Very high-dimensional data: greedy boosting

(and convex Lasso-optimization)

Peter Bühlmann

ETH Zürich

1. High-dimensional data

$(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. $\underbrace{\text{or stationary}}_{\text{e.g. times series}}$

$X_i \in \mathbb{R}^p$ predictor variable

Y_i univariate response variable, e.g. $Y_i \in \mathbb{R}$ or $Y_i \in \{0, 1\}$

high-dimensional: $p \gg n$

areas of application: astronomy, biology, imaging, marketing research, text classification,...

High-dimensional linear models

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i, \; i = 1, \ldots, n$$

$p \gg n$

How should we fit this model?

approaches include:

Ridge regression (Tikhonov regularization); variable selection via AIC, BIC, gMDL (in a forward manner); Bayesian methods for regularization, ...

Boosting

2. Greedy is "quite good" for $p \gg n$: L_2Boosting

boosting has been advocated as an ensemble method

(multiple prediction and aggregation)

specify a base procedure ("weak learner"):

$$\text{data} \quad \xrightarrow{\text{base procedure}} \quad \hat{\theta}(\cdot) \qquad \text{(a function estimate)}$$

e.g. tree (CART)

principle:

use many base procedure estimates from "reweighted data" to improve prediction

2.1. L_2Boosting

with base procedure $\hat{\theta}(\cdot)$

$\rightsquigarrow$ amounts to repeated fitting of residuals

$$m = 1: \ (X_i, Y_i)_{i=1}^n \ \rightsquigarrow \hat{\theta}_1(\cdot), \ \hat{f}_1 = \underbrace{\nu}_{\text{e.g. } = 0.1} \hat{\theta}_1 \quad \rightsquigarrow \text{ resid. } U_i = Y_i - \hat{f}_1(X_i)$$

$$m = 2: \ (X_i, U_i)_{i=1}^n \ \rightsquigarrow \hat{\theta}_2(\cdot), \ \hat{f}_2 = \hat{f}_1 + \nu\hat{\theta}_2 \quad \rightsquigarrow \text{ resid. } U_i = Y_i - \hat{f}_2(X_i)$$

$$\ldots \qquad\qquad \ldots$$

$\hat{f}_{m_{stop}}(\cdot) = \nu \sum_{m=1}^{m_{stop}} \hat{\theta}_m(\cdot)$ (greedy fitting of residuals)

Tukey (1977): twicing for $m_{stop} = 2$ and $\nu = 1$

2.1. L_2Boosting for linear models

base procedure: componentwise linear least squares

linear OLS regression against the one predictor variable which reduces residual sum of squares most

$$\hat{\theta}(x) = \hat{\beta}_{\hat{\mathcal{S}}} x^{(\hat{\mathcal{S}})}, \ \hat{\beta}_j = \sum_{i=1}^{n} Y_i X_i^{(j)} / \sum_{i=1}^{n} (X_i^{(j)})^2, \ \hat{\mathcal{S}} = \arg\min_j \sum_{i=1}^{n} (Y_i - \hat{\beta}_j X_i^{(j)})^2$$

first round of estimation: selected predictor variable $X^{(\hat{\mathcal{S}}_1)}$ (e.g. $= X^{(3)}$)

corresponding $\hat{\beta}_{\hat{\mathcal{S}}_1} \rightsquigarrow$ fitted function $\hat{f}_1(x)$

second round of estimation: selected predictor variable $X^{(\hat{\mathcal{S}}_2)}$ (e.g.$= X^{(21)}$)

corresponding $\hat{\beta}_{\hat{\mathcal{S}}_2} \rightsquigarrow$ fitted function $\hat{f}_2(x)$

etc.

yields linear model fit, i.e. structured model fit for $\nu = 1$, this is known as

Matching Pursuit (Mallat and Zhang, 1993)

Weak greedy algorithm (deVore & Temlyakov, 1997)

a version of Boosting (Schapire, 1992; Freund & Schapire, 1996)

Gauss-Southwell algorithm

C.F. Gauss in 1803

“Princeps Mathematicorum”

R.V. Southwell in 1933

Professor in engineering, Oxford

Properties

variable selection

shrinkage towards zero for coefficients of selected variables

$\rightsquigarrow$ often much better performance than OLS on selected variables
("more stable" in Breiman's terminology)

computational complexity:

$$O(npm_{stop}) = O(p) \text{ if } p \gg n, \text{ i.e. linear in dimension } p$$

statistically consistent for very high-dimensional, sparse problems

Theorem (PB, 2004)

L_2Boosting with comp. linear LS regression is consistent (for suitable number of boosting iterations) if:

