Estimation and Accuracy after Model Selection

Bradley Efron

Stanford University

Estimation after Model Selection

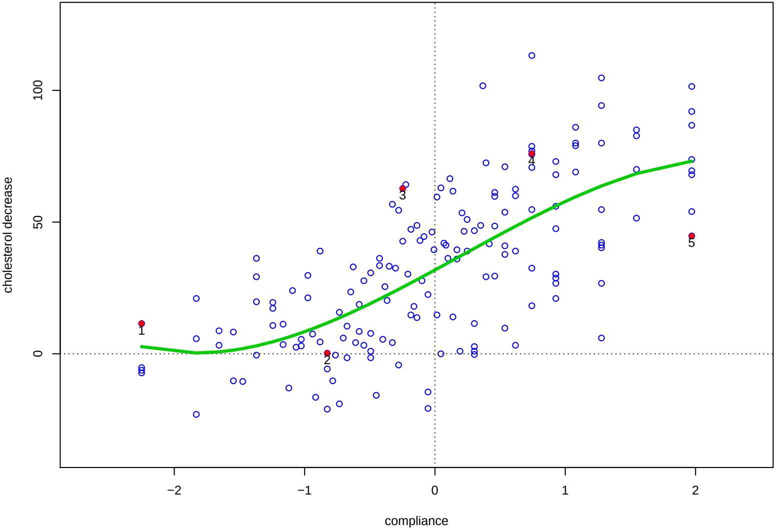
- My usual bad practice:
  (a) look at data
  (b) choose model (linear, quad, cubic . . . ?)
  (c) fit estimates using chosen model
  (d) analyze as if pre-chosen
- Today: include model selection process in the analysis
- Question   Effects on standard errors, confidence intervals, etc.?
- *Two examples*: nonparametric, parametric

Cholesterol Data

- $n = 164$ men took Cholestyramine for ~ 7 years
- $x =$ *compliance measure* (adjusted: $x \sim \mathcal{N}(0,1)$)
- $y =$ *cholesterol decrease*
- Regression y on x?

  [wish to estimate $\mu_j = E\{y|x_j\},\ j = 1, 2, \ldots, n$]

Cholesterol data, n=164 subjects: cholesterol decrease plotted versus adjusted compliance; Green curve is OLS cubic regression; Red points indicate 5 featured subjects



C_p *Selection Criterion*

- Regression Model $\underset{n\times 1}{\boldsymbol{y}} = \underset{n\times m}{X} \underset{m\times 1}{\beta} + \underset{n\times 1}{\boldsymbol{e}}$ $\left[e_i \sim (0, \sigma^2)\right]$
- *C_p Criterion* $\boxed{\left\|\boldsymbol{y} - X\hat{\beta}\right\|^2 + 2m\sigma^2}$

  $\hat{\beta}$ = OLS estimate, m = "degrees of freedom"
- Model selection: from possible models $X_1, X_2, X_3, \ldots,$

  choose the one minimizing C_p.
- Then use OLS estimate from chosen model.

C_p *for Cholesterol Data*

Model	df	$C_p - 80000$	(Boot %)
$\mathcal{M}_1$ (linear)	2	1132	(19%)
$\mathcal{M}_2$ (quad)	3	1412	(12%)
$\mathcal{M}_3$ (cubic)	4	667	(34%)
$\mathcal{M}_4$ (quartic)	5	1591	(8%)
$\mathcal{M}_5$ (quintic)	6	1811	(21%)
$\mathcal{M}_6$ (sextic)	7	2758	(6%)

($\sigma = 22$ from "full model" $\mathcal{M}_6$)

