_i^2$$

  where

$$s_i^2 = [(x_i - \hat{\mu}_i)^2 + (y_i - \hat{\mu}_i)^2]/2 = (x_i - y_i)^2/4$$

- Then

$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2/2$$

  and the MLE is inconsistent as $n \to \infty$

High-dimensional logistic regression

- The logistic model assumes that the probability of a case conditional on the covariates is given by

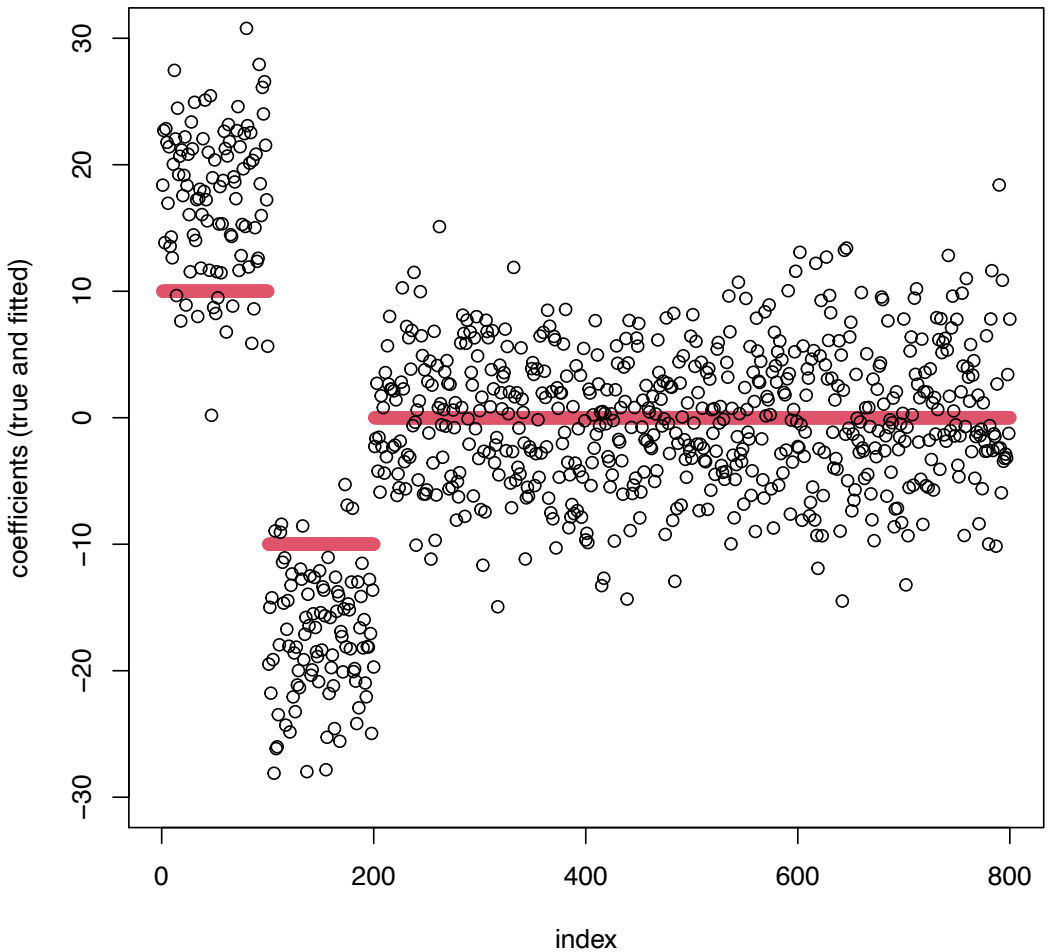
$$\mathrm{P}(y_i = 1|x_i) = \rho(x_i^t \beta)$$

  where $\rho(z) = e^z/(1 + e^z)$ is the sigmoidal function
- When p is fixed and $n \to \infty$, the MLE obeys

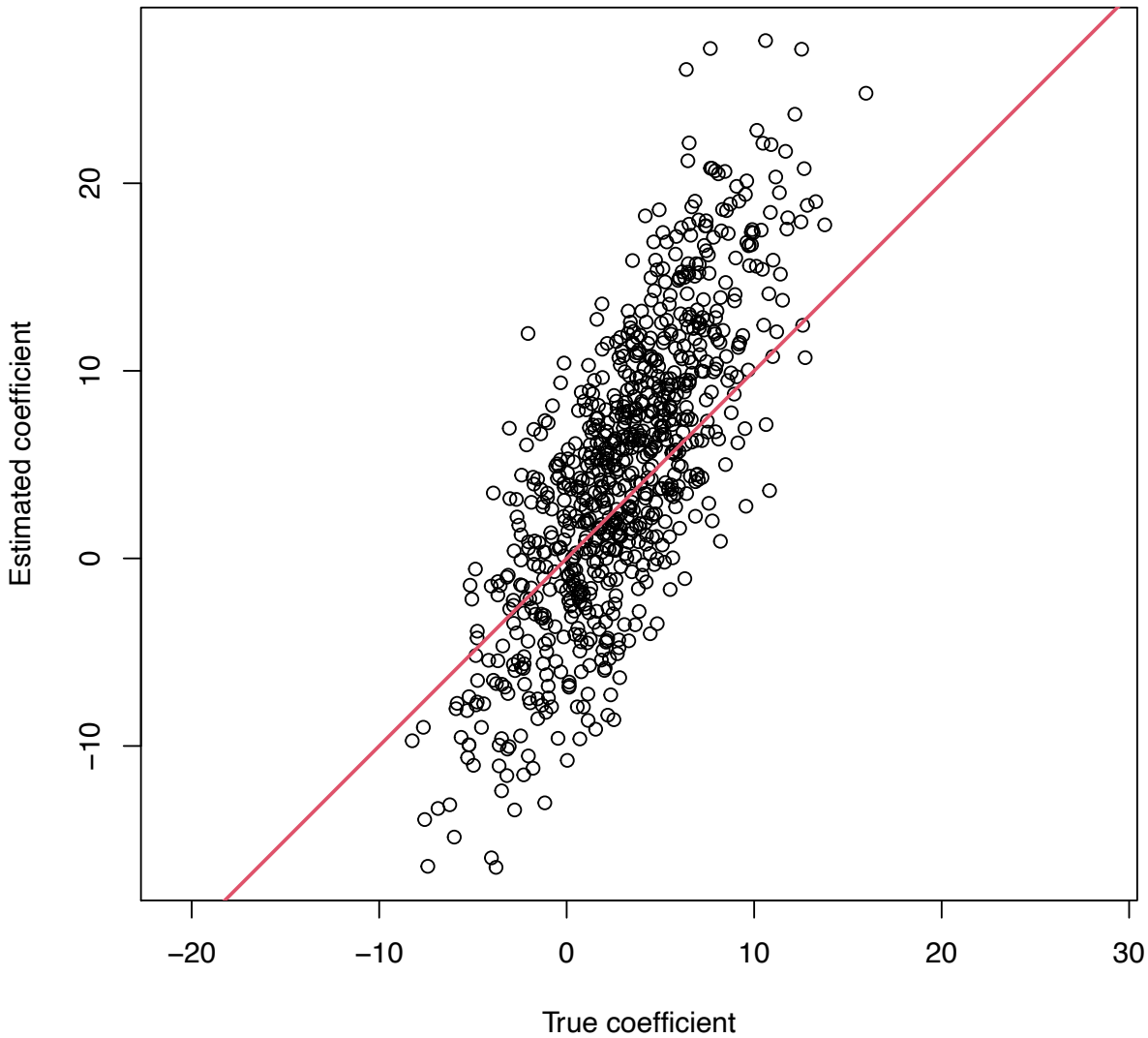
$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathrm{d}} N(0, \mathcal{I}_\beta^{-1})$$

  where $\mathcal{I}_\beta$ is the $p \times p$ Fisher information matrix evaluated at the true β
- Does this approximation holds in high-dimensional settings where p is not vanishingly small compared with n?
- We set $n = 4000$ and $p = 800$, so that $p/n = 1/5$, with $x_{ij} \overset{\text{iid}}{\sim} N(0, 1/n)$

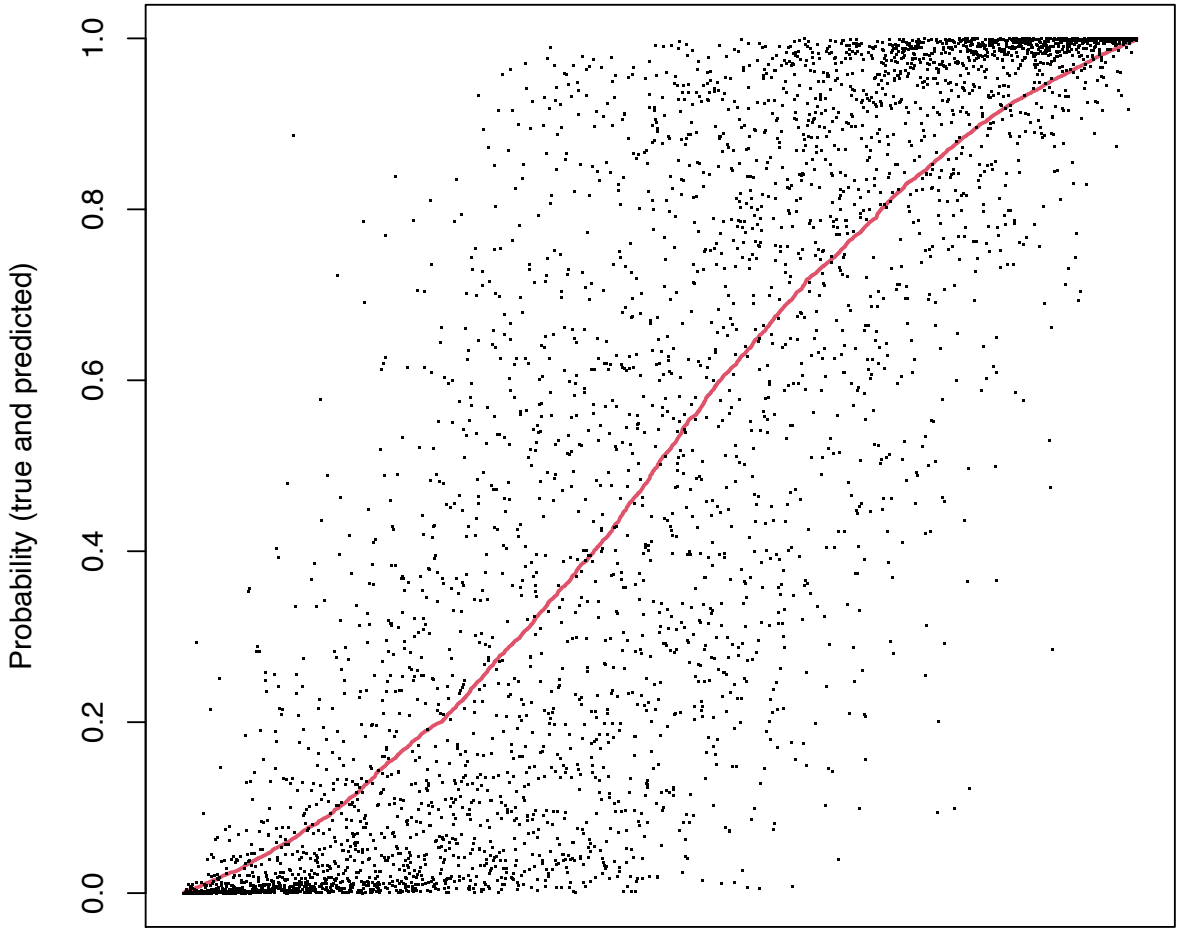
Unbiasedness?



$$\beta_1 = \ldots = \beta_{100} = 10, \beta_{101} = \ldots = \beta_{200} = -10, \beta_{201} = \ldots = \beta_{800} = 0$$



$$\beta_i \overset{\text{iid}}{\sim} N(3, 16)$$



$$\beta_i \overset{\text{iid}}{\sim} N(3, 16)$$

Distribution of the LRT?

- Testing $H_j : \beta_j = 0$ against $\beta_j \neq 0$
- Log-likelihood ratio statistic

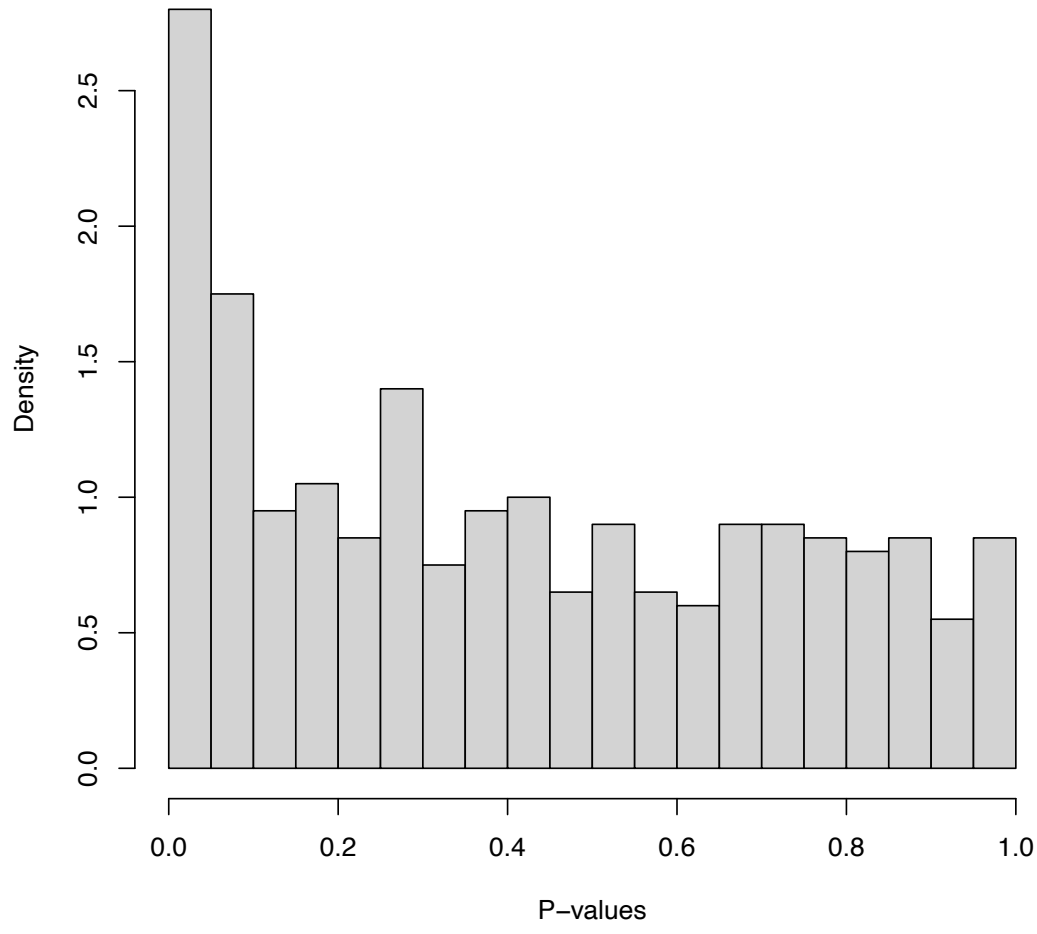
$$\Lambda_j = \ell(\hat{\beta}) - \ell(\hat{\beta}_{-j})$$

- Wilks Theorem: with p fixed and $n \to \infty$, the test has asymptotic null distribution

$$2\Lambda_j \xrightarrow{\text{d}} \chi^2_1$$

- If the χ^2 approximation were true, then we would expect to observe uniformly distributed null p-values

Distribution of the null p-values?