- $p_n = O(\exp(Cn^{1-\xi}))$ $(0 < \xi < 1)$ (high-dimensional)
  essentially exponentially many variables relative to n
- $\sup_n \sum_{j=1}^{p_n} |\beta_{j,n}| < \infty$ ℓ^1-sparseness of true function

i.e. for suitable, slowly growing $m = m_n$:

$$\mathbf{E}_X|\hat{f}_{m_n,n}(X) - f_n(X)|^2 = o_P(1) \ (n \to \infty)$$

analogous result also for multivariate autoregressive time series (Lutz & PB, 2005)
(assuming some polynomial decay for α-mixing coefficients)

binary lymph node classification in breast cancer using gene expressions:

a high noise problem

$n = 49$ samples, $p = 7129$ gene expressions

	L_2Boosting	FPLR	Pelora	1-NN	DLDA	SVM
CV-misclassif.err.	17.7%	35.25%	27.8%	43.25%	36.12%	36.88%

multivariate gene selection — best 200 genes from Wilcox.

L_2Boosting selected 42 out of $p = 7129$ genes

for this data-set: not good prediction, with any of the methods

but L_2Boosting may be a reasonable(?) multivariate gene selection method

3. Lasso and L_2Boosting

Efron et al. (2004): intriguing relation between L_2Boosting and Lasso for linear model satisfying a positive cone condition for the design matrix: roughly,

L_2Boosting with comp.wise linear LS and "infinitesimally" small ν
yields a path (as iterations increase)
which contains all Lasso solutions when varying λ

$\leadsto$ computationally interesting to produce all Lasso solutions in one sweep of boosting

for linear models: LARS (Efron et al., 2004) is computationally very clever and efficient for computing all Lasso solutions

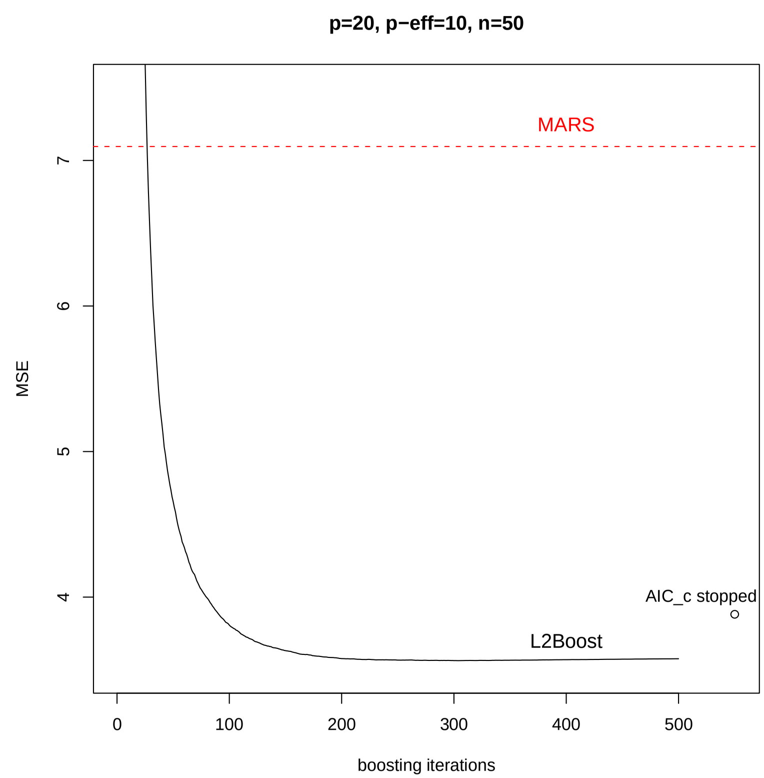
Boosting is algorithmically much more generic than Lasso

other loss function than L_2, nonparametric models, qualitative constraints, ...

Boosting with nonparametric first-order interactions

base procedure: pairwise smoothing splines ($\mathbb{R}^2 \to \mathbb{R}$) which selects the pair of predictors such that corresponding spline smooth reduces RSS most (fixed d.f.)

$\leadsto$ nonparametric model fit with first-order interactions (structured model fit!)



Friedman #1 model:

$Y = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \mathcal{N}(0,1)$

$X = (X_1, \ldots, X_{20}) \sim \text{Unif.}([0,1]^{20})$

Sample size $n = 50$

Dimension $p = 20, p_{eff} = 5$

4. Sparser than Boosting

consider linear model $Y = X\beta + \varepsilon$

for orthonormal design: $\mathbf{X}^T\mathbf{X} = I$:

L_2Boosting with comp.wise linear LS yields the soft-threshold estimator

Is soft-thresholding a good thing?