Nonparametric Bootstrap Analysis

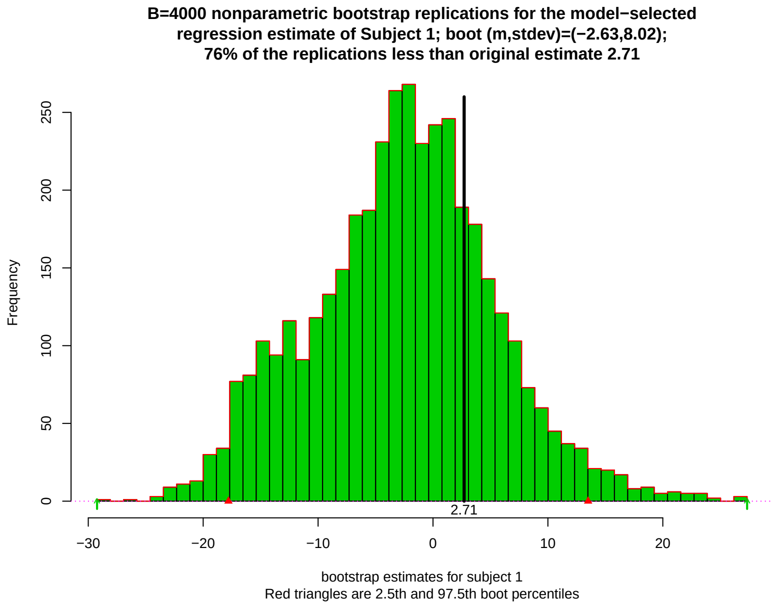
- data $= \{(x_i, y_i),\ i = 1, 2, \ldots, n = 164\}$ gave original estimate

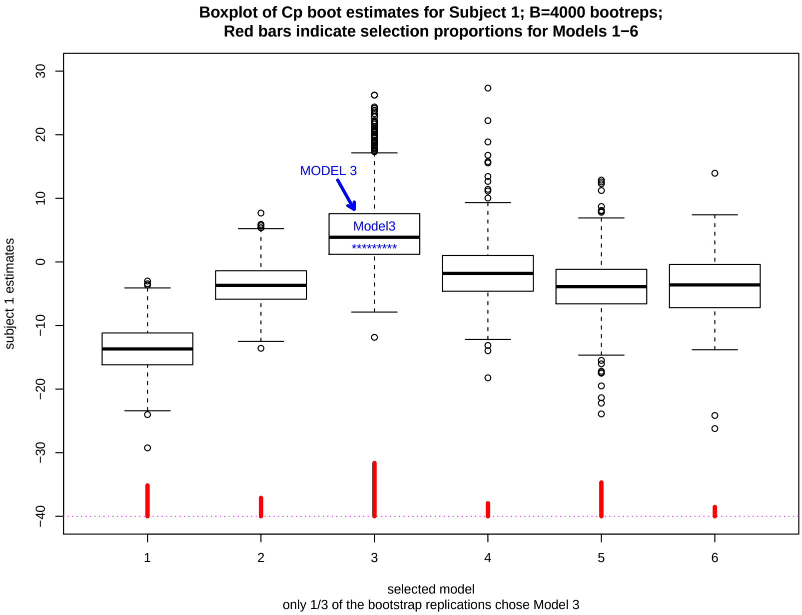
$$\hat{\mu} = X_3 \hat{\beta}_3$$

- Bootstrap data set data$^* = \left\{(x_j, y_j)^*,\ j = 1, 2, \ldots, n\right\}$ where $(x_j, y_j)^*$ drawn randomly and with replacement from data:

$$\text{data}^* \underset{C_p}{\longrightarrow} m^* \underset{\text{OLS}}{\longrightarrow} \hat{\beta}^*_{m^*} \longrightarrow \hat{\mu}^* = X_{m^*} \hat{\beta}^*_{m^*}$$

- I did this all $B = 4000$ times.


B=4000 nonparametric bootstrap replications for the model−selected
regression estimate of Subject 1; boot (m,stdev)=(−2.63,8.02);
76% of the replications less than original estimate 2.71
Frequency
0
50
100
150
200
250
2.71
−30
−20
−10
0
10
20
bootstrap estimates for subject 1
Red triangles are 2.5th and 97.5th boot percentiles


Boxplot of Cp boot estimates for Subject 1; B=4000 bootreps;
Red bars indicate selection proportions for Models 1−6
MODEL 3
Model3
****
subject 1 estimates
30
20
10
0
−10
−20
−30
−40
1
2
3
4
5
6
selected model
only 1/3 of the bootstrap replications chose Model 3