Half of β are i.i.d. $N(7, 1)$, the other half o

Linear discriminant analysis in high-dimensions

Classification problem

- Let's turn to the classification problem involving the allocation of the observed unit x to one of two classes A and B
- For a Bayesian analysis suppose that the prior probabilities are $\pi_A \equiv \mathrm{P}(y = A)$ and $\pi_B \equiv \mathrm{P}(y = B)$ with $\pi_A + \pi_B = 1$. Then the posterior probabilities satisfy

$$\frac{\mathrm{P}(y = B|x)}{\mathrm{P}(y = A|x)} = \frac{\pi_B}{\pi_A} \frac{f_B(x)}{f_A(x)}$$

  giving the class with the higher posterior probability
- As a special case, suppose that the two classes are distributed as multivariate Gaussians $x_A \sim N(\mu_A, I_p)$ and $x_B \sim N(\mu_B, I_p)$, with $\pi_A = \pi_B = 1/2$

Optimal decision

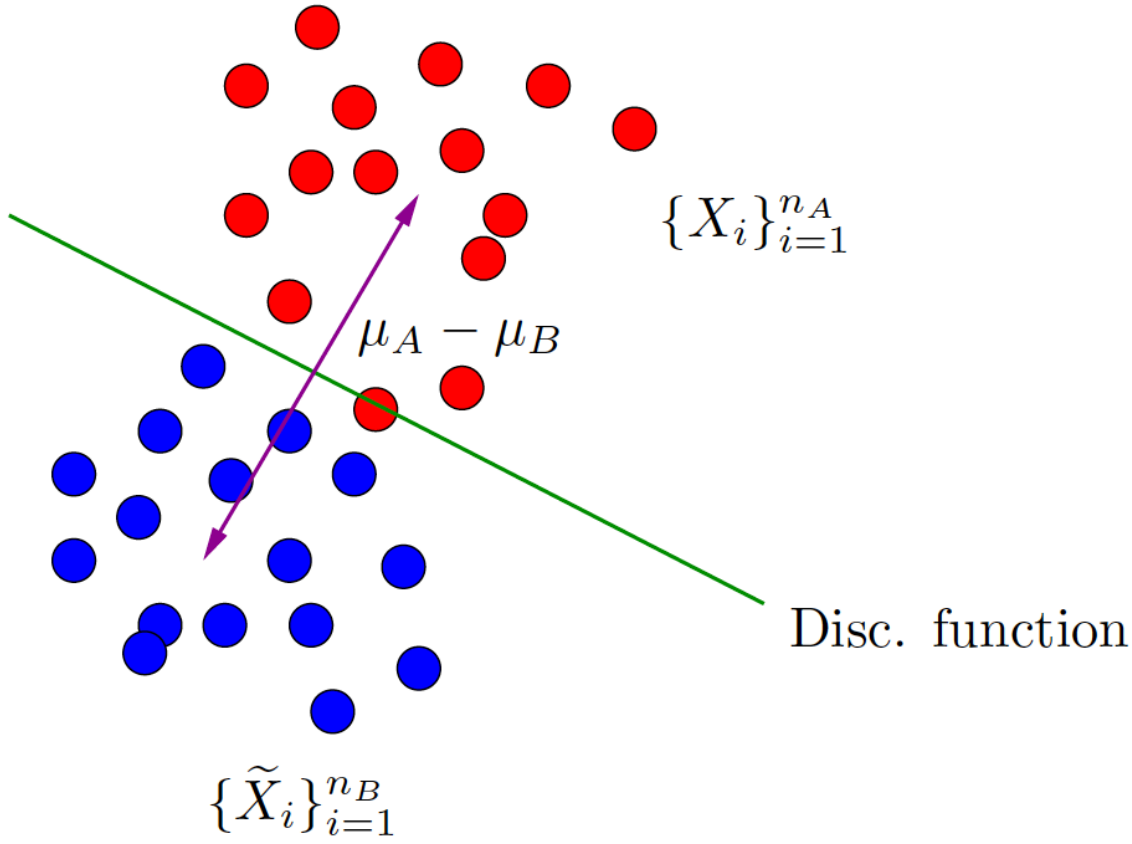
- The optimal decision rule is to threshold the log-likelihood ratio

$$\Psi(x) = \langle \mu_A - \mu_B, \left(x - \frac{\mu_A + \mu_B}{2} \right) \rangle$$

  where $\langle x, z \rangle = x^t z = \sum_{j=1}^{p} x_j z_j$ denotes the Euclidean inner product in $\mathbb{R}^p$
- If $\Psi(x) > 0$ then classify A, otherwise B
- Error probability of the optimal rule:

$$\mathrm{Err}(\Psi) = \frac{1}{2}\mathrm{P}(\Psi(x_A) < 0) + \frac{1}{2}\mathrm{P}(\Psi(x_B) \geq 0) = \Phi\Big(-\frac{\gamma}{2} \Big)$$

  where $\gamma = \|\mu_A - \mu_B\|$ and Φ is the cdf of a standard normal variable



$$\langle \mu_A - \mu_B, \left(x - \frac{\mu_A + \mu_B}{2} \right) \rangle = 0$$

source: Wainwright

Linear Discriminant Analysis

- Fisher's LDA: uses the plug-in principle based on n_A samples from class A and n_B samples from class B

$$\hat{\Psi}(x) = \langle \hat{\mu}_A - \hat{\mu}_B, x - \frac{\hat{\mu}_A + \hat{\mu}_B}{2} \rangle$$

- Error probability of LDA (is itself a random variable)

$$\mathrm{Err}(\hat{\Psi}) = \frac{1}{2}\mathrm{P}(\hat{\Psi}(x_A) < 0) + \frac{1}{2}\mathrm{P}(\hat{\Psi}(x_B) \geq 0)$$

- Classical theory: if $(n_A, n_B) \to \infty$ and p remains fixed, then $\hat{\mu}_A \overset{prob.}{\to} \mu_A$, $\hat{\mu}_B \overset{prob.}{\to} \mu_B$ and the asymptotic error probability is $\mathrm{Err}(\hat{\Psi}) \overset{prob.}{\to} \mathrm{Err}(\Psi) = \Phi(-\gamma/2)$

– If $x \sim N(\mu, I_p)$, then

$$\begin{aligned}\hat{\Psi}(x) &= \langle \hat{\mu}_A - \hat{\mu}_B, x - \frac{\hat{\mu}_A + \hat{\mu}_B}{2} \rangle \\ &= \hat{d}^t(x - \hat{m}) \sim N(\hat{d}^t(\mu - \hat{m}), \hat{d}^t\hat{d})\end{aligned}$$

where $\hat{d} = \hat{\mu}_A - \hat{\mu}_B$ and $\hat{m} = \frac{\hat{\mu}_A + \hat{\mu}_B}{2}$

High-Dimensional Theory

- What happens if $(n_A, n_B, p) \to \infty$ with
  - $p/n_A \to \delta$ with $\delta \geq 0$
  - $p/n_B \to \delta$
  - $\|\mu_A - \mu_B\|_2 \to \gamma > 0$
- Kolmogorov (1960) showed that

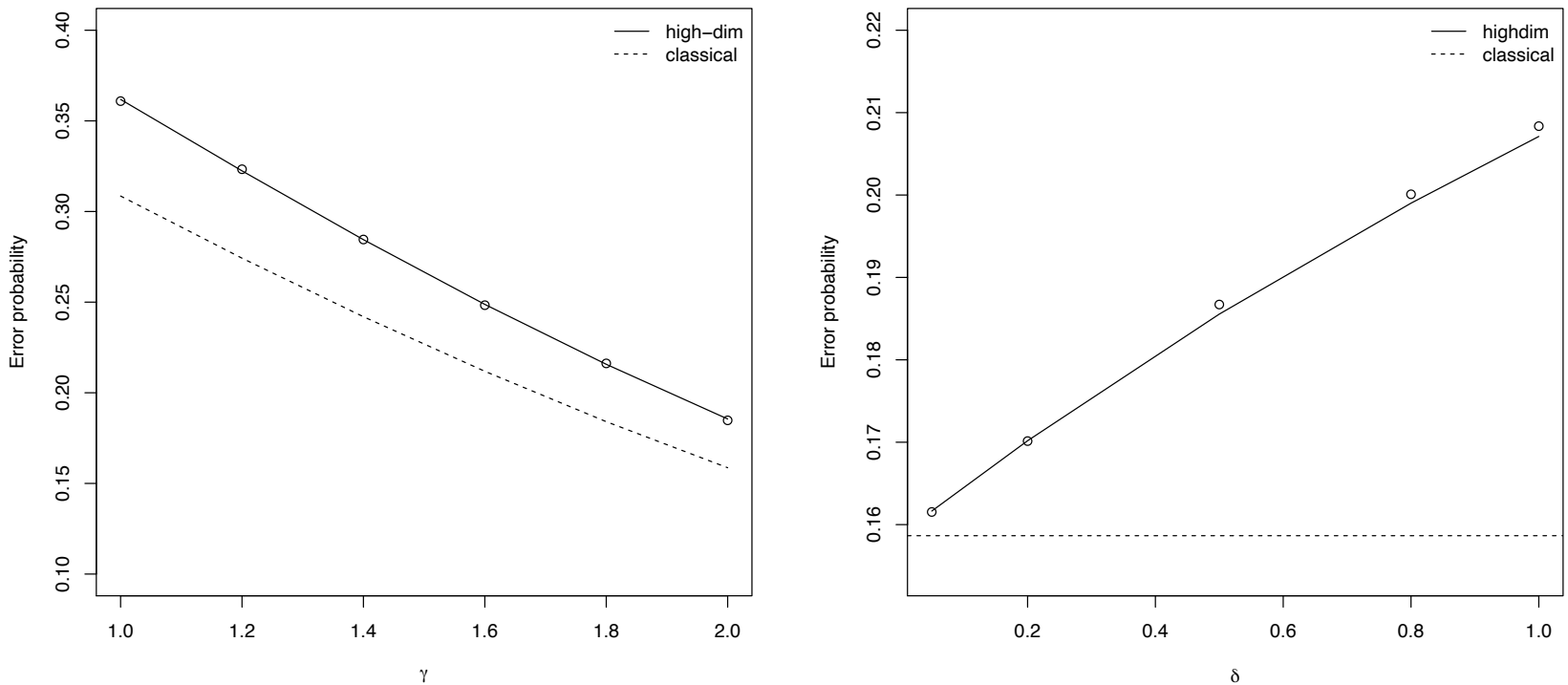
$$\mathrm{Err}(\hat{\Psi}) \overset{prob.}{\to} \Phi\Big(- \frac{\gamma^2}{2\sqrt{\gamma^2 + 2\delta}} \Big)$$

- If $p/n \to 0$, then the asymptotic error probability is $\Phi(-\gamma/2)$ as is predicted by classical theory
- If $p/n \to \delta > 0$, then the asymptotic error probability is strictly larger than $\Phi(-\gamma/2)$

The error probability of $\hat{\Phi}$, for the finite triple

$$(p, n_A, n_B) = (400, 800, 800)$$

is better described by the classical $\Phi(-\gamma/2)$, or the high-dimensional analog $\Phi(-\gamma^2/(2\sqrt{\gamma^2 + 2\delta}))$?



circles correspond to the empirical error probabilities, averaged over 10 trials

What can help us in high dimensions?

- An important fact is that high-dimensional phenomena are unavoidable
- If the ratio p/n stays bounded strictly above zero, then it is not possible to achieve the optimal classification rate
- Our only hope is that the data is endowed with some form of *low-dimensional structure*

- What is the underlying cause of the inaccuracy of the prediction for the LDA in high-dimensions?
- The squared Euclidean error

$$\|\hat{\mu} - \mu\|^2 = \sum_{j=1}^{p} (\hat{\mu}_j - \mu_j)^2$$

concentrates sharply around p/n, i.e. for $t \in (0, 1)$

$$\mathrm{P}\left(\left|\|\hat{\mu} - \mu\|^2 - \frac{m}{n}\right| \geq \frac{p}{n} t\right) = \mathrm{P}\left(\left|\frac{1}{p} \sum_{j=1}^{p} Z_j^2 - 1\right| \geq t\right) \leq 2e^{-\frac{pt^2}{8}}$$

where $Z_j = \sqrt{n}(\hat{\mu}_j - \mu_j) \sim N(0, 1)$; for the upper bound see Wainwright (2019), Example 2.11

Sparsity

- Suppose that the p-vector μ is *sparse*, with only s of its p entries being nonzero, for some sparsity parameter $s \ll p$
- If sparsity holds, we can obtain a better estimator by thresholding the sample means

$$\tilde{\mu}_j = \hat{\mu}_j \mathbb{1}\{|\hat{\mu}_j| > \lambda\}$$

where

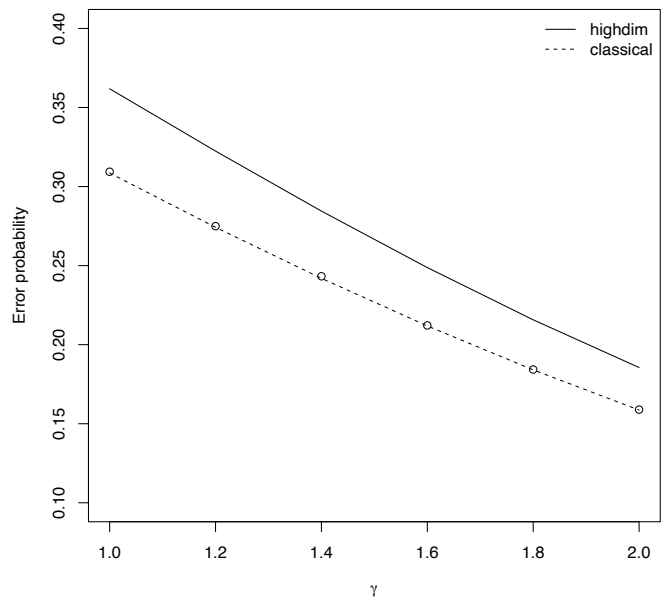
$$\lambda = \sqrt{\frac{2 \log p}{n}}$$

Thresholded mean

Suppose to replace $\hat{\mu}$ by the thresholded mean $\tilde{\mu}$, then

$$\tilde{\Psi}(x) = \langle \tilde{\mu}_A - \tilde{\mu}_B, x - \frac{\tilde{\mu}_A + \tilde{\mu}_B}{2} \rangle$$

approaches the optimal $\mathrm{Err}(\Psi)$ if $\log \binom{p}{s} / n \to 0$. For $s = 5$:



circles correspond to the empirical error probabilities, averaged over 10 trials

Sparse Modeling:
Best Subset and the Lasso

Statistical Learning

CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Tibshirani, Wasserman (2017). Sparsity, the Lasso, and Friends. Lecture notes on Statistical Machine Learning

Three norms: ℓ_0, ℓ_1 and ℓ_2

- Let's consider three canonical choices: the ℓ_0, ℓ_1 and ℓ_2 norms:

$$\|\beta\|_0 = \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\}, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$$

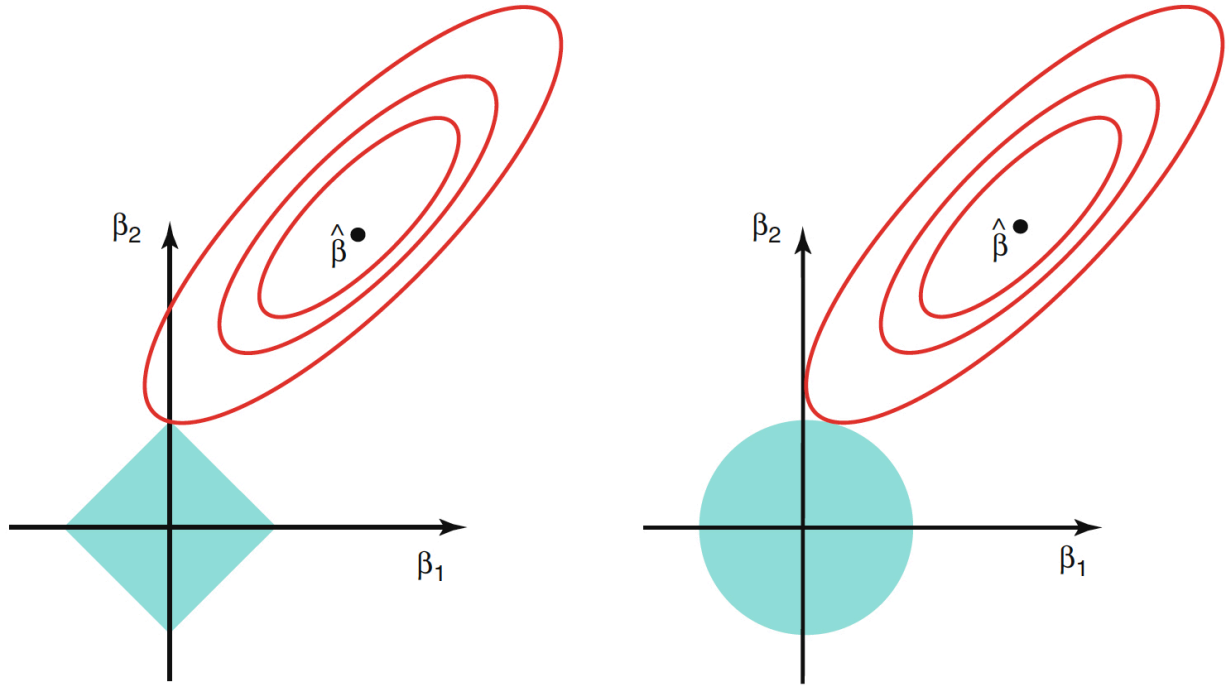
- ℓ_0 is not a proper norm: it does not satisfy positive homogeneity, i.e. $\|a\beta\|_0 \neq |a|\|\beta\|_0$ for $a \in \mathbb{R}$

Constrained form

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_0 \leq c \qquad \text{Best Subset Selection}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_1 \leq c \qquad \text{Lasso Regression}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_2^2 \leq c \qquad \text{Ridge Regression}$$



The “classic” illustration comparing lasso and ridge constraints.
From Chapter 3 of ESL

Sparsity

- *Signal sparsity* is the assumption that only a small number of predictors have an effect, i.e. have $\beta_j \neq 0$
- In this case we would like our estimator $\hat{\beta}$ to be sparse, meaning that $\hat{\beta}_j = 0$ for many components $j \in \{1, \ldots, p\}$
- Sparse estimators are desirable because perform variable selection and improve interpretability of the result
- The best subset selection and the lasso estimators are sparse, the ridge estimator is not sparse

Penalized form

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0 \qquad \text{Best Subset Selection}$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \qquad \text{Lasso Regression}$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \qquad \text{Ridge Regression}$$

- Suppose that $y \sim N(\mu, 1)$
- ℓ_0 penalty

$$\min_{\mu} \frac{1}{2}(y-\mu)^2 + \lambda \mathbb{1}\{\mu \neq 0\}, \qquad \hat{\mu} = H_{\sqrt{2\lambda}}(y)$$

  where $H_a(y) = y\mathbb{1}\{|y| > a\}$ is the hard-thresholding operator
- ℓ_1 penalty