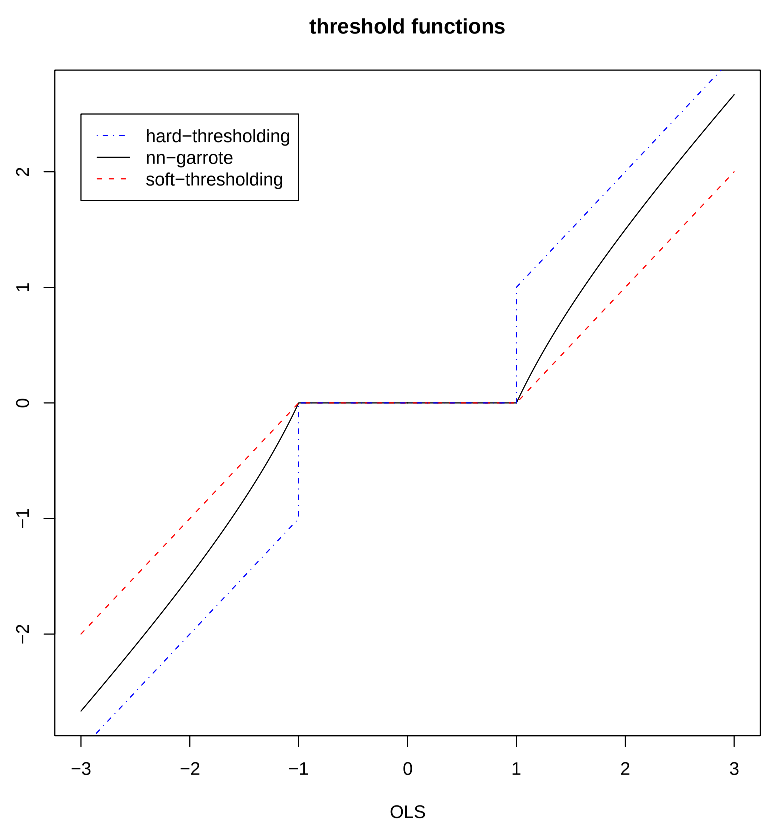
quite a few “yes”-answers (Donoho & Johnstone)

a different story in the very high-dimensional sparse case

$\leadsto$ very slow convergence rates for soft-thresholding (Meinshausen, 2005) suppose that p_{eff} (number of effective predictors) is small but p very large

need large threshold parameter to control the non-effective predictors

$\rightsquigarrow$ strong bias of soft-thresholding



and "analogously" for non-orthogonal design

4.1. SparseL_2Boosting

(PB and Yu, 2005)

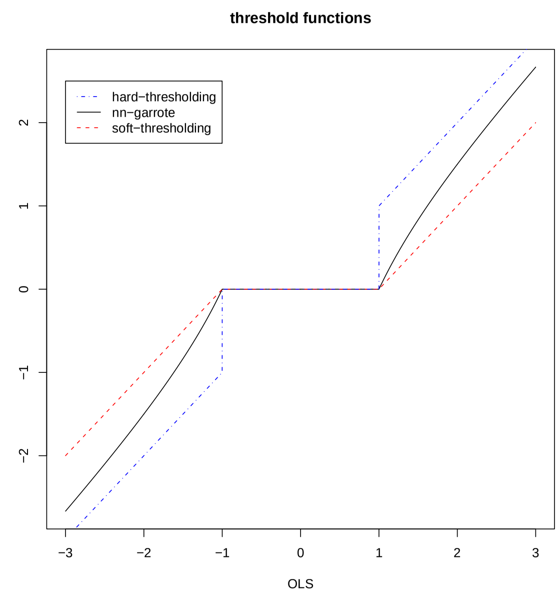
instead of minimizing RSS in every iteration,

minimize a final prediction error (FPE) criterion: we propose gMDL,

$$\hat{\theta}_m = \arg\min_{\theta(\cdot)} \sum_{i=1}^{n} (Y_i - \hat{f}_{m-1}(X_i) - \theta(X_i))^2 + \underbrace{\text{gMDL-penalty}}_{\text{requires d.f. for boosting}}$$

d.f. for boosting via trace of hat-matrices for orthonormal linear model:

SparseL_2Boosting with componentwise linear least squares yields Breiman's nonnegative garrote estimator (PB & Yu, 2005)