Bootstrap Smoothing

- Idea Replace original estimator $t(\boldsymbol{y})$ with bootstrap average

$$\boxed{s(\boldsymbol{y}) = \sum_{i=1}^{B} t\left(\boldsymbol{y}_i^*\right)\Big/B}$$

- Model averaging
- Same as *bagging* ("bootstrap aggregation" / Breiman)
- Removes discontinuities • Reduces variance

Bootstrap Confidence Intervals

- Standard: $\hat{\mu} \pm 1.96\, \widehat{\text{se}}$
- Percentile: $\left[\hat{\mu}^{*(.025)}, \hat{\mu}^{*(.975)}\right]$
- Smoothed Standard: $\tilde{\mu} \pm 1.96\, \widetilde{\text{se}}$
- BCa/ABC: corrects percentiles for bias and changing se

Accuracy Theorem

- Notation $\quad s_0 = s(\boldsymbol{y}), \quad t_i^* = t(\boldsymbol{y}_i^*),\ i = 1, 2, \ldots B$
- $N_{ij}^* = \#$ of times jth data point appears in ith boot sample
- $\text{cov}_j = \sum_{i=1}^{B} N_{ij}^* \cdot \left(t_i^* - s_0\right) \big/ B \quad \left[\text{covariance } N_{ij}^* \text{ with } t_i^*\right]$

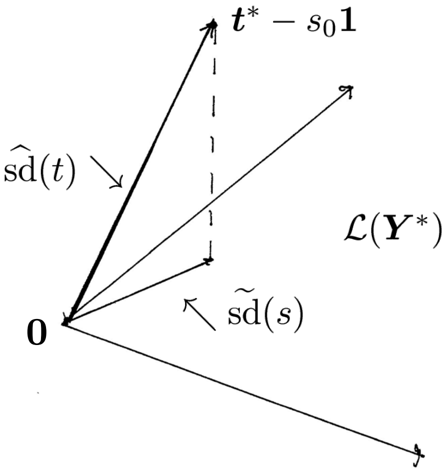
Theorem

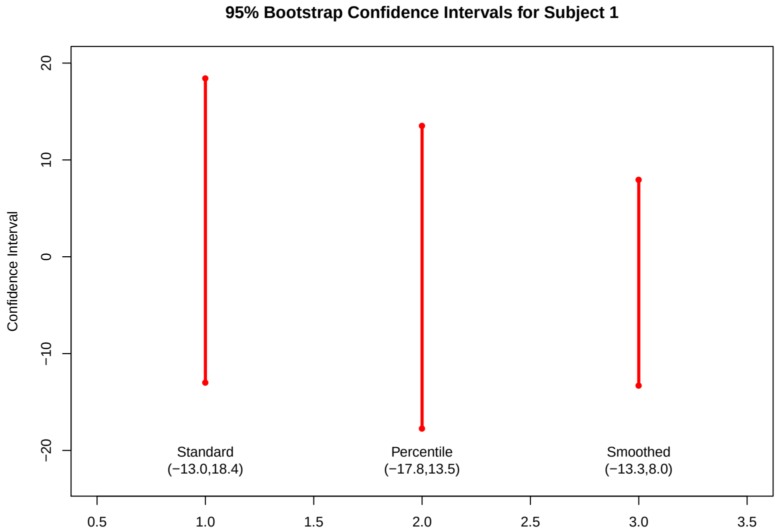
The delta method standard deviation estimate for s_0 is