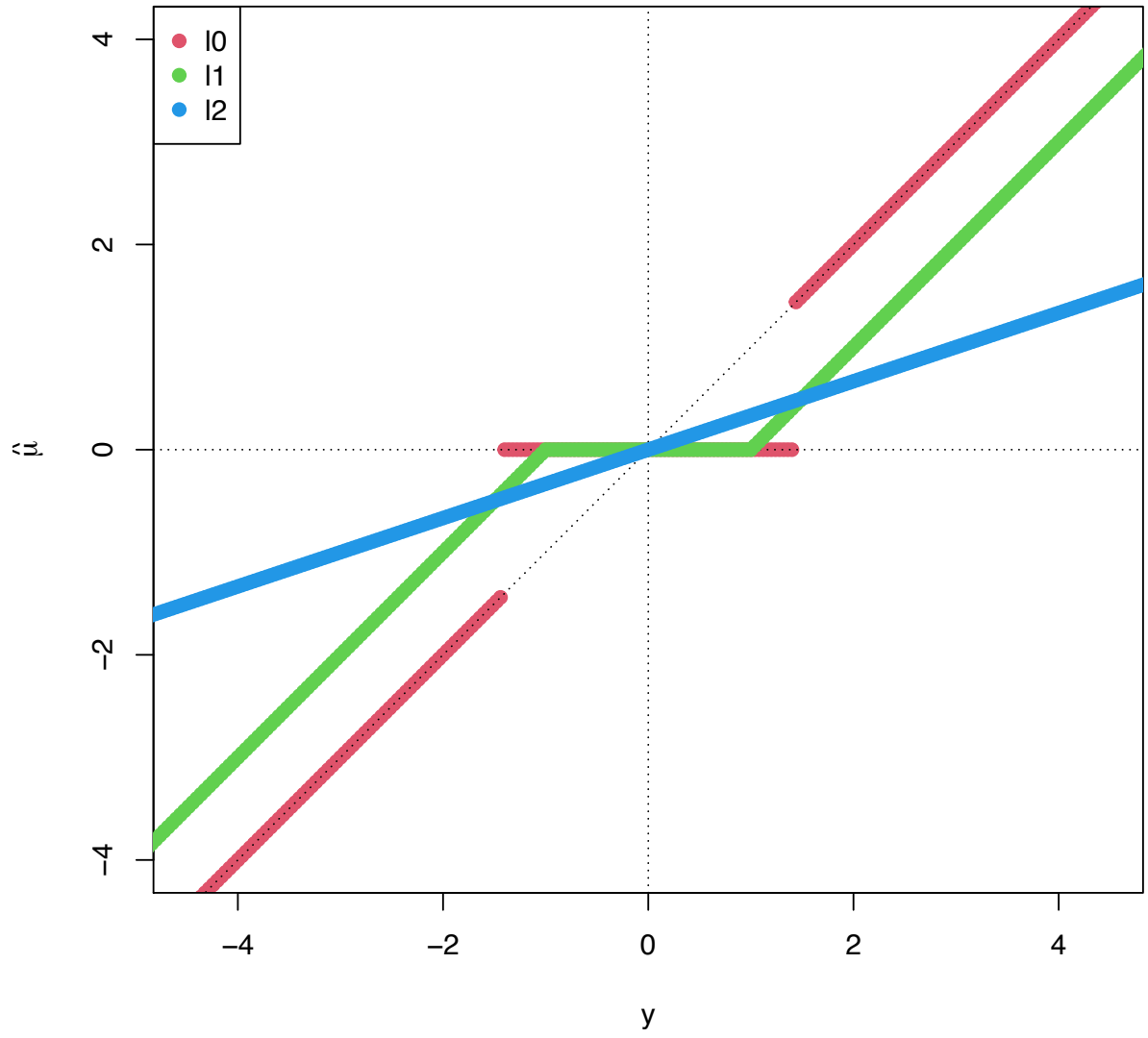
$$\min_{\mu} \frac{1}{2}(y-\mu)^2 + \lambda|\mu|, \qquad \hat{\mu} = S_\lambda(y)$$

  where

$$S_a(y) = \begin{cases} y-a & \text{if } y > a \\ 0 & \text{if } -a \leq y \leq a \\ y+a & \text{if } y > a \end{cases}$$

  is the soft-thresholding operator
- ℓ_2 penalty

$$\min_{\mu} \frac{1}{2}(y-\mu)^2 + \lambda\mu^2, \qquad \hat{\mu} = \left(\frac{1}{1+2\lambda}\right)y$$



$\lambda = 1$

Hard and soft thresholding

- ℓ_0 penalty creates a zone of sparsity but it is discontinuous (hard thresholding)
- ℓ_1 penalty creates a zone of sparsity but it is continuous (soft thresholding)
- ℓ_2 penalty creates a nice smooth estimator but it is never sparse

Orthogonal case

- Suppose $X^tX = I_p$
- OLS estimator

$$\hat{\beta} = X^t y$$

- BSS estimator

$$\hat{\beta} = H_{\sqrt{2\lambda}}(X^t y)$$

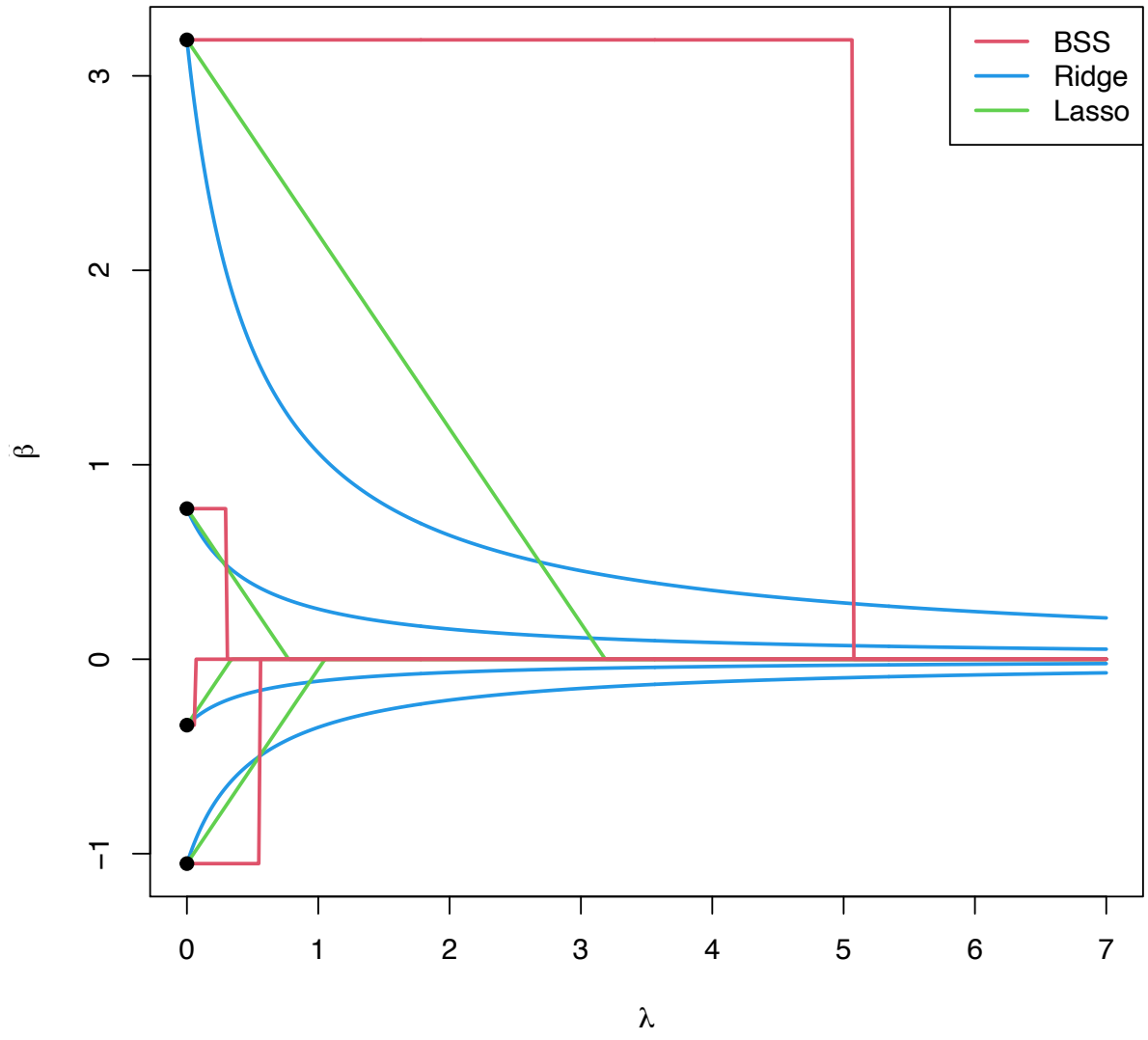
- Lasso estimator

$$\hat{\beta} = S_\lambda(X^t y)$$

- Ridge estimator

$$\hat{\beta} = \left(\frac{1}{1+2\lambda}\right) X^t y$$

where $H_a(\cdot)$, $S_a(\cdot)$ are the componentwise hard- and soft-thresholding operators



Solution paths of ℓ_0, ℓ_1 and ℓ_2 penalties as a function of λ

Convexity

- Consider using the norm $\|\beta\|_q = (\sum_{j=1}^{q} |\beta_j|^q)^{1/q}$ as a penalty. Sparsity requires $q \leq 1$ and convexity requires $q \geq 1$. The only norm that gives sparsity and convexity is $q = 1$
- The lasso and ridge regression are *convex optimization problems*, best subset selection is not
- The ridge regression optimization problem is always *strictly convex* for $\lambda > 0$
- The best subset selection optimization problem is N-P-complete because of its combinatorial complexity (there are 2^p subsets), the worst kind of non convex problem

Best Subset Selection

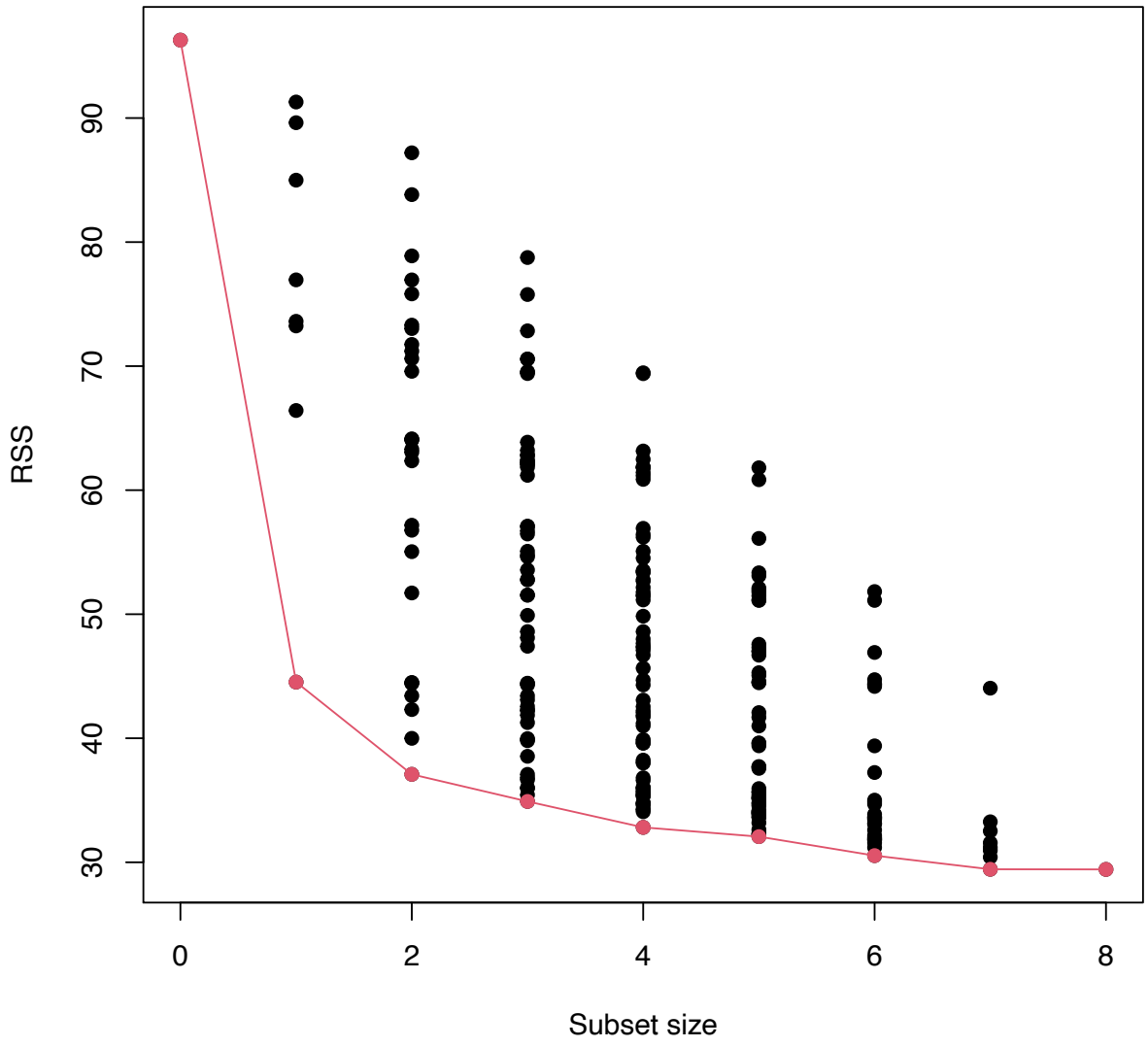
BSS algorithm

A natural approach is to consider all possible regression models each involving regressing the response on a different set of predictors

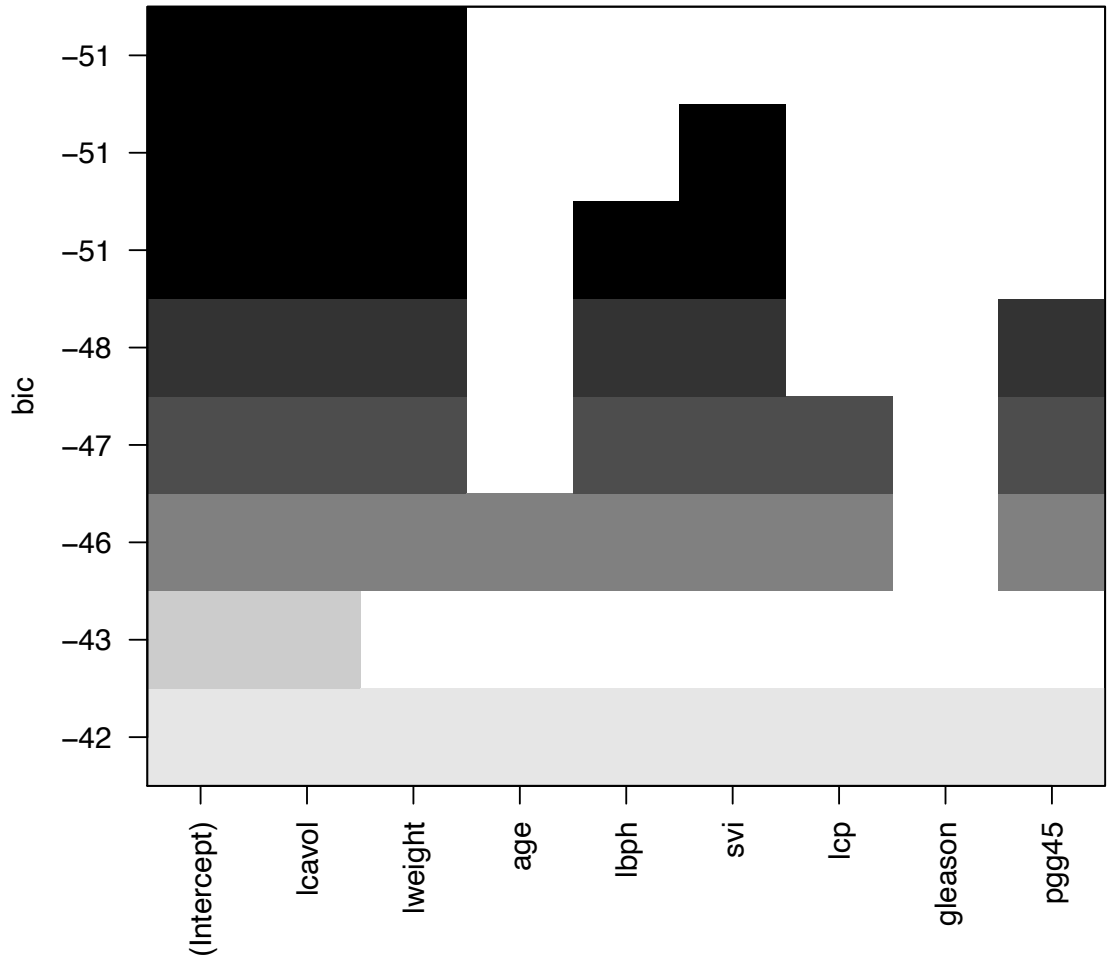
- Set B_0 as the null model (intercept only)
- For $k = 1, \ldots, p$
  1. Fit all $\binom{p}{k}$ models that contain exactly k predictors
  2. Pick the best among these $\binom{p}{k}$ models, and call it B_k, where best is defined having the smallest residual sum of squares
- Select a single best model from among $B_0, B_1, \ldots, B_p$ (e.g. using Cp, BIC, Cross-Validation, validation set, etc.)

prostate

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	RSS
B_1	*								44.53
B_2	*	*							37.09
B_3	*	*			*				34.91
B_4	*	*		*	*				32.81
B_5	*	*		*	*			*	32.07
B_6	*	*		*	*	*		*	30.54
B_7	*	*	*	*	*	*		*	29.44
B_8	*	*	*	*	*	*	*	*	29.43