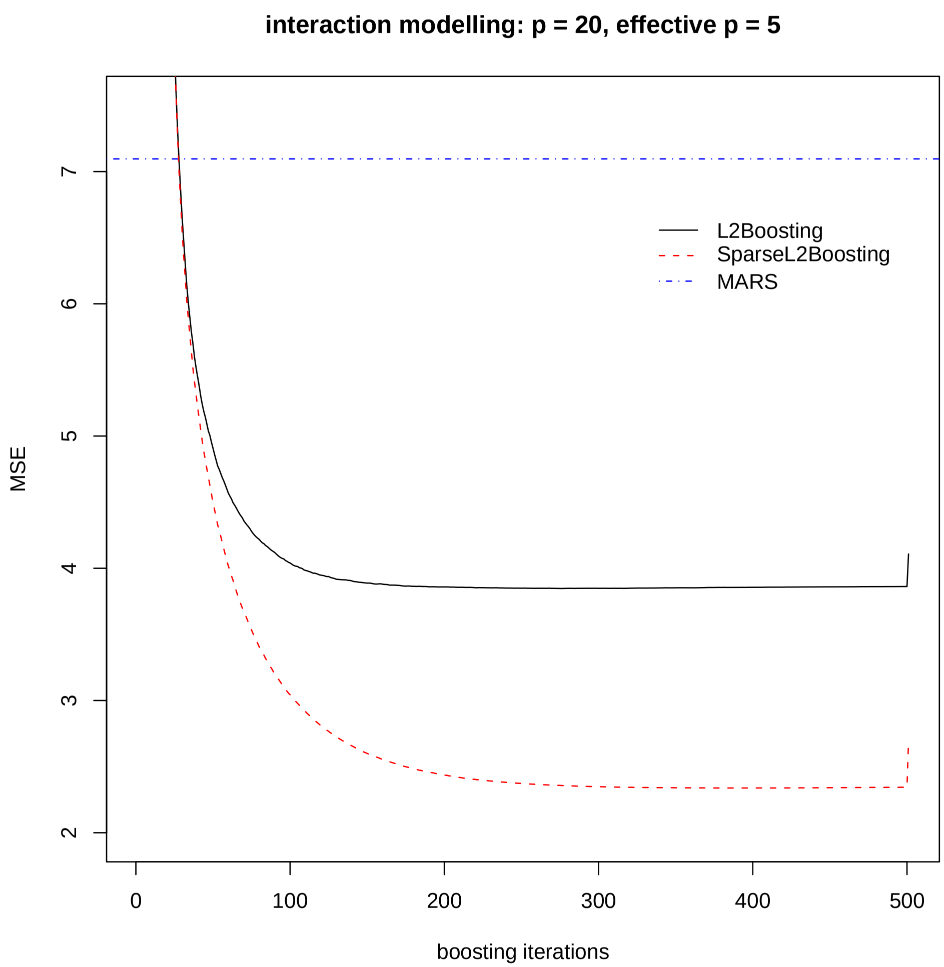
- SparseL_2Boosting yields sparser solutions than L_2Boosting
- SparseL_2Boosting still very generic (although less generic than L_2Boosting)
  e.g. nonparametric problems, non-quadratic loss functions

Linear modeling: L_2Boosting with componentwise linear LS

sample size $n = 50$, dimension $p = 50$

model	SparseL_2Boosting	L_2Boosting
$Y = 1 + 5X^{(1)} + 2X^{(2)} + X^{(3)} + \mathcal{N}(0,1)$ $X = (X^{(1)}, \ldots, X^{(49)}) \sim \mathcal{N}_{49}(0, I)$	0.16 (0.0018)	0.46 (0.0041)
$Y = \sum_{j=1}^{50} \beta_j X^{(j)} + \mathcal{N}(0,1)$ $\beta_1, \ldots, \beta_{50} \sim$ Double-Exponential; X as above	3.64 (0.188)	2.19 (0.083)

Nonparametric first-order interaction modeling



Friedman #1 model:

$Y = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 +$

$10X_4 + 5X_5 + \mathcal{N}(0, 1)$

$X = (X_1, \ldots, X_{20}) \sim$ Unif.$([0, 1]^{20})$

Sample size $n = 50$

Dimension $p = 20, p_{eff} = 5$

5. Conclusions

Boosting can be used as an estimation and regularization method
within some structured models

- Boosting is generic
- Boosting is computationally attractive, in particular in complex situations
- Boosting has some good asymptotic properties
  consistency in very high-dimensional problems
  minimax rate optimal for one-dimensional curve estimation (PB & Yu, 2003)
- SparseL_2Boosting can be very worthwhile if the truth is very sparse