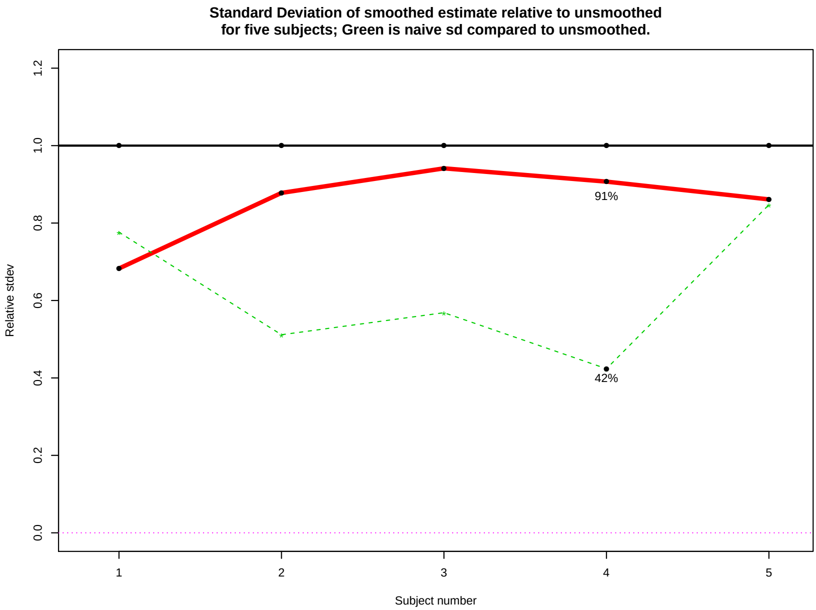
$$\boxed{\widetilde{\text{sd}} = \left[\sum_{j=1}^{n} \text{cov}_j^2\right]^{1/2}},$$

always $\leq \left[\sum_{i=1}^{B} \left(t_i^* - s_0\right)^2 \big/ B\right]^{1/2}$*, the boot stdev for $t(\boldsymbol{y})$.*

Projection Interpretation




95% Bootstrap Confidence Intervals for Subject 1
Confidence Interval
20
10
0
−10
−20
Standard
(−13.0,18.4)
Percentile
(−17.8,13.5)
Smoothed
(−13.3,8.0)
0.5
1.0
1.5
2.0
2.5
3.0
3.5



Standard Deviation of smoothed estimate relative to unsmoothed
for five subjects; Green is naive sd compared to unsmoothed.
Relative stdev
0.0
0.2
0.4
0.6
0.8
1.0
1.2
91%
42%
1
2
3
4
5
Subject number

How Many Bootstrap Replications Are Enough?

- How accurate is $\widetilde{\text{sd}}$? *Jackknife*
- Divide the 4000 bootreps t_i^* into 20 groups of 200 each
- Recompute $\widetilde{\text{sd}}$ with each group removed in turn
- Jackknife gave coef variation $\left(\widetilde{\text{sd}}\right) \doteq 0.02$ for all 164 subjects

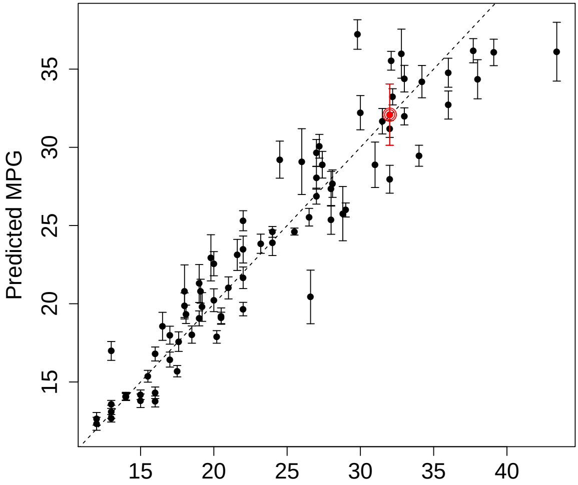
  (could have stopped at $B = 500$)

Model Probability Estimates

- 34% of the 4000 bootreps chose the cubic model
- Poor man's Bayes posterior prob for "cubic"
- How accurate is that 34%?
- Apply accuracy theorem to indicator function for choosing "cubic"

Model	Boot %	± Standard Error
$\mathcal{M}_1$ (linear)	19%	±24
$\mathcal{M}_2$ (quad)	12%	±18
$\mathcal{M}_3$ (cubic)	34%	±24
$\mathcal{M}_4$ (quartic)	8%	±14
$\mathcal{M}_5$ (quintic)	21%	±27
$\mathcal{M}_6$ (sextic)	6%	±6