All possible subset models for the prostate cancer example



BIC Best Subset = B_2

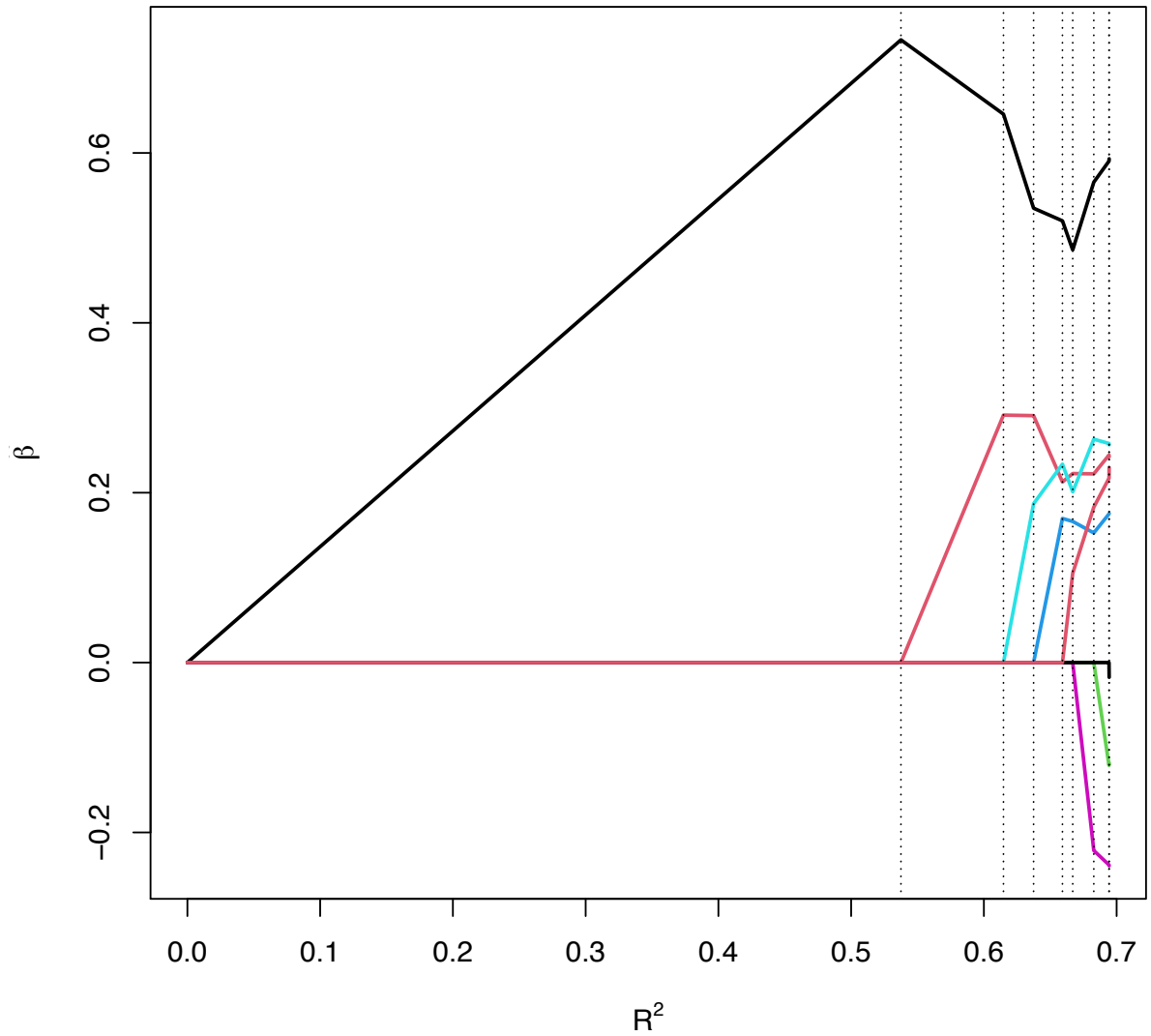
Computational bottleneck

- Furnival and Wilson (1974) and Hofmann et al. (2006) solve with $p \approx 30$ by using branch and bound algorithms. Implemented in the R packages `leaps` and `lmSubsets`
- Bertsimas et al. (2016) solve with $p \approx 100$ by using a mixed integer quadratic program along with the gurobi solver. Implemented in the R package `bestsubset`

Forward Stepwise Selection

Greedy forward algorithm, sub-optimal but feasible alternative to BSS and applicable when $p > n$

- Set S_0 as the null model (intercept only)
- For $k = 0, \ldots, \min(n-1, p-1)$:
  1. Consider all $p - k$ models that augment the predictors in S_k with one additional predictor
  2. Choose the best among these $p - k$ models and call it S_{k+1}, where best is defined having the smallest RSS
- Select a single best model from among $S_0, S_1, S_2, \ldots$ (e.g. using Cp, BIC, Cross-Validation, validation set, etc.)



Forward Stepwise solution path as a function of training R^2

The Lasso

- The name "lasso" was also introduced as an acronym for *Least Absolute Selection and Shrinkage Operator* (Tibshirani, 1996)
- The lasso finds the solution $(\hat{\alpha}, \hat{\beta})$ to the optimization problem

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - 1\alpha - X\beta\|_2^2 + \lambda \|\beta\|_1$$

- Typically, we first standardize the predictors X so that each column is centered ($(1/n)\sum_{i=1}^n x_{ij} = 0$) and has unit variance ($(1/n)\sum_{i=1}^n x_{ij}^2 = 0$)
- Without standardization, the lasso solutions would depend on the units (e.g., feet versus meters) used to measure the predictors. On the other hand, we typically would not standardize if the features were measured in the same units
- For convenience, we also assume that the outcome values y_i have been centered ($(1/n)\sum_{i=1}^n y_i = 0$). Centering is convenient, since we can omit the intercept term α in the lasso optimization, and given the solution $\hat{\beta}$

$$\hat{\alpha} = \bar{y} - \sum_{j=1}^p \bar{x}_j \hat{\beta}_j$$

- Lagrange form

$$\frac{1}{2n} \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

- Intercept term omitted (center / scale y and the columns of X)
- The solution satisfies the subgradient / Karush-Kuhn-Tuker conditions

$$\frac{1}{n} X^t (y - X\hat{\beta}) = \lambda s$$

where $s \in \partial \|\beta\|_1$, a subgradient of the ℓ_1 norm evaluated at $\hat{\beta}$

- The solution satisfies

$$-\frac{1}{n}\langle X_j, y - X\hat{\beta}\rangle + \lambda s_j = 0 \quad j = 1, \ldots, p$$

  where

$$s_j \in \begin{cases} 1 & \text{if } \hat{\beta}_j > 0 \\ [\text{-}1,1\,] & \text{if } \hat{\beta}_j = 0 \\ -1 & \text{if } \hat{\beta}_j < 0 \end{cases}$$

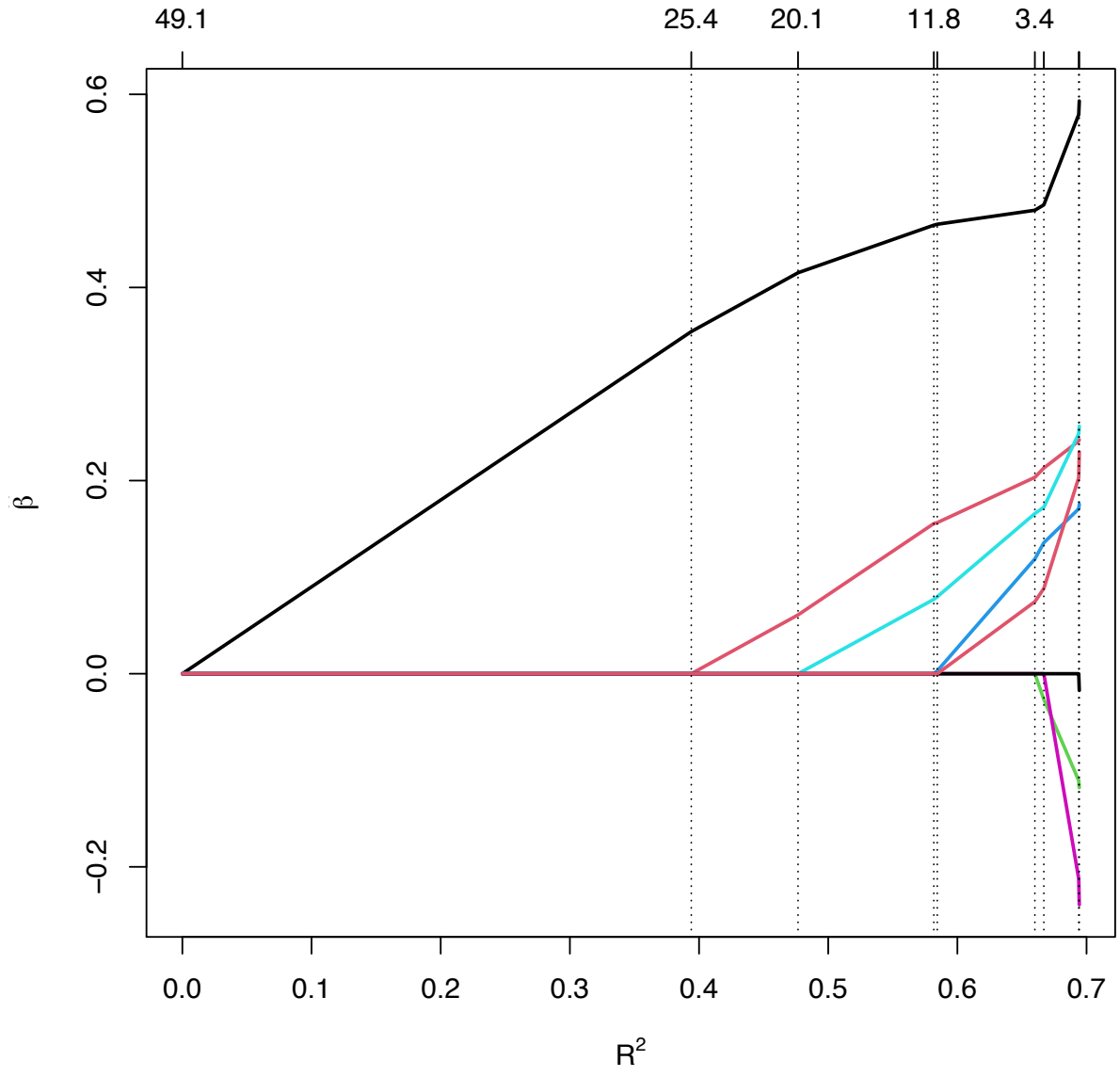
- Each of the variables in the model (with nonzero coefficient) has the same covariance with the residuals (in absolute value), i.e.

$$\frac{1}{n}|\langle X_j, y - X\hat{\beta}\rangle| = \lambda$$

- For all variables with zero coefficient

$$\frac{1}{n}|\langle X_j, y - X\hat{\beta}\rangle| \leq \lambda$$

- The coefficient profiles for the lasso are continuous and piecewise linear over the range of λ, with knots occurring whenever the *active set* changes, or the sign of the coefficients changes



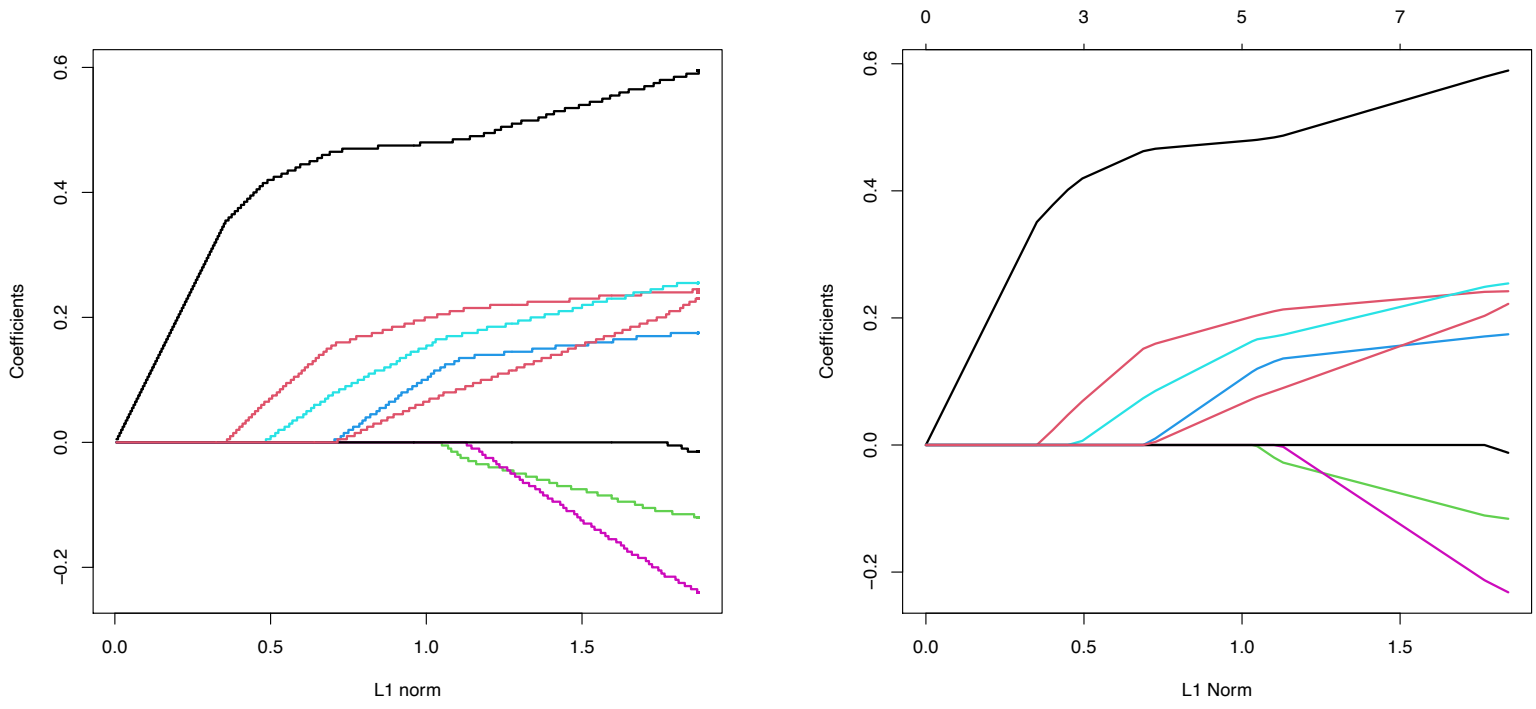
Lasso solution path as a function of training R^2

Boosting with componentwise linear least squares

- Response and predictors are standardized to have mean zero and unit norm
- Initialize $\hat{\beta}^{(0)} = 0$
- For $b = 1, \ldots, B$
  - compute the residuals $r = y - X\hat{\beta}^{(b-1)}$
  - find the predictor X_j most correlated with the residuals r
  - update $\hat{\beta}^{(b-1)}$ to $\hat{\beta}^{(b)}$ with

$$\hat{\beta}_j^{(b)} = \hat{\beta}_j^{(b-1)} + \epsilon \cdot s_j$$

    where s_j is the sign of the correlation
- This is known as *forward stagewise regression* and converges to the least squares solution when $n > p$
- Forward stagewise regression with infinitesimally small step-sizes, i.e. $\epsilon \to 0$, produces a set of solutions which is approximately equivalent to the set of Lasso solutions



Left: forward stagewise regression with $\epsilon = 0.005$; Right: lasso

Degrees of freedom

- Let $A(\lambda) = \{j \in \{1, \ldots, p\} : \beta_j(\lambda) \neq 0\}$ denotes the active set
- The degrees of freedom of the Lasso are the