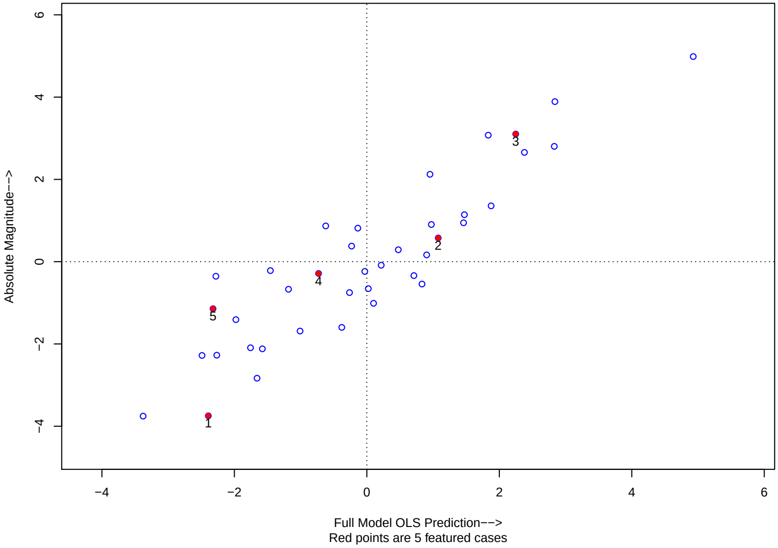
Random Forest Predictions & Bagged Stdevs: Automobile Data



The Supernova Data

- data $= \left\{(x_j, y_j),\ j = 1, 2, \ldots, n = 39\right\}$
- $y_j =$ absolute magnitude of Type Ia supernova
- $x_j =$ vector of 10 spectral energies (350–850nm)
- Full Model $\quad \boldsymbol{y} = \underset{39\times 10}{X}\beta + \boldsymbol{e} \qquad \left[\boldsymbol{e}_i \overset{\text{ind}}{\sim} \mathcal{N}(0,1)\right]$

Adjusted absolute magnitudes for 39 Type1A supernovas plotted versus OLS predictions from 10 spectral measurments; Naive R2 (squared correlation)=.82; Adjusted R2=.62



Lasso Model Selection

- Lasso estimate is $\hat{\beta}$ minimizing $\left\| \boldsymbol{y} - X\beta \right\|^2 + \lambda \sum_1^p |\beta_k|$
- *Shrinks* OLS estimates toward zero (all the way for some)
- Degrees of freedom "m" $=$ number of nonzero $\hat{\beta}_k$'s
- Model selection: Choose λ (or m) to maximize adjusted R^2.
- Then $\hat{\boldsymbol{\mu}} = X\hat{\beta}_m$.

Lasso for the Supernova Data

λ	m (# nonzero $\hat{\beta}_k$'s)	Naive R^2	Adjusted R^2	
63.0	**1**	.17	**.12**	
12.9	**4**	.77	**.72**	
3.56	**7**	.81	**.73**	*(Selected)*
0.50	**9**	.82	**.71**	
0	**10**	.82	**.69**	*(OLS)*

Parametric Bootstrap Smoothing

- *Original estimates*

$$\boldsymbol{y} \xrightarrow{\text{Lasso}} m, \hat{\beta}_m \longrightarrow \hat{\boldsymbol{\mu}} = X\hat{\beta}_m$$

- Full model bootstrap $\quad \boldsymbol{y}^* \sim \mathcal{N}_{39}\left(\hat{\boldsymbol{\mu}}_{\text{OLS}}, I\right)$

$$\boldsymbol{y}^* \longrightarrow m^*, \hat{\beta}^*_{m^*} \longrightarrow \hat{\boldsymbol{\mu}}^* = X\hat{\beta}^*_{m^*}$$

- I did this all $B = 4000$ times giving boots for super k

$$\hat{\boldsymbol{\mu}}^*_{ik}, \qquad i = 1, 2, \ldots, 4000$$

- Smoothed estimates $\quad s_k = \sum_{i=1}^{4000} \mu^*_{ik} \Big/ 4000 \qquad [k = 1, 2, \ldots, 39]$