$$\mathrm{df}(\lambda) = |A(\lambda)|$$

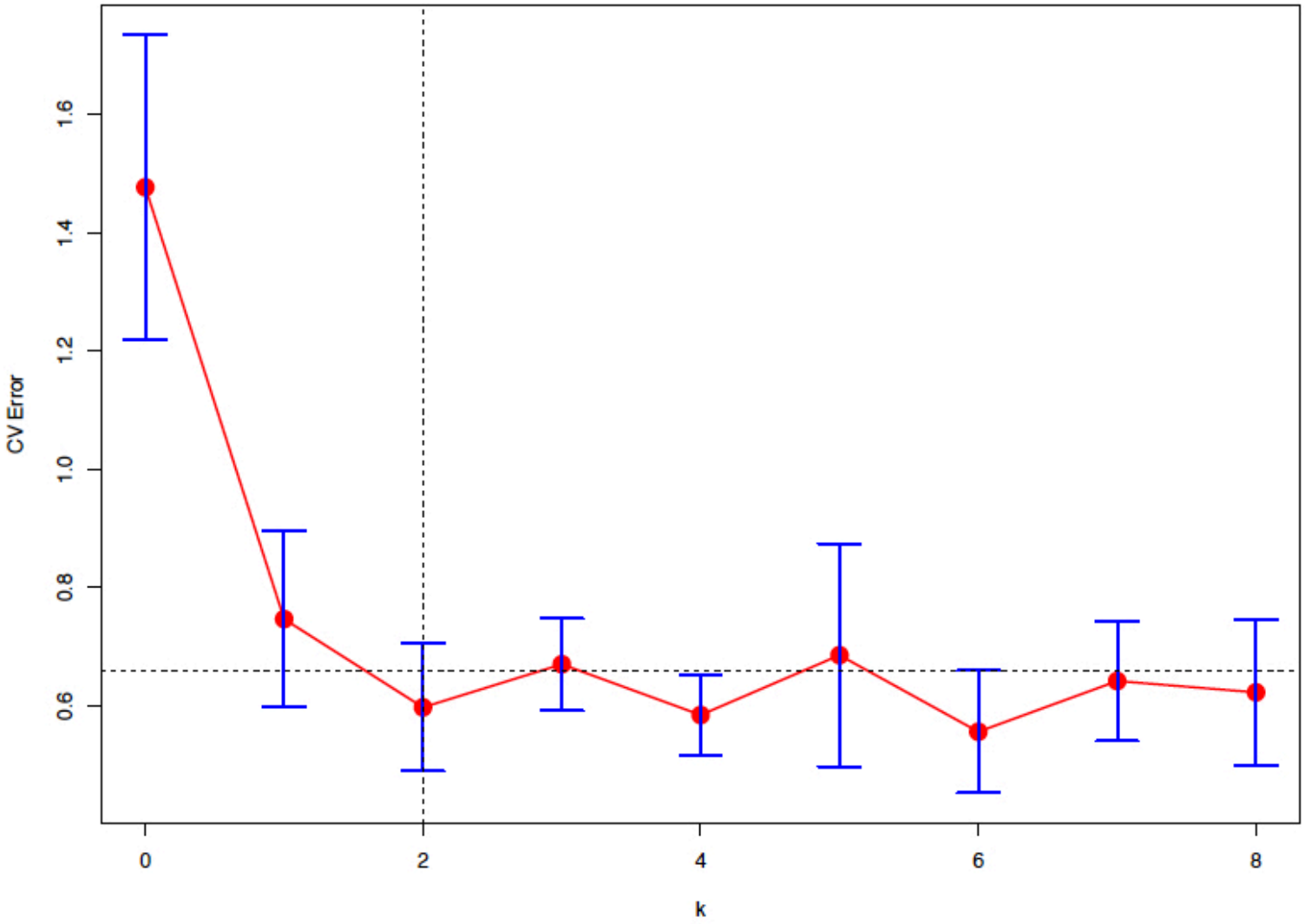
  i.e. the size of the active set

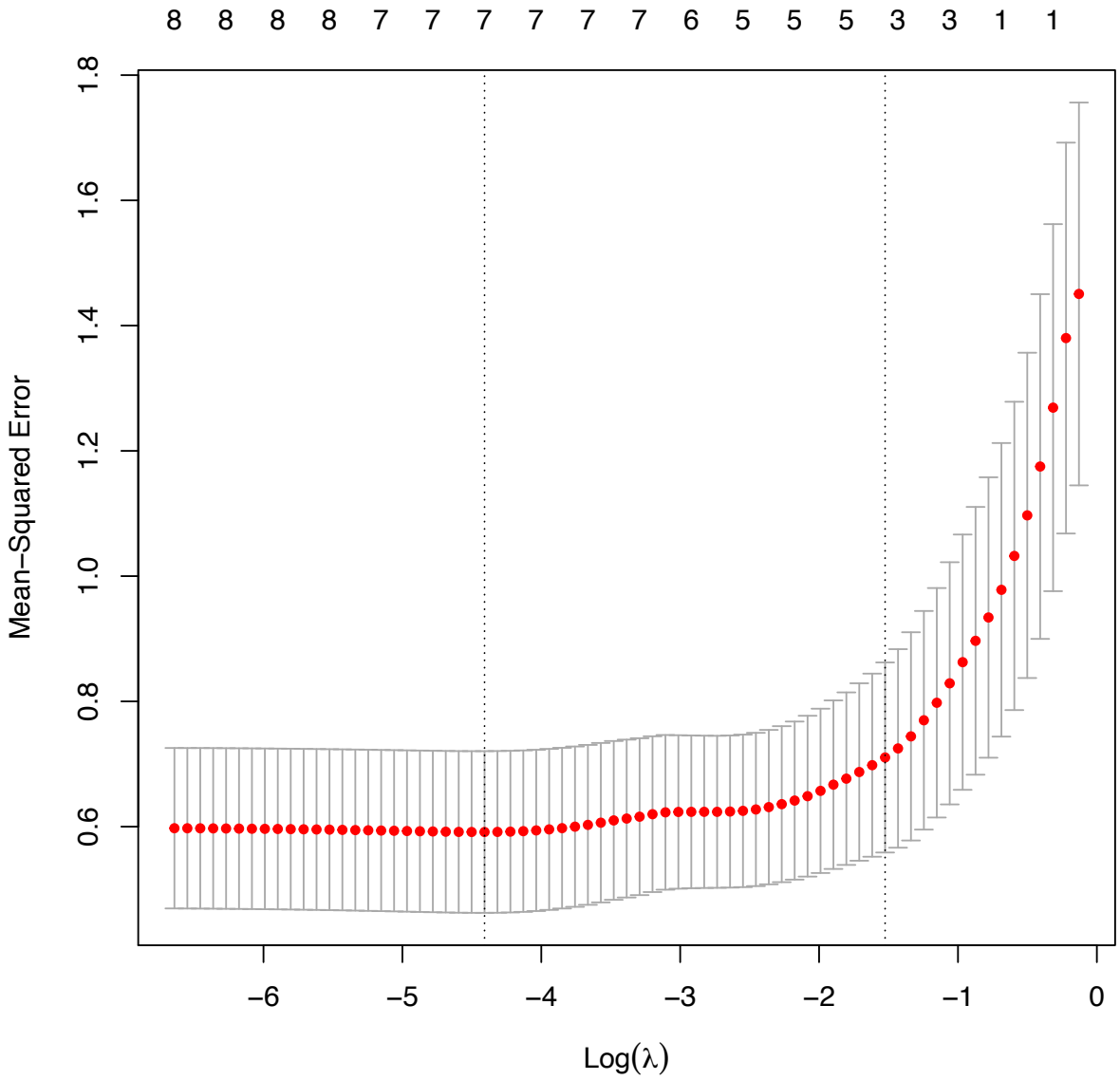
Cross-validation

- `lambda.min`: λ that minimize the cross-validation error
- `lambda.1se`: largest value of lambda such that error is within 1 standard error of the minimum (*one standard error rule*). To compute cross-validation "standard errors"

$$\text{se} = \frac{1}{\sqrt{K}} \text{sd}(\text{Err}^{-1}, \ldots, \text{Err}^{-K})$$

where Err^{-k} denotes the error incurred in predicting the observations in the k hold-out fold, $k = 1, \ldots, K$.


CV Error
0.6
0.8
1.0
1.2
1.4
1.6
0
2
4
6
8
k



$\lambda_{\min} = 0.012$ (7 nonzero), $\lambda_{1\text{se}} = 0.21$ (3 nonzero)

Bayesian interpretation

- A Bayesian viewpoint assumes that β has a double-exponential (Laplace) prior distribution with mean zero and scale parameter a function of λ

$$(1/2\tau)\exp(-\|\beta\|_1/\tau)$$

  with $\tau = 1/\lambda$
- It follows that the posterior mode for β is the lasso solution
- However, the lasso solution is not the posterior mean and, in fact, the posterior mean does not yield a sparse coefficient vector

Extensions of the lasso

Group Lasso

- Suppose we have a partition $G_1, \ldots, G_q$ of $\{1, \ldots, p\}$
- The group Lasso penalty (Yuan and Lin, 2006) is given by

$$\lambda \sum_{k=1}^{q} m_k \|\beta_{G_k}\|_2$$

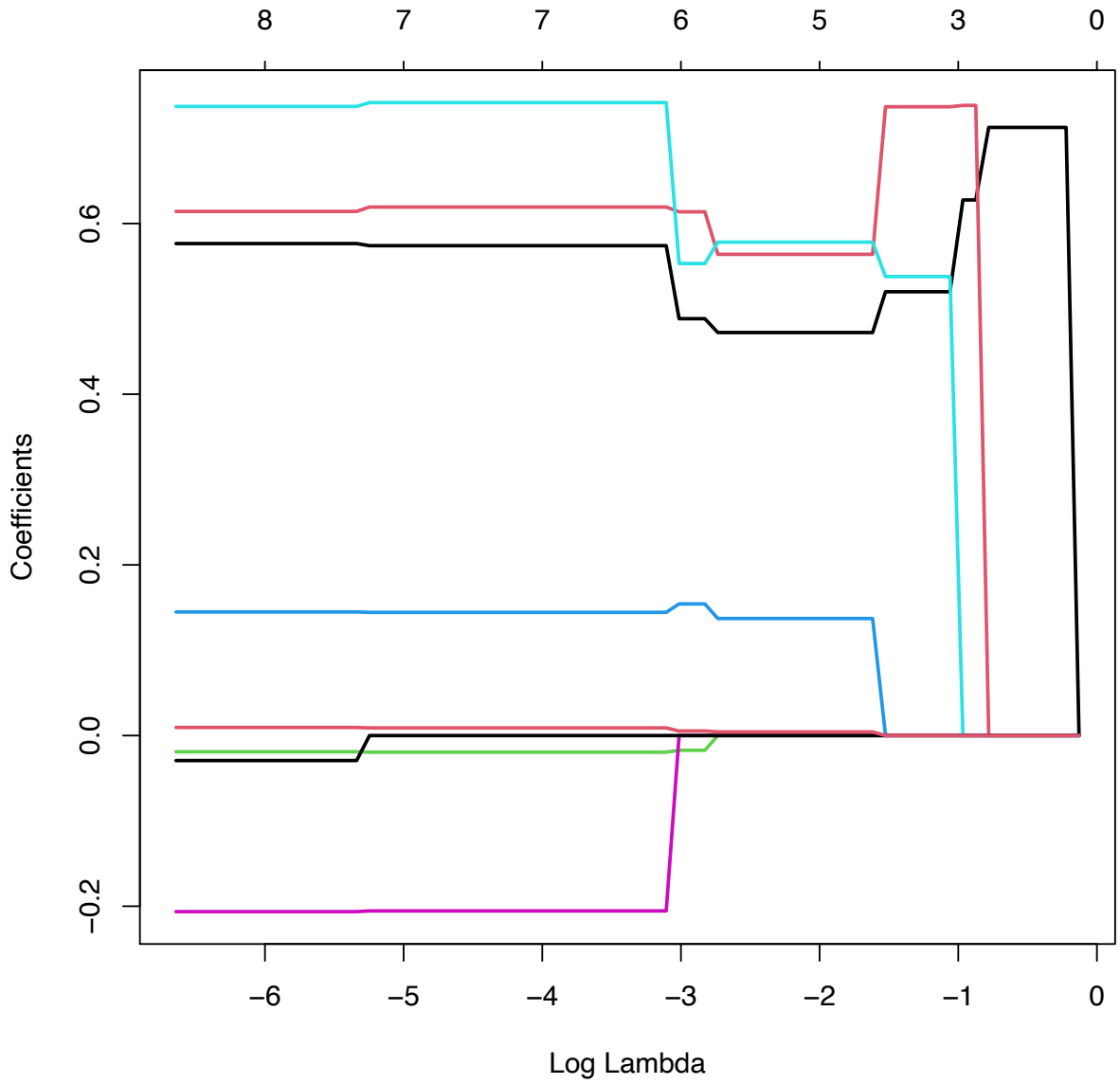
  The multipliers $m_k > 0$ serve to balance cases where the groups are of very different sizes; typically we choose $m_k = \sqrt{|G_k|}$
- This penalty encourages either an entire group G to have $\hat{\beta}_G = 0$ or $\hat{\beta}_j \neq 0$ for all $j \in G$
- Such a property is useful when groups occur through coding for categorical predictors or when expanding predictors using basis functions.

Relaxed Lasso

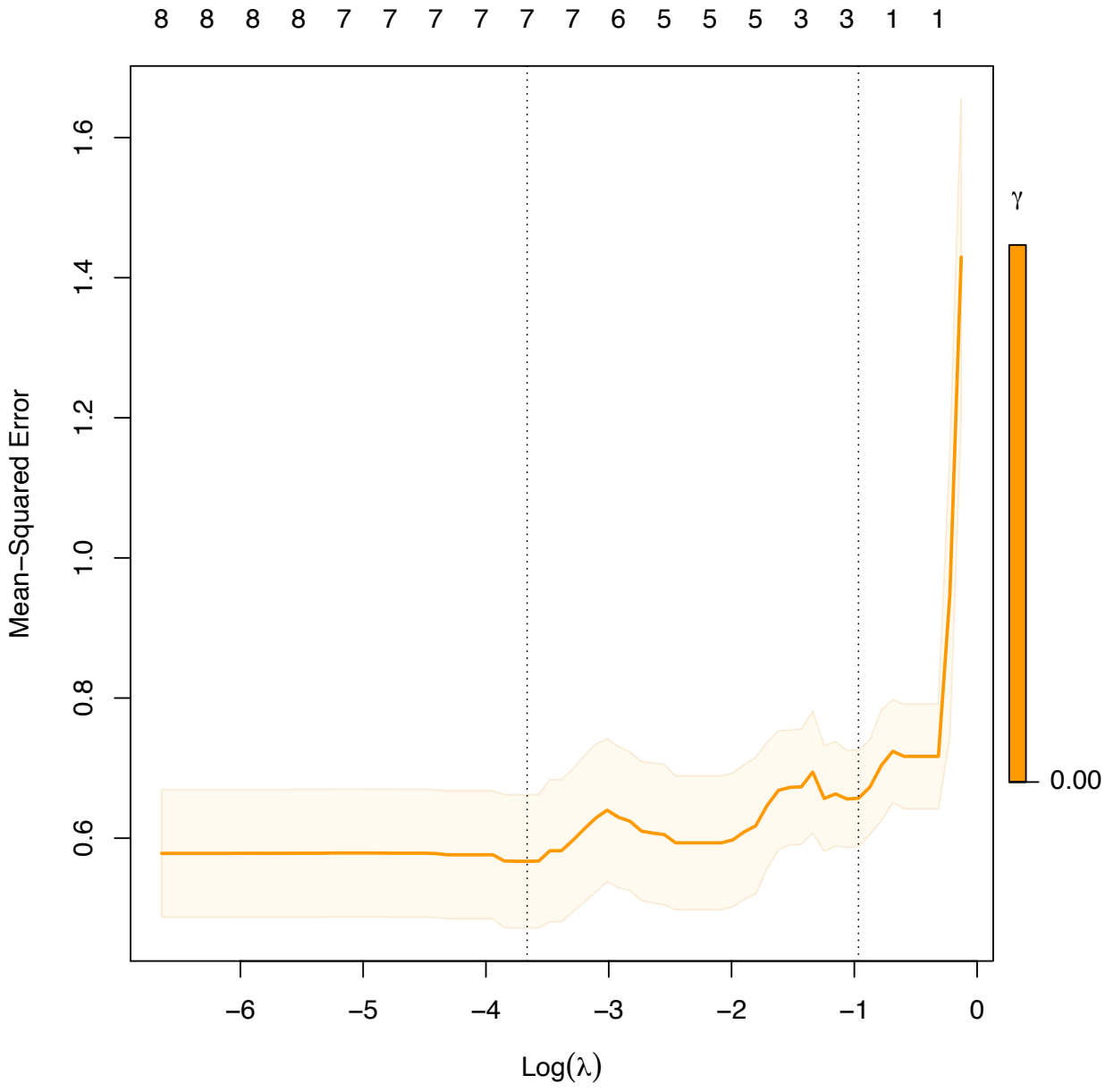
- Originally proposed by Meinshausen (2006). We present a simplified version.
- Suppose $\hat{\beta}_\lambda$ is the lasso solution at λ and let $\hat{A}$ be the active set of indices with nonzero coefficients in $\hat{\beta}_\lambda$
- Let $\hat{\beta}^{\text{LS}}$ be the coeffcients in the least squares fit, using only the variables in $\hat{A}$. Let $\hat{\beta}_\lambda^{\text{LS}}$ be the full-sized version of this coeffcient vector, padded with zeros. $\hat{\beta}_\lambda^{\text{LS}}$ debiases the lasso, while maintaining its sparsity.
- Define the Relaxed Lasso

$$\hat{\beta}_\lambda^{\text{RELAX}} = \gamma\hat{\beta}_\lambda + (1-\gamma)\hat{\beta}_\lambda^{\text{LS}}$$

with $\gamma \in [0, 1]$ is an additional tuning parameter which can be selected by cross-validation



$\gamma = 0$


8 8 8 8 7 7 7 7 7 7 6 5 5 5 3 3 1 1
Mean−Squared Error
0.6
0.8
1.0
1.2
1.4
1.6
−6
−5
−4
−3
−2
−1
0
Log(λ)
γ
0.00

$\gamma = 0$

Elastic Net

- Define the objective function f for some $\lambda > 0$ and $\alpha \in [0, 1]$ as

$$f(\beta; \lambda, \alpha) = \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda \left((1-\alpha)\frac{1}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1 \right)$$

  and the corresponding *elastic net* estimator as

$$\hat{\beta}_{\lambda,\alpha} = \arg\min_{\beta} f(\beta; \lambda, \alpha)$$

- Setting α to 1 yields the Lasso regression and setting it to 0 the ridge regression.
- Adding a small ℓ_2-penalty preserves the variable selection and convexity properties of the ℓ_1-penaltized regression, while reducing the variance of the model when X contains sets of highly correlated variables.

Data splitting for variable selection

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Dezeure, Buhlmann, Meier, Meinshausen (2015). High dimensional inference: Confidence intervals, p-values and r-software `hdi`. Statistical Science, 533–558

High-dimensional inference

- Consider the gaussian linear model

$$y \sim N_n(1_n\beta_0 + X\beta, \sigma^2 I_n)$$

  with $n \times p$ design matrix X and $p \times 1$ vector of coefficients β
- When $p \geq n$, classical approaches for estimation and inference of β cannot be directly applied
- How to perform inference on β (e.g. confidence intervals and p-values for individual regression parameters β_j, $j = 1, \ldots, p$) in a high-dimensional setting?