Parametric Accuracy Theorem

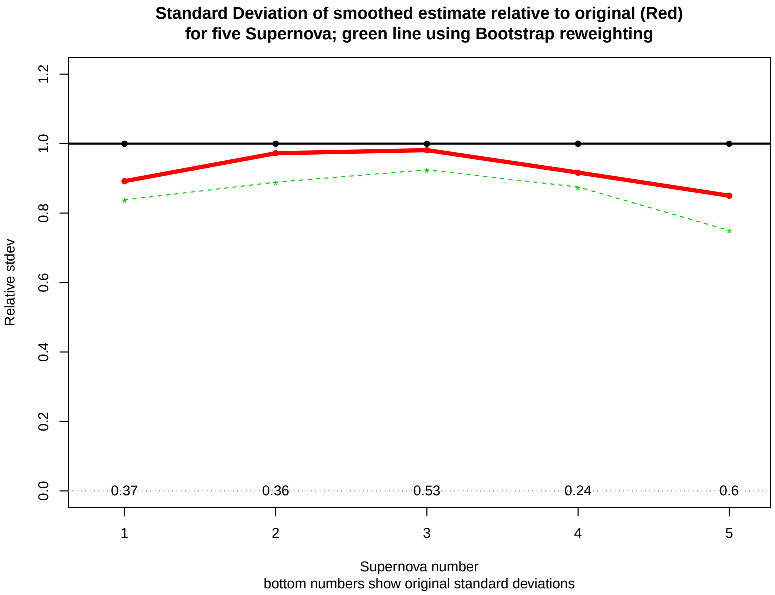
Theorem

The delta method standard deviation estimate for s_k is

$$\boxed{\widehat{\mathrm{sd}}_k = \left[\widehat{\mathbf{cov}}_k' \, \mathcal{G} \, \widehat{\mathbf{cov}}_k\right]^{1/2}},$$

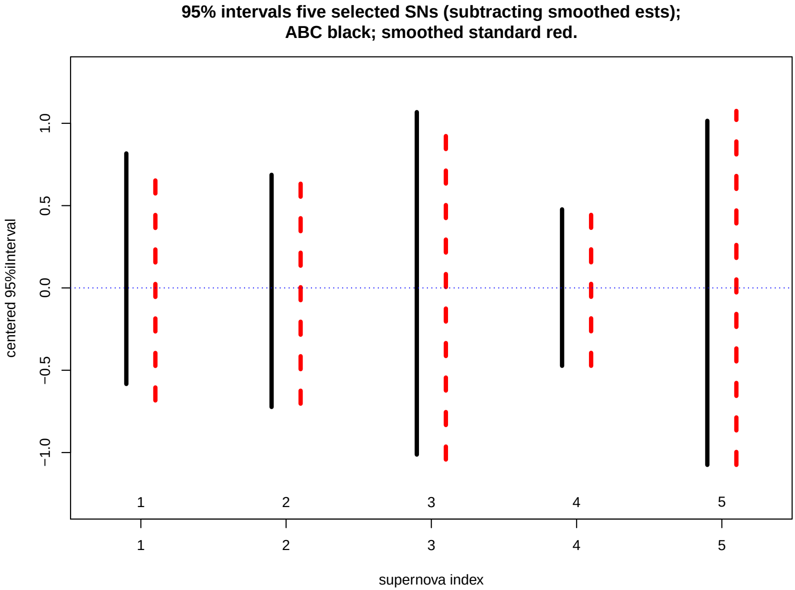
where $\mathcal{G} = X'X$ and $\widehat{\mathbf{cov}}_k$ is bootstrap covariance between $\hat{\beta}^*_{\mathrm{OLS}}$ and $\hat{\boldsymbol{\mu}}^*_{ik}$.