Support set

- The *support set* is

$$S = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$$

  with cardinality $s = |S|$, and its complement is the *null set*, i.e.

$$N = \{j \in \{1, \ldots, p\} : \beta_j = 0\}$$

- Let $\hat{S} \subseteq \{1, \ldots, p\}$ be an estimator of S. Then

$$|\hat{S} \cap N|$$

  is the number of the wrong selections (type I errors) and

$$|S \setminus \hat{S}|$$

  is the number of wrong deselections (type II errors)

Error rates

- Define the *False Discovery Proportion* (FDP) by

$$\mathrm{FDP}(\hat{S}) = \frac{|\hat{S} \cap N|}{|\hat{S}|}$$

  with $\mathrm{FDP}(\emptyset) = 0$
- *FamilyWise Error Rate* (FWER)

$$\mathrm{P}(\mathrm{FDP}(\hat{S}) > 0) = \mathrm{P}(\hat{S} \cap N \neq \emptyset)$$

- *False Discovery Rate* (FDR)

$$\mathbb{E}(\mathrm{FDP}(\hat{S}))$$

Error control

- We would like to *control* the chosen error rate at level α, i.e.

$$\mathrm{P}(\hat{S} \cap N \neq \emptyset) \leq \alpha \quad \text{or} \quad \mathbb{E}(\mathrm{FDP}(\hat{S})) \leq \alpha$$

  while maximizing some notion of power e.g. the average power

$$\mathrm{AvgPower} = \frac{\sum_{j \in S} \mathrm{P}(\hat{S} \in j)}{|S|}$$

- We are dealing with the trade-off between type I and type II errors, and since FWER is more stringent than FDR, i.e.

$$\mathbb{E}(\mathrm{FDP}(\hat{S})) \leq \mathrm{P}(\hat{S} \cap N \neq \emptyset)$$

  methods that control FWER are less powerful

Simulate data as described in Section 3.1 of Hastie et al. (2020)

Given n (number of observations), p (problem dimensions), s (sparsity level), beta-type (pattern of sparsity), ρ (predictor autocorrelation level), and ν (signal-to-noise ratio (SNR) level)

1. we define coefficients $\beta \in \mathbb{R}^p$ according to s and the beta-type; e.g. beta-type 2: β has its first s components equal to 1, and the rest equal to 0
2. we draw the rows of the predictor matrix $X \in \mathbb{R}^{n \times p}$ i.i.d. from $N_p(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$ has entry (i, j) equal to $\rho^{|i-j|}$ (Toeplitz matrix)
3. we draw the response vector $y \in \mathbb{R}^n$ from $N_n(X\beta, \sigma^2 I_n)$ with σ^2 defined to meet the desired SNR level, i.e. $\sigma^2 = \beta^t \Sigma \beta / \nu$

Lasso active set

Lasso with λ chosen by e.g. the 1-se rule

$$\hat{S} = \{j \in \{1, \ldots, p\} : \hat{\beta}_j \neq 0\}$$

Simulated data with $n = 200$, $p = 1000$, $s = 10$, $\rho = 0$, $\nu = 2.5$:

Size $\lvert\hat{S}\rvert$	# Type I $\lvert\hat{S} \cap N\rvert$	# Type II $\lvert S \setminus \hat{S}\rvert$	FDP $\lvert\hat{S} \cap N\rvert / \lvert\hat{S}\rvert$	Sensitivity $\lvert\hat{S} \cap S\rvert / \lvert S\rvert$
23	13	0	56.5%	100%

100 replications

	1	2	3	4	5	6	7
Size	23	20	13	25	23	21	11
# Type I	13	10	3	15	13	11	4
# Type II	0	0	0	0	0	0	3
FDP	0.57	0.50	0.23	0.60	0.57	0.52	0.36
Sensitivity	1	1	1	1	1	1	0.7

FWER = 99%, FDR = 54.2%, AvgPower = 99.6%

Naïve two-step procedure

1. Perform the lasso in order to obtain the active set

$$\hat{M} = \{j \in \{1, \ldots, p\} : \hat{\beta}_j \neq 0\}$$

2. Use least squares to fit the submodel containing just the variables in $\hat{M}$, i.e. linear regression of the $n \times 1$ response y on the reduced $n \times |\hat{M}|$ submatrix $X_{\hat{M}}$. Obtain

$$\hat{S} = \{j \in \hat{M} : p_j \leq \alpha\}$$

   where p_j is the p-value for testing the null hypothesis $H_j : \beta_j = 0$ in the linear model including only the selected variables

Simulation with $n = 200$, $p = 1000$, $s = 10$, $\rho = 0$, $\nu = 2.5$, $\alpha = 5\%$:

Size $\lvert\hat{S}\rvert$	# Type I $\lvert\hat{S} \cap N\rvert$	# Type II $\lvert S \setminus \hat{S}\rvert$	FDP $\lvert\hat{S} \cap N\rvert / \lvert\hat{S}\rvert$	Sensitivity $\lvert\hat{S} \cap S\rvert / \lvert S\rvert$
15	5	0	33.3%	100%

100 replications

	1	2	3	4	5	6	7
Size	15	18	12	17	18	17	11
# Type I	5	8	2	7	8	7	4
# Type II	0	0	0	0	0	0	3
FDP	0.33	0.44	0.17	0.41	0.44	0.41	0.36
Sensitivity	1	1	1	1	1	1	0.7

FWER = 99%, FDR = 42.1%, AvgPower = 99.6%

j	p_j	Selected
1	0.00	*
2	0.00	*
3	0.00	*
4	0.00	*
5	0.00	*
6	0.00	*
7	0.00	*
8	0.00	*
9	0.00	*
10	0.00	*
37	0.29	
53	0.06	
273	0.00	*
417	0.04	*
427	0.12	
525	0.04	*
577	0.24	
590	0.06	
636	0.16	
673	0.01	*
698	0.31	
721	0.12	
829	0.01	*

- The main problem with the naïve two-step procedure is that it peeks at the data twice: once to select the variables to include in $\hat{M}$, and then again to test hypotheses associated with those variables
- Here $\hat{M}$ is a random variable (it is a function of the data), but inference for linear model assumes it fixed (given a priori)
- A secondary problem is the multiplicity of the tests performed
- A simple idea is to use data-splitting to break up the dependence of variable selection and hypothesis testing (Cox, 1975)

Data-split

The *single-split* approach (Wasserman and Roeder, 2009) splits the data into two parts I and L of equal sizes $n_I = n_L = n/2$:

1. Use variable selection on the L portion (X^L, y^L) to obtain

$$\hat{M}^L \subseteq \{1, \ldots, p\}$$

2. Use the I portion (X^I, y^I) for constructing p-values

$$p_j = \begin{cases} p_j^I & \text{if } j \in \hat{M}^L \\ 1 & \text{if } j \notin \hat{M}^L \end{cases}$$

where p_j^I is the p-value testing $H_j : \beta_j = 0$ in the linear model including only the selected variables, i.e. based on the linear regression of the reduced $n_I \times 1$ response y^I on the reduced $n_I \times |\hat{M}^L|$ matrix $X^I_{\hat{M}^L}$

3. Adjust the p-values for their multiplicity $|\hat{M}^L|$, by e.g. Bonferroni

$$\tilde{p}_j = \min(|\hat{M}^L| \cdot p_j, 1), \quad j = 1, \ldots, p$$

4. Selected variables

$$\tilde{S} = \{j \in \hat{M}^L : \tilde{p}_j \leq \alpha\}$$

j	p_j^L	p_j^I	$\tilde{p}_j^I$	Selected
1	0.00	0.08	1.00	
2	0.00	0.00	0.00	*
3	0.00	0.00	0.00	*
4	0.03	0.01	0.09	
6	0.00	0.00	0.00	*
8	0.00	0.00	0.01	*
9	0.16	0.00	0.00	*
10	0.00	0.00	0.00	*
37	0.03	0.38	1.00	
390	0.15	0.79	1.00	
398	0.01	0.21	1.00	
720	0.24	0.04	0.60	
721	0.02	0.82	1.00	
742	0.04	0.21	1.00	
824	0.02	0.24	1.00	
829	0.01	0.38	1.00	
943	0.15	0.66	1.00	

Theorem

Assume that

1. *the linear model* $y \sim N_n(1\beta_0 + X\beta, \sigma^2 I)$ *holds*
2. *the variable selection procedure satisfies the screening property for the first half of the sample, i.e.*

$$P(\hat{M}^L \supseteq S) \geq 1 - \delta$$

   for some $\delta \in (0, 1)$.

3. *The reduced design matrix for the second half of the sample satisfies* $\text{rank}(X^I_{\hat{M}^L}) = |\hat{M}^L|$.

Then the single-split procedure yields FWER control at α *against inclusion of null predictors up to the additional (small) value* δ*, i.e.*

$$P(\tilde{S} \cap N \neq \emptyset) \leq \alpha + \delta$$

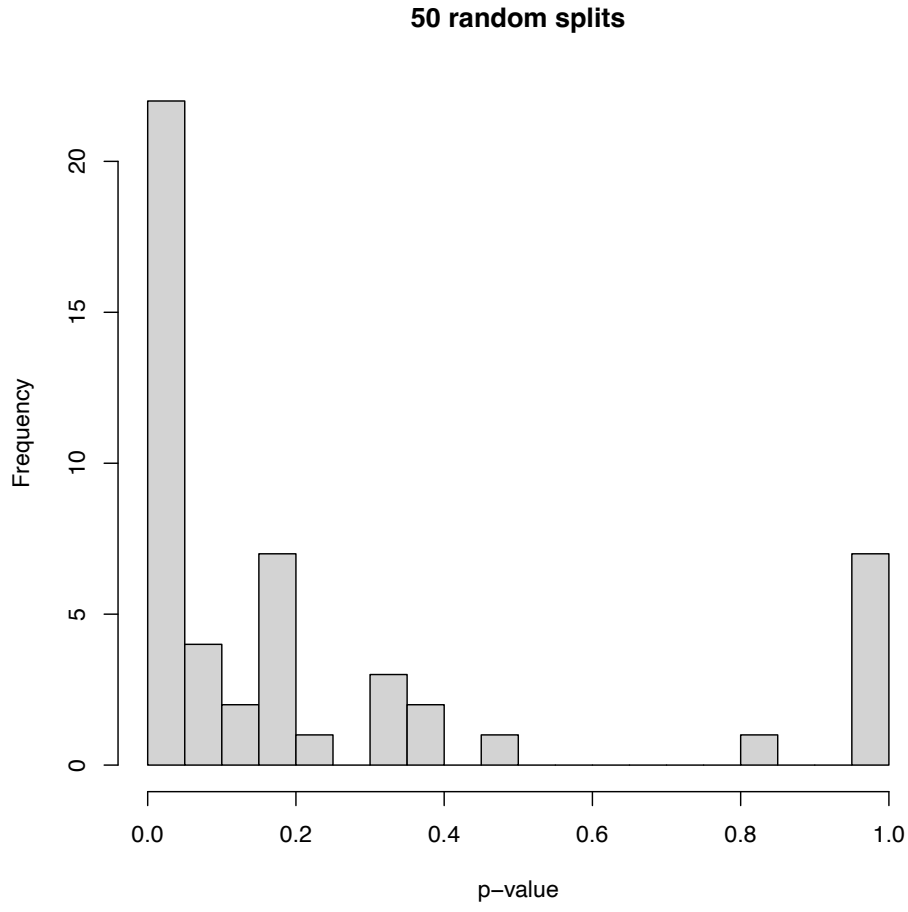
Proof.
Let $E = \{\hat{M}^L \supseteq S\}$ with $\mathrm{P}(E^c) \leq \delta$ by assumption. If E happens, then p_j^I is a valid p-value, i.e. $\mathrm{P}(p_j^I \leq u|E) \leq u$ for $j \in N \cap \hat{M}^L$. We have

$$
\begin{aligned}
& \mathrm{P}(\tilde{S} \cap N \neq \emptyset) = \mathrm{P}\Big(\bigcup_{j \in \hat{M}^L \cap N} \{\tilde{p}_j \leq \alpha\} \Big) \\
= \; & \mathrm{P}\Big(\bigcup_{j \in \hat{M}^L \cap N} \{\tilde{p}_j \leq \alpha\} | E \Big) \mathrm{P}(E) + \mathrm{P}\Big(\bigcup_{j \in \hat{M}^L \cap N} \{\tilde{p}_j \leq \alpha\} | E^c \Big) \mathrm{P}(E^c) \\
\leq \; & \Big[\sum_{j \in \hat{M}^L \cap N} \mathrm{P}\Big(p_j^I \leq \frac{\alpha}{|\hat{M}^L|} | E\Big) \Big] \mathrm{P}(E) + \mathrm{P}\Big(\bigcup_{j \in \hat{M}^L \cap N} \mathbb{1}\{\tilde{p}_j \leq \alpha\} | E^c \Big) \mathrm{P}(E^c) \\
\leq \; & |\hat{M}^L \cap N| \frac{\alpha}{|\hat{M}^L|} \cdot 1 + 1 \cdot \delta \\
\leq \; & \alpha + \delta
\end{aligned}
$$

□

P-value lottery

A major problem of the single data-splitting method is that different data splits lead to different p-values



Multi-split

The *multi-split* approach (Meinshausen et al., 2009)

1. For $b = 1, \ldots, B$
   apply the single-split procedure (L^b, I^b) to obtain

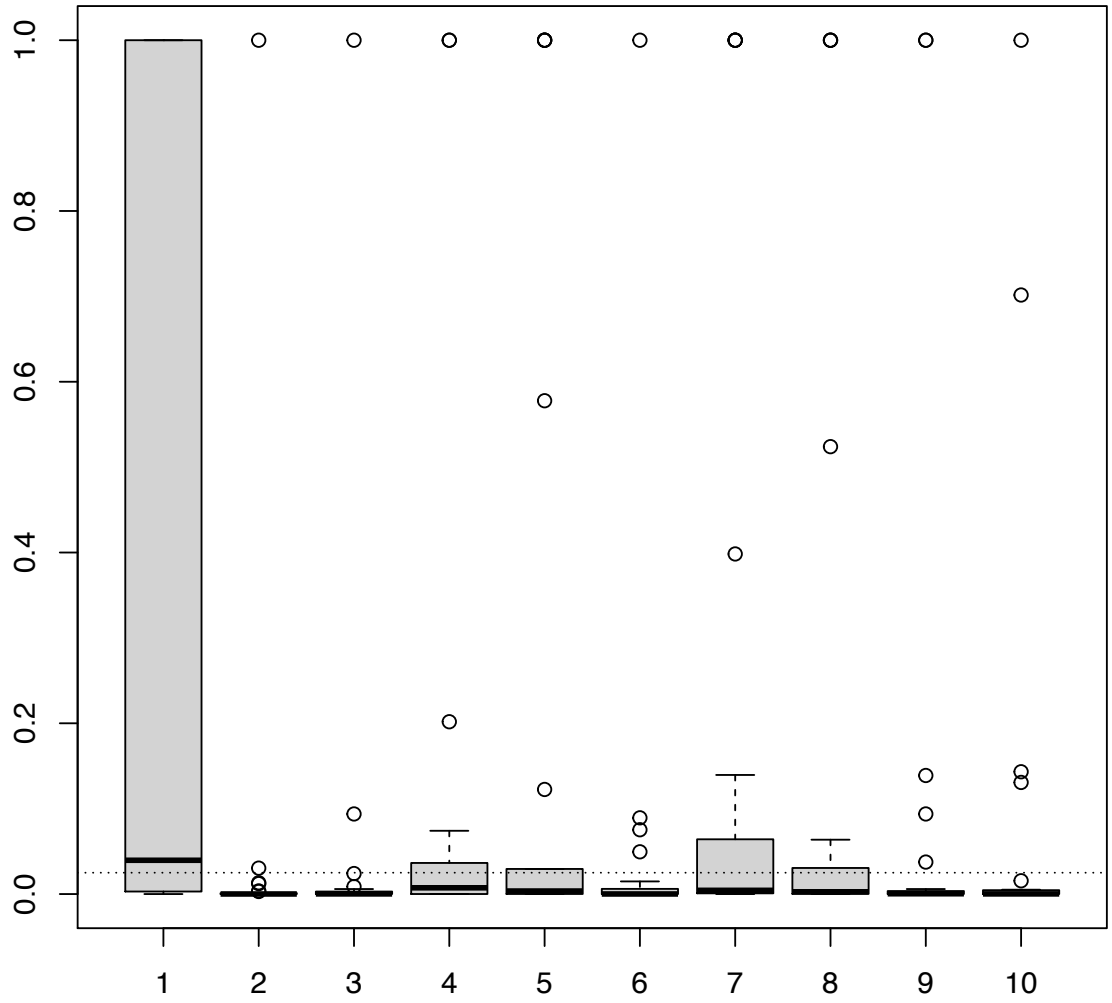
$$\{\tilde{p}_j^b, j = 1, \ldots, p\}$$

2. Aggregate the p-values as

$$\bar{p}_j = 2 \cdot \text{median}(\tilde{p}_j^1, \ldots, \tilde{p}_j^B), \quad j = 1, \ldots, p$$

3. Selected predictors:

$$\bar{S} = \{j \in \{1, \ldots, p\} : \bar{p}_j \leq \alpha\}$$


1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10

Simultaneous confidence intervals

$$\mathrm{P}(\beta_j \in [\hat{L}_j, \hat{U}_j] \ \forall j \in \{1, \ldots, p\}) \geq 1 - \alpha$$

j	$\hat{L}_j$	$\hat{U}_j$
1	-∞	∞
2	0.69	1.84
3	0.48	1.73
4	0.36	1.49
5	0.47	1.70
6	0.56	1.78
7	0.27	1.57
8	0.40	1.69
9	0.41	1.56
10	0.44	1.56
11	-∞	∞
...		

Stability Selection

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Meinshausen, Buhlmann (2010). Stability selection. JRSS-B, 72:417–473
- Shah, Samworth (2013). Variable selection with error control: another look at stability selection. JRSS-B, 75:55–80.

Stability path

- The *regularisation path* of the lasso is

$$\{\hat{\beta}_j(\lambda), j = 1, \ldots, p, \lambda \in \Lambda\}$$

- The *stability path* is

$$\{\hat{\pi}_j(\lambda), j = 1, \ldots, p, \lambda \in \Lambda\}$$

  where $\hat{\pi}_j(\lambda)$ is the estimated probability for the jth predictor to be selected by the lasso(λ) when randomly resampling from the data

Algorithm 1 Stability Path Algorithm with the Lasso

Require: $B \in \mathbb{N}$, Λ grid, $\tau \in (0.5, 1)$

1: **for** $b = 1, \ldots, B$ **do**

2: Randomly select $n/2$ indices from $\{1, \ldots, n\}$;

3: Perform the lasso on the $n/2$ observations to obtain

$$\hat{S}_{n/2}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0\} \qquad \forall\ \lambda \in \Lambda$$

4: **end for**

5: Compute the relative selection frequencies:

$$\hat{\pi}_j(\lambda) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{j \in \hat{S}_{n/2}(\lambda)\} \qquad \forall\ \lambda \in \Lambda$$

6: The set of *stable predictors* is given by

$$\hat{S}_{\text{stab}} = \{j : \max_{\lambda \in \Lambda} \hat{\pi}_j(\lambda) \geq \tau\}$$

Algorithm 2 (Complementary Pairs) Stability Selection

Require: A variable selection procedure $\hat{S}_n$, $B \in \mathbb{N}$, $\tau \in (0.5, 1)$

1: **for** $b = 1, \ldots, B$ **do**

2: Split $\{1, \ldots, n\}$ into (I^{2b-1}, I^{2b}) of size $n/2$, and for each get

$$\hat{S}^{2b-1}_{n/2} \subseteq \{1, \ldots, p\}, \qquad \hat{S}^{2b}_{n/2} \subseteq \{1, \ldots, p\}$$

3: **end for**

4: Compute the relative selection frequencies:

$$\hat{\pi}_j = \frac{1}{2B} \sum_{b=1}^{B} \left(\mathbb{1}\{j \in \hat{S}^{2b-1}_{n/2}\} + \mathbb{1}\{j \in \hat{S}^{2b}_{n/2}\} \right)$$

5: The set of *stable predictors* is given by

$$\hat{S}_{\text{stab}} = \{j : \hat{\pi}_j \geq \tau\}$$

- The relative selection frequency $\hat{\pi}_j$ is an unbiased estimator of

$$\pi_j^{n/2} = \mathrm{P}(j \in \hat{S}_{n/2})$$

  but, in general, a biased estimator of

$$\pi_j^n = \mathrm{P}(j \in \hat{S}_n) = \mathbb{E}(\mathbb{1}\{j \in \hat{S}_n\})$$

- The key idea of stability selection is to improve on the simple estimator $\mathbb{1}\{j \in \hat{S}_n\}$ of π_j^n through subsampling.
- By means of averaging involved in $\hat{S}_{\text{stab}}$, we hope that $\hat{\pi}_j$ will have reduced variance compared to $\mathbb{1}\{j \in \hat{S}_n\}$ and this increased stability will more than compensate for the bias incurred.

Theorem

Assume that

1. $\{\mathbb{1}\{j \in \hat{S}_{n/2}\}, j \in N\}$ *is exchangeable;*
2. *The variable selection procedure is not worse than random guessing, i.e.*

$$\frac{\mathbb{E}(|\hat{S}_{n/2} \cap S|)}{\mathbb{E}(|\hat{S}_{n/2} \cap N|)} \geq \frac{|S|}{|N|}.$$

Then, for $\tau \in (1/2, 1]$

$$\mathbb{E}(|\hat{S}_{\text{stab}} \cap N|) \leq \frac{1}{(2\tau - 1)} \frac{q^2}{p}$$

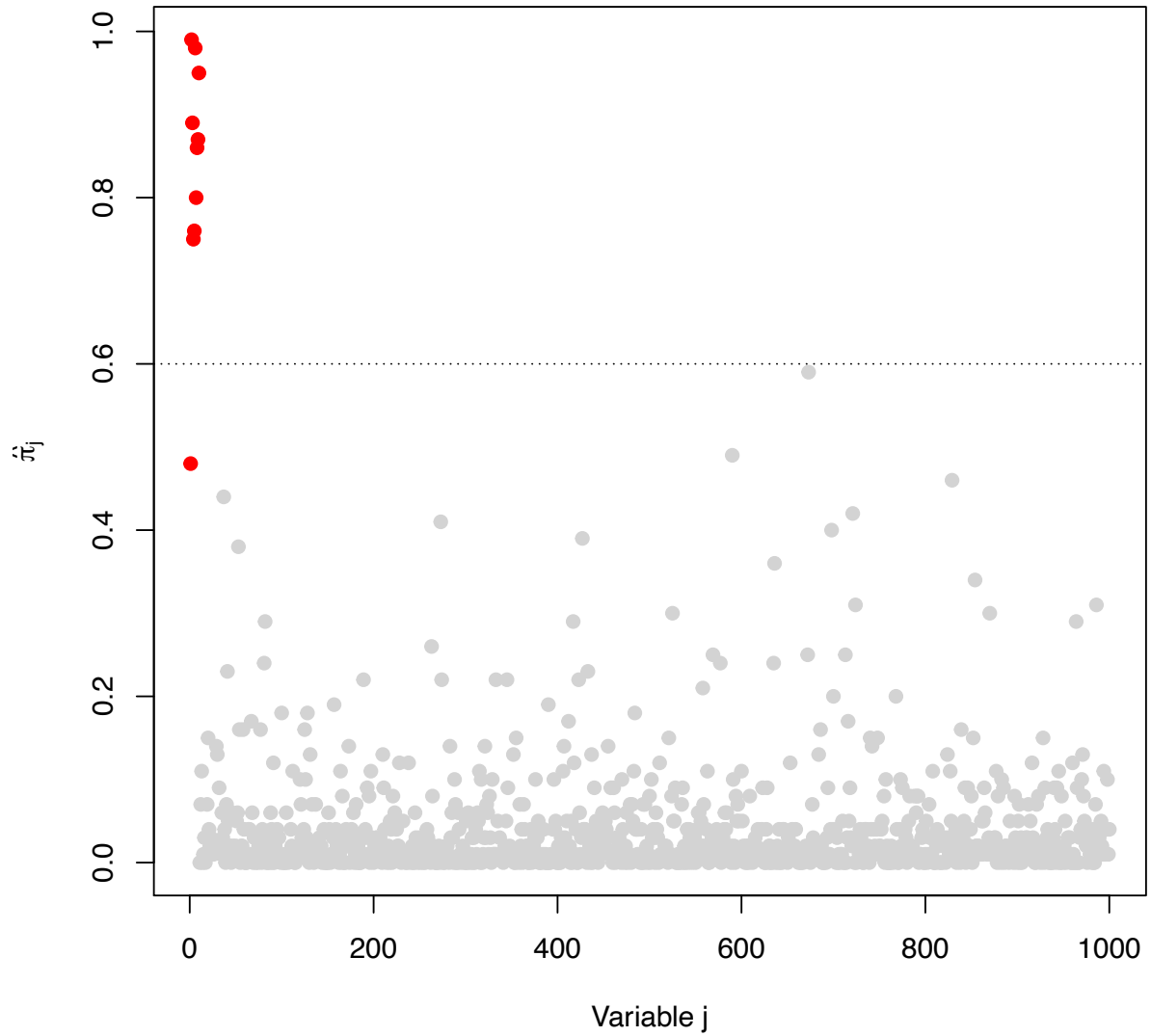
where $q = \mathbb{E}(|\hat{S}_{n/2}|)$

- The choice of the number of subsamples B is of minor importance
- It is possible to fix $q = \mathbb{E}(|\hat{S}_{n/2}|)$ and run the variable selection procedure until it selects q variables. However, if q is too small, one would select only a subset of the signal variables as

$$|\hat{S}_{\text{stab}}| \leq |\hat{S}_{n/2}| = q$$

- For example, with $p = 1000$, $q = 50$ and $\tau = 0.6$ then

$$\mathbb{E}(|\hat{S}_{\text{stab}} \cap N|) \leq 12.5$$


1.0
0.8
0.6
0.4
0.2
0.0
$\hat{\pi}_j$
0
200
400
600
800
1000
Variable j

Conformal prediction

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (2018) Distribution-free predictive inference for regression. JASA,113:1094–1111

Suppose we have fitted a Gaussian linear model based on the training data $(\mathbf{y}, \mathbf{X})$, obtaining the estimates

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}, \quad \hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2/(n-p)$$

There are (at least) two levels at which we can make predictions

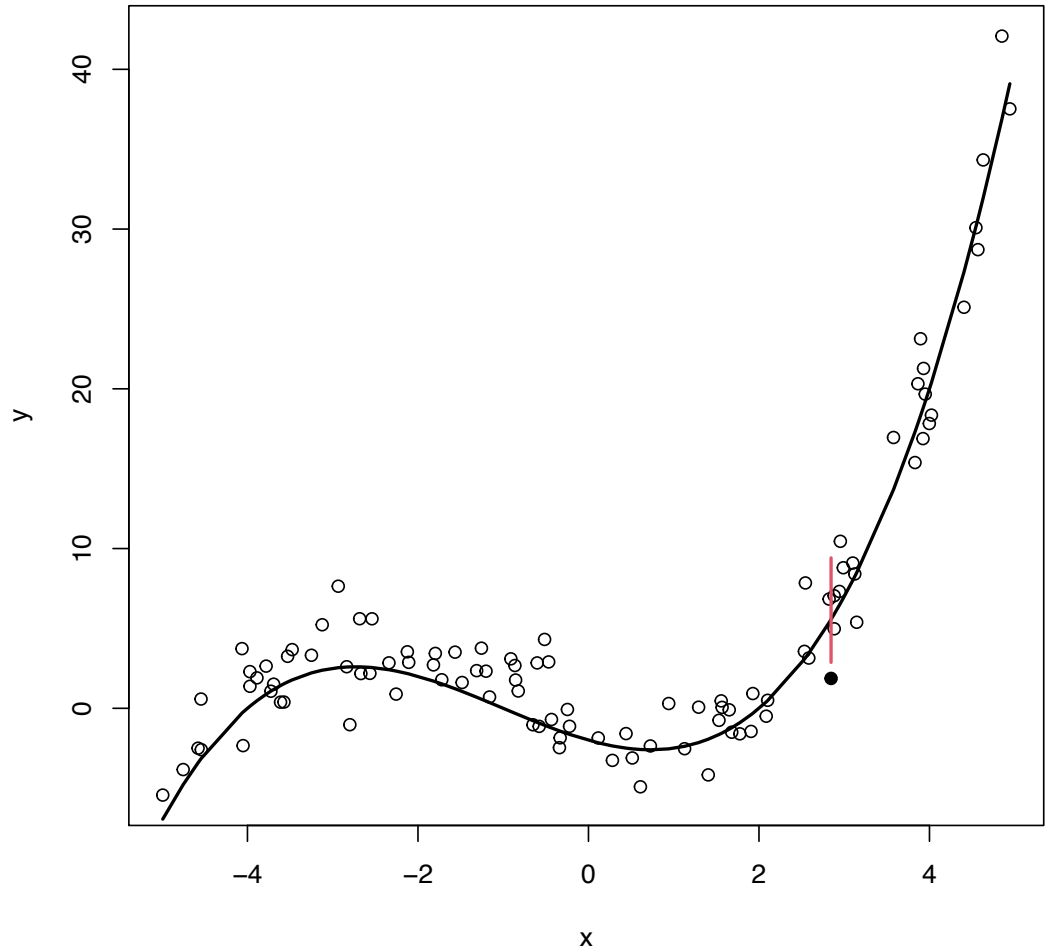
1. A *point prediction* is a single best guess about what a new Y will be when $X = x$
2. A *prediction interval*

   $$C_\alpha(x) = x^t\hat{\beta} \pm t_{n-p}^{1-\alpha/2}\hat{\sigma}\sqrt{x^t(\mathbf{X}^t\mathbf{X})^{-1}x + 1}$$

   for $Y|X = x$ with $(1-\alpha)$ *conditional coverage* guarantee, i.e.

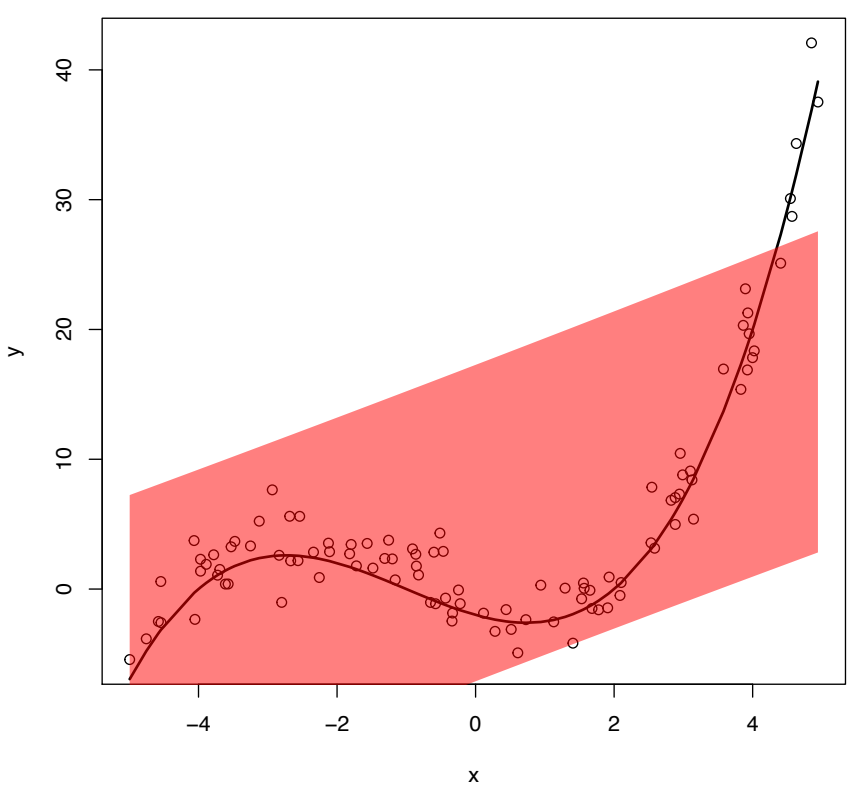
   $$\mathrm{P}(Y \in C_\alpha(x)|X = x) = 1 - \alpha$$

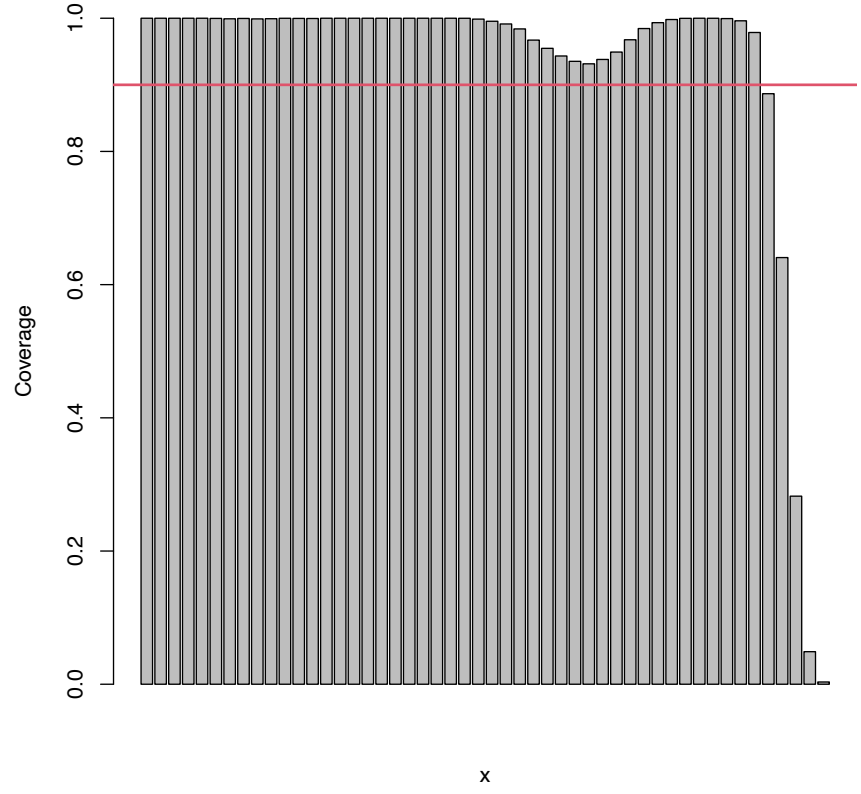
   where the probability is with respect to the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, and the new response Y at a fixed test point $X = x$



$$f(x) = \frac{1}{4}(x+4)(x+1)(x-2)$$

Model miss-specification





$1 - \alpha = 90\%$, marginal coverage $\approx 93\%$

Marginal and conditional coverage

- $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ follows some *unknown* joint distribution P_{XY}
- Training $(X_1, Y_1), \ldots, (X_n, Y_n)$ and test (X_{n+1}, Y_{n+1}) i.i.d. (X, Y)
- C_α satisfies *distribution-free marginal coverage* at level $1 - \alpha$ if

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha \qquad \forall\ P_{XY}$$

  where the probability is w.r.t. $(X_1, Y_1), \ldots, (X_n, Y_n)$ and (X_{n+1}, Y_{n+1})
- C_α satisfies *distribution-free conditional coverage* at level $1 - \alpha$ if

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1}) | X_{n+1} = x) \geq 1 - \alpha \qquad \forall\ P_{XY},\ \ \forall\ x$$

  where the probability is w.r.t. $(X_1, Y_1), \ldots, (X_n, Y_n)$, and Y_{n+1} at a fixed test point $X_{n+1} = x$

Conformal prediction

Conformal prediction (Vovk, Gammerman, Saunders, Vapnik, 1996-1999) is a general framework for constructing prediction intervals by using *any* algorithm with finite sample and distribution-free *exact* marginal coverage, i.e.