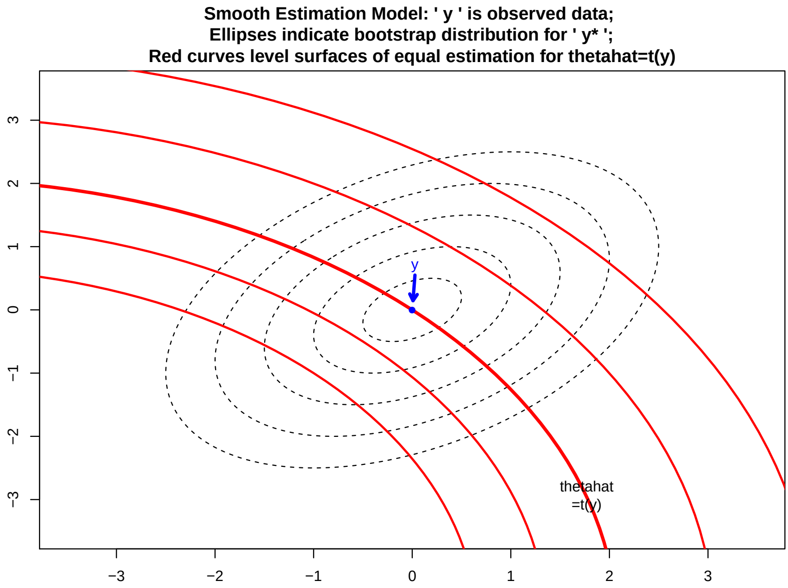
This is always less than the bootstrap estimate of standard deviation for unsmoothed intervals.


Standard Deviation of smoothed estimate relative to original (Red)
for five Supernova; green line using Bootstrap reweighting
Relative stdev
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.37
0.36
0.53
0.24
0.6
1
2
3
4
5
Supernova number
bottom numbers show original standard deviations

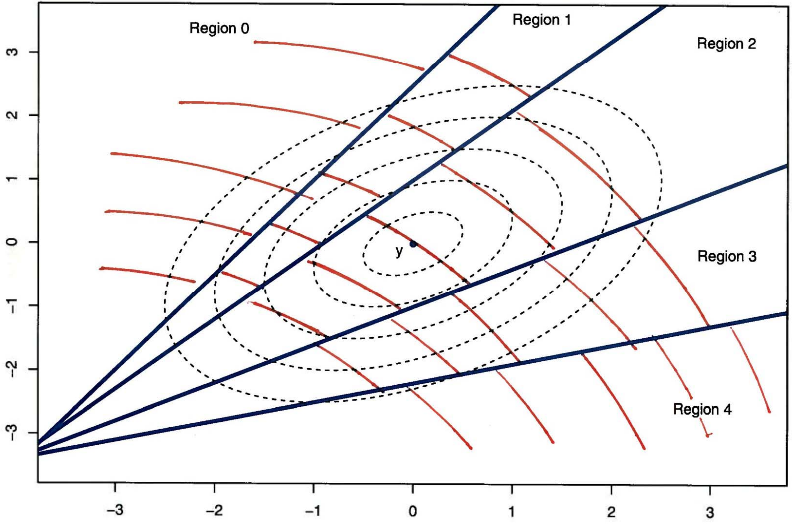
Better Confidence Intervals

- Smoothed standard intervals and percentile intervals have coverage errors of order $O\left(1\middle/\sqrt{n}\right)$.
- "ABC" intervals have errors $O(1/n)$: corrects for bias and "acceleration" (change in stdev as estimate varies).
- Uses local reweighting for 2nd order correction.


95% intervals five selected SNs (subtracting smoothed ests);
ABC black; smoothed standard red.
centered 95%iInterval
1.0
0.5
0.0
−0.5
−1.0
1
2
3
4
5
supernova index


Smooth Estimation Model: ' y ' is observed data;
Ellipses indicate bootstrap distribution for ' y* ';
Red curves level surfaces of equal estimation for thetahat=t(y)
y
thetahat
=t(y)
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

Estimation with model selection: now the curves of equal estimation are discontinuous across the Model region boundaries



References

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.* 41: 802–837, **conservative frequentist intervals à la Tukey, Scheffé**.

Buja, A. and Stuetzle, W. (2006). Observations on bagging. *Statist. Sinica* 16: 323–351, **more on bagging**.

DiCiccio, T. J. and Efron, B. (1992). More accurate confidence intervals in exponential families. *Biometrika* 79: 231–245, **ABC confidence intervals**.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* 98: 879–899, **model selection asymptotics**.

Sexton, J. and Laake, P. (2009). Standard errors for bagged and random forest estimators. *Comput. Statist. Data Anal.* 53: 801–811, **close development to current work**.

Wager, S., Hastie, T. and Efron, B. (2013). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *ArXiv 1311.4555,* **the Automobile example**.