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) = 1 - \alpha \qquad \forall\ P_{XY}$$

Two main versions:

- *Full* conformal prediction
- *Split* conformal prediction

Algorithm 1 Full conformal prediction

Require: Training $(x_1, y_1), \ldots, (x_n, y_n)$, test x_{n+1}, algorithm $\hat{\mu}$, level α, grid of values $\mathcal{Y} = \{y, y', y'', \ldots\}$

for $y \in \mathcal{Y}$ **do**
    Train $\hat{\mu}^y(x) = \hat{\mu}(x; (x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y))$
    Compute $R_i^y = |y_i - \hat{\mu}^y(x_i)|$ for $i = 1, \ldots, n$
    Sort $R_1^y, \ldots, R_n^y$ in increasing order: $R_{(1)}^y \leq \ldots \leq R_{(n)}^y$
    Compute $R_\alpha^y = R_{(k)}^y$ with $k = \lceil (1-\alpha)(n+1) \rceil$
    Compute $R^y = |y - \hat{\mu}^y(x_{n+1})|$
end for
$C_\alpha(x_{n+1}) = \{y \in \mathcal{Y} : R^y \leq R_\alpha^y\}$

- Assume that (X_i, Y_i), $i = 1, \ldots, n+1$ are i.i.d. from a probability distribution P_{XY} on the sample space $\mathbb{R}^p \times \mathbb{R}$. This is the only assumption of the method
- The prediction interval

$$C_\alpha(x_{n+1}) = \{y \in \mathbb{R} : R^y \leq R^y_\alpha\},$$

  satisfies

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) = 1 - \alpha$$

  if and only if $\alpha \in \{1/(n+1), 2/(n+1), \ldots, n/(n+1)\}$
- Informally, the null hypothesis that the random variable Y_{n+1} will have the outcome y, i.e.

$$H_y : Y_{n+1} = y$$

  is rejected when $R^y > R^y_\alpha$

Algorithm 2 Split conformal prediction

Require: Training $(x_1, y_1), \ldots, (x_n, y_n)$, x_{n+1}, algorithm $\hat{\mu}$, validation sample size m, level α

1: Split $\{1, \ldots, n\}$ into L of size w and I of size $m = n - w$
2: Train $\hat{\mu}_L(x) = \hat{\mu}(x; (x_l, y_l), l \in L)$
3: Compute $R_i = |y_i - \hat{\mu}_L(x_i)|$ for $i \in I$
4: Sort $\{R_i, i \in I\}$ in increasing order: $R_{(1)} \leq \ldots \leq R_{(m)}$
5: Compute $R_\alpha = R_{(k)}$ with $k = \lceil (1-\alpha)(m+1) \rceil$

$$\begin{aligned} C_\alpha(x_{n+1}) &= \{y \in \mathbb{R} : |y - \hat{\mu}_L(x_{n+1})| \leq R_\alpha\} \\ &= [\hat{\mu}_L(x_{n+1}) - R_\alpha, \hat{\mu}_L(x_{n+1}) + R_\alpha] \end{aligned}$$

- Assume that (X_i, Y_i), $i = 1, \ldots, n+1$ are i.i.d. from a probability distribution P_{XY} on the sample space $\mathbb{R}^p \times \mathbb{R}$
- The prediction interval

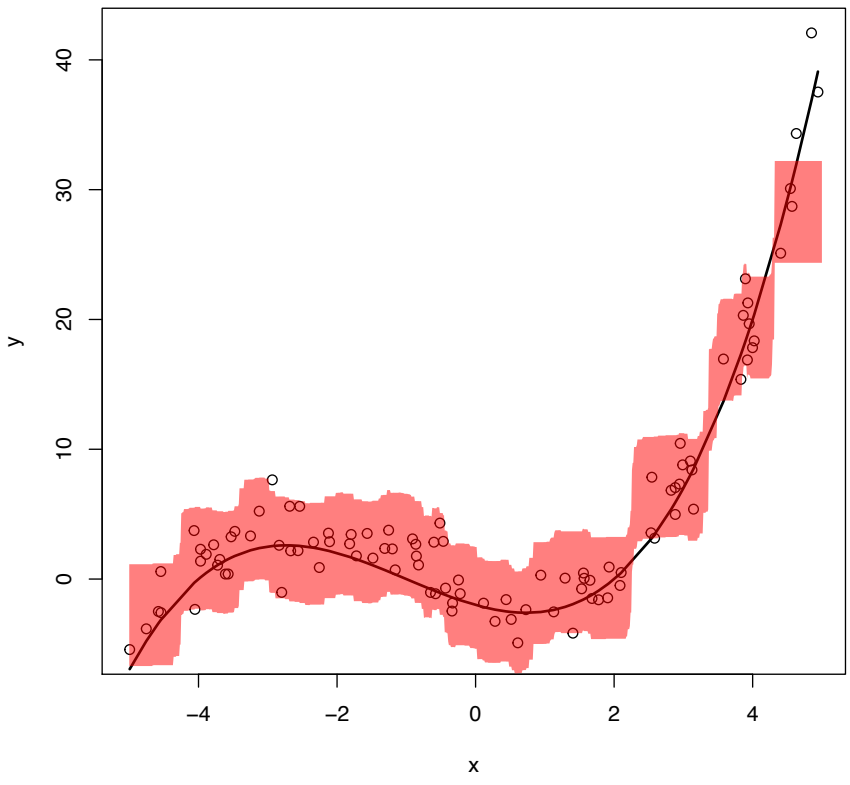
$$C_\alpha(x_{n+1}) = [\hat{\mu}_L(x_{n+1}) - R_\alpha, \hat{\mu}_L(x_{n+1}) + R_\alpha]$$

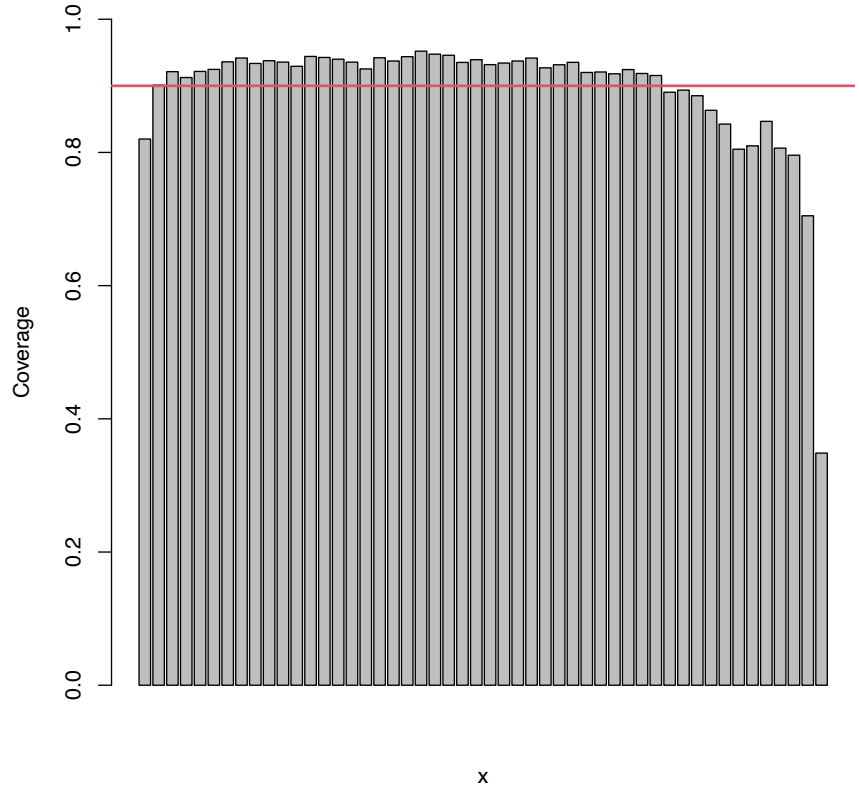
  satisfies

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) = 1 - \alpha$$

  if and only if $\alpha \in \{1/(m+1), 2/(m+1), \ldots, m/(m+1)\}$
- Note that in computing the critical value $R_\alpha = R_{(k)}$ with $k = \lceil (1-\alpha)(m+1) \rceil$, we need to have $k \leq m$, which happens if $\alpha \geq 1/(m+1)$ (otherwise if $k > m$ we need to set $R_\alpha = +\infty$)

Random Forest





Conformity scores

- In the previous algorithm we used a statistic, called *conformity score*, which is the absolute value of the residual

$$R_i = |y_i - \hat{\mu}_L(x_i)|, \quad i \in I$$

  where $\hat{\mu}_L$ is an estimator of $\mathbb{E}(Y \mid X)$ based on $\{(X_i, Y_i), i \in L\}$
- The oracle knows the conditional distribution of $Y \mid X$. The oracle prediction interval

$$C^*_\alpha(x) = [q^{\alpha/2}(x), q^{1-\alpha/2}(x)]$$

  where $q^\gamma(x)$ is the γ-quantile of $Y \mid X = x$, guarantees exact conditional coverage

$$\mathrm{P}(Y \in C^*_\alpha(X) | X = x) = 1 - \alpha \quad \forall\ x$$

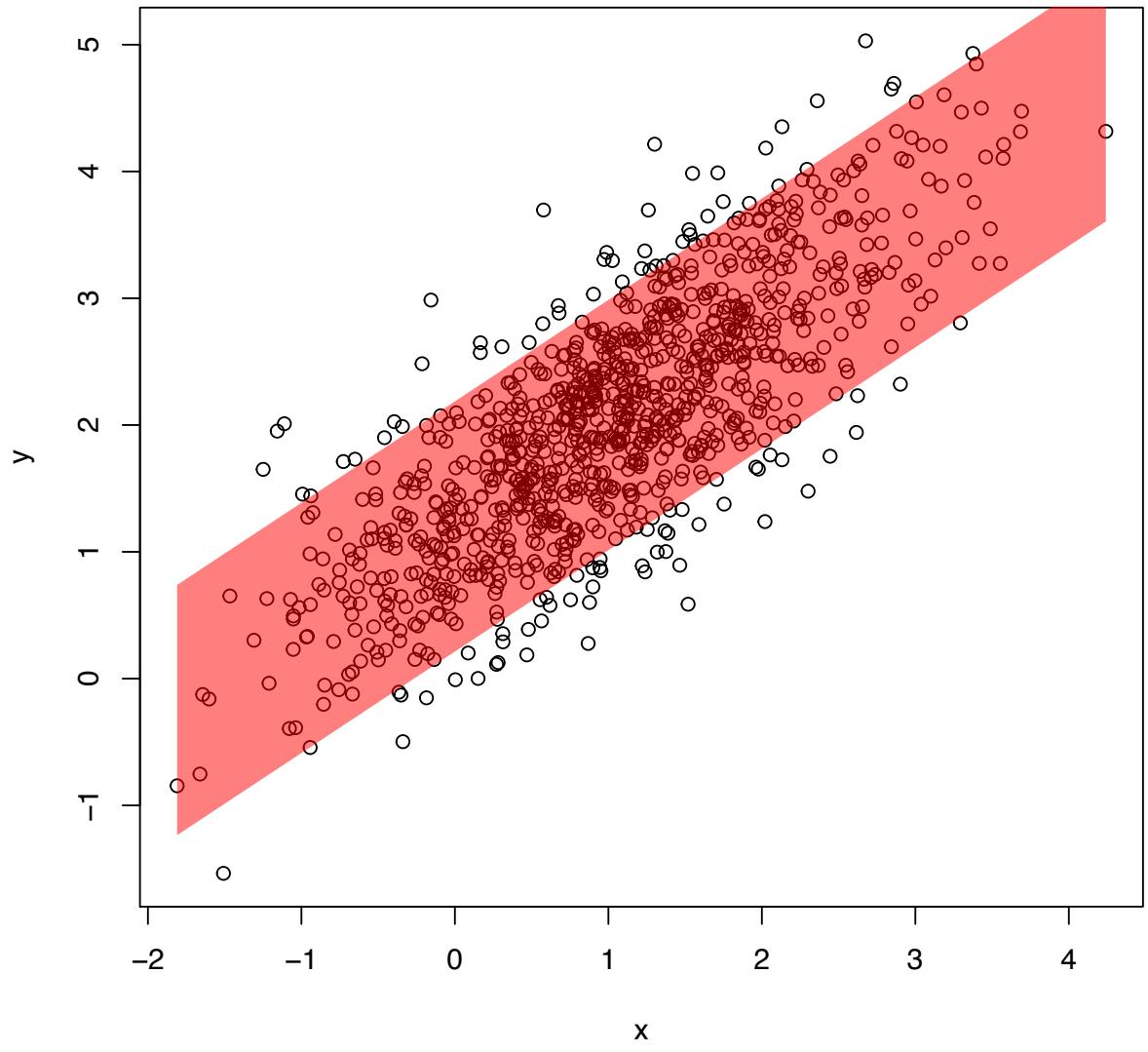
Suppose that

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}\right)$$

then the conditional distribution of $Y \mid X = x$ is

$$(Y|X = x) \sim N\Big(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y^2(1 - \rho^2)\Big)$$

from which we can compute the quantile $q^\gamma(x)$



$C^*_\alpha(x) = [q^{\alpha/2}(x), q^{1-\alpha/2}(x)]$ as a function of x

Conformal quantile regression

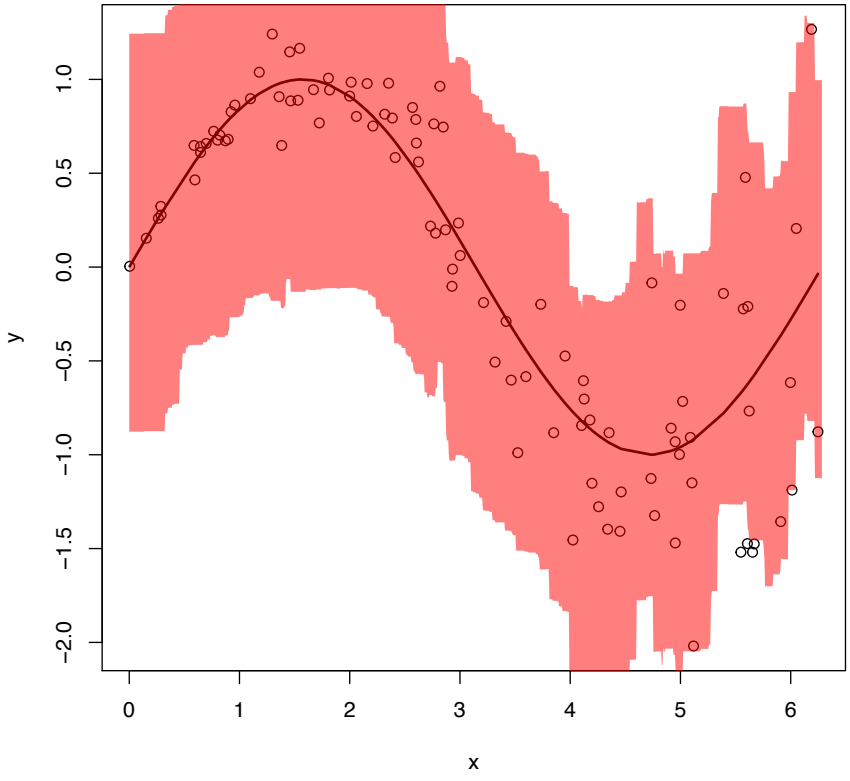
- Compute conformity scores

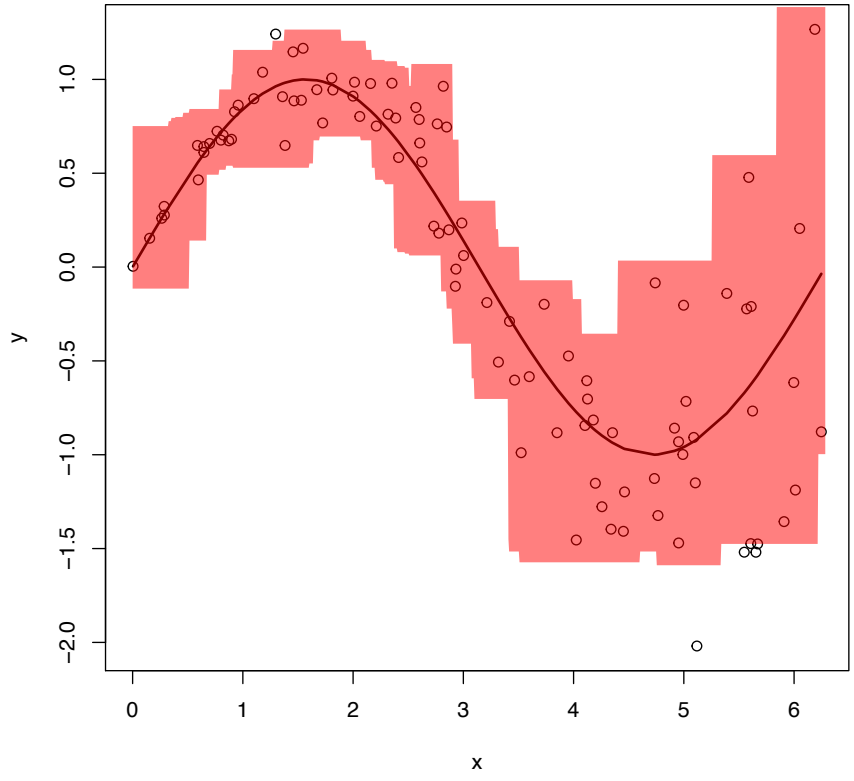
$$R_i = \max\left\{\hat{q}_L^{\gamma}(X_i) - Y_i, Y_i - \hat{q}_L^{1-\gamma}(X_i)\right\}, \quad i \in I$$

  where $\hat{q}_L^{\gamma}$ is an estimator of the γ-quantile of $Y \mid X$ based on $\{(X_i, Y_i), i \in L\}$
- Sort $\{R_i, i \in I\}$ in increasing order, obtaining $R_{(1)} \leq \ldots \leq R_{(m)}$, and compute $R_\alpha = R_{(k)}$ with $k = \lceil (1-\alpha)(m+1) \rceil$
- Compute the prediction interval

$$\begin{aligned} C_\alpha(x_{n+1}) &= \{y \in \mathbb{R} : \max\left\{\hat{q}_L^{\gamma}(x_{n+1}) - y, y - \hat{q}_L^{1-\gamma}(x_{n+i})\right\} \leq R_\alpha\} \\ &= [\hat{q}_L^{\gamma}(x_{n+1}) - R_\alpha, \hat{q}_L^{1-\gamma}(x_{n+1}) + R_\alpha] \end{aligned}$$

  or $C_\alpha(x_{n+1}) = \emptyset$ if $R_\alpha < (1/2)(\hat{q}_L^{\gamma}(x_{n+1}) - \hat{q}_L^{1-\gamma}(x_{n+1}))$





$$X_i \sim U(0, 2\pi), \epsilon_i \sim N(0, 1), Y_i = \sin(X_i) + \frac{\pi |X_i|}{20} \epsilon_i$$

Multi-split conformal prediction

Algorithm

1. Choose a number B of splits
2. Choose a threshold $\tau \in \{0, 1/B, 2/B, \ldots, (B-1)/B\}$
3. Compute B split conformal prediction intervals with coverage level $1 - \beta$

   $$C_\beta^{[1]}(x_{n+1}), \ldots, C_\beta^{[B]}(x_{n+1})$$

   where

   $$\beta = \alpha(1 - \tau)$$

4. Compute the aggregated prediction interval

   $$C_\alpha^\tau(x_{n+1}) = \{y \in \mathbb{R} : \Pi_\beta^y > \tau\}$$

   with

   $$\Pi_\beta^y = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{y \in C_\beta^{[b]}(x_{n+1})\}$$

- The multi-split prediction interval guarantees

$$\mathrm{P}(Y_{n+1} \in C_\alpha^\tau(X_{n+1})) \geq 1 - \alpha \qquad \forall\ P_{XY}$$

- The parameter τ can be regarded as a tuning parameter, and proper choice of τ is essential for good performance
- On the one hand, setting $\tau = 1 - 1/B$ gives the Bonferroni-intersection method with $C_\alpha^{(B-1)/B} = \bigcap_b C_{\alpha/B}^{[b]}$.
- On the other hand, setting $\tau = 0$ gives an unadjusted-union $C_\alpha^0 = \bigcup_b C_\alpha^{[b]}$.
- For B even, an intermediate choice $\tau = 1/2$ amounts to constructing B single split confidence intervals at level $\alpha/2$, that is $C_{\alpha/2}^{[b]}$, which is a small but not negligible price to pay for using multiple splits rather than just one split. In practice, however, $\tau = 1/2$ and $C_\alpha^{[b]}$ may give marginal coverage $\approx 1 - \alpha$