ROBUST LOCATION AND SCALE ESTIMATION WITH CENSORED OUTCOMES

Jerome H. Friedman

Stanford University

MACHINE LEARNING

$$y = F(\mathbf{x}, \mathbf{z})$$

$y =$ outcome variable

$\mathbf{x} = (x_1 \cdot \cdot\cdot, x_p)$ observed predictor variables

$\mathbf{z} = (z_1, z_2, \cdots)$ other variables

Goal: estimate $E\,[y \,|\, \mathbf{x}]$ given data $\{y_i, \mathbf{x}_i\}_{i=1}^N$

STATISTICAL MODEL

$$y = f(\mathbf{x}) + s(\mathbf{x}) \cdot \epsilon$$

$f(\mathbf{x}) = E\left[y \,|\, \mathbf{x}\right]$ location function

$s(\mathbf{x}) > 0$ scale function

$\varepsilon =$ random variable, $E\left[\varepsilon \,|\, \mathbf{x}\right] = 0$

Prediction: $\hat{y} = f(\mathbf{x})$

$s(\mathbf{x}) \cdot \epsilon =$ "irreducible error" (unavoidable)

REDUCIBLE ERROR

$$r(\mathbf{x}) = E\,|\,f(\mathbf{x}) - \hat{f}(\mathbf{x})\,|$$

$f(\mathbf{x})$ = optimal location (target) function

$\hat{f}(\mathbf{x})$ = estimate based on training data & ML method

ML goal: methods to reduce $r(\mathbf{x})$

Statistics goal: methods to estimate $r(\mathbf{x})$

Prediction error (y) = Reducible + Irreducible

Usually: Irreducible $s(\mathbf{x}) >> $ Reducible $r(\mathbf{x})$

USUAL ASSUMPTIONS

$s(\mathbf{x}) = s =$ constant (homoscedasticity)

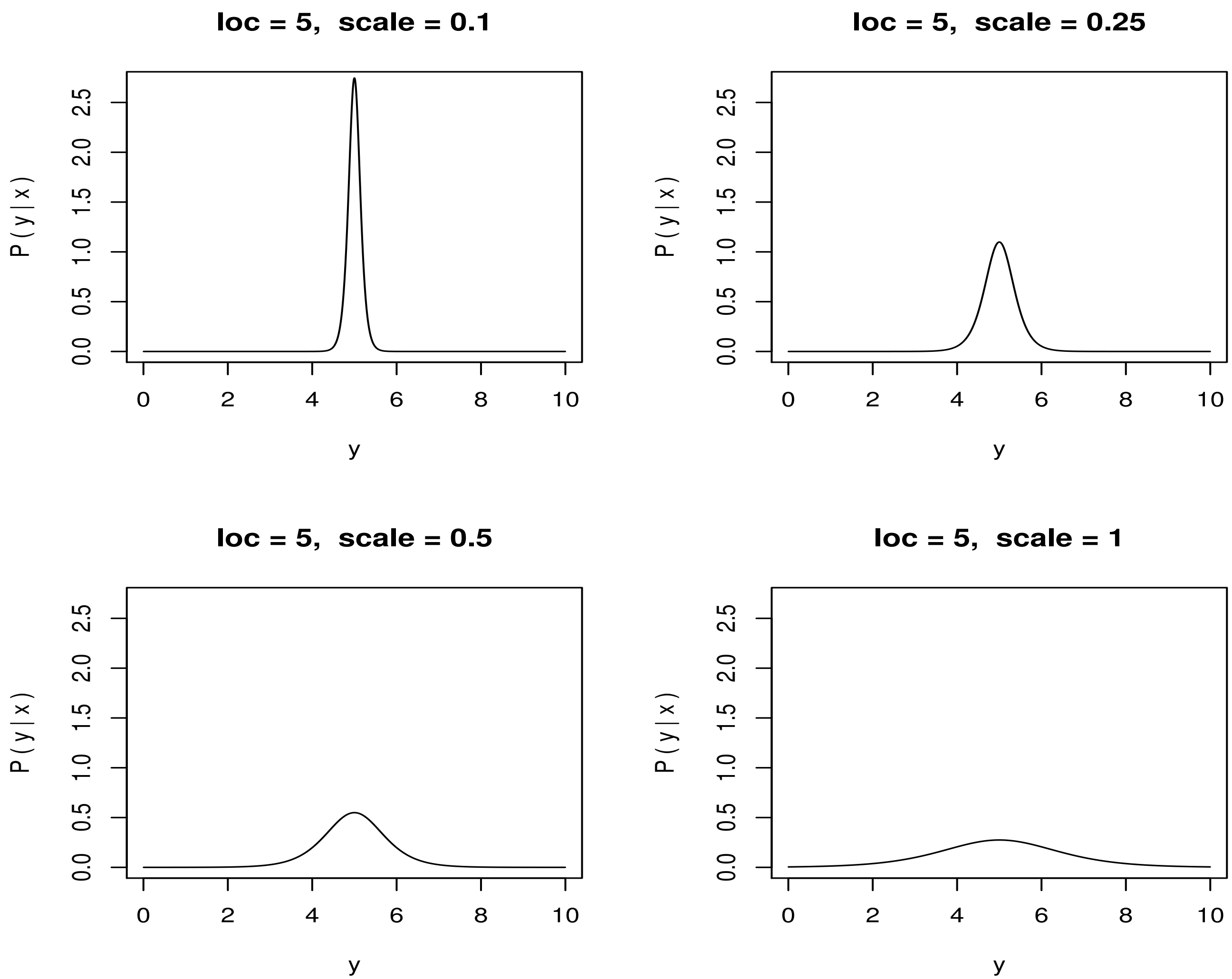
$\epsilon \sim N(0,1)$ (normality)

HOMOSCEDASTICITY

$$F(\mathbf{x}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z}) \quad \text{additive}$$

$$p(\mathbf{x}, \mathbf{z}) \Longrightarrow scale\,[g(\mathbf{z}) \mid \mathbf{x}] = \text{constant}$$

Not very likely


loc = 5, scale = 0.1
loc = 5, scale = 0.25
loc = 5, scale = 0.5
loc = 5, scale = 1
P (y | x)
y
0.0
0.5
1.0
1.5
2.0
2.5
0
2
4
6
8
10

NORMALITY - not very likely either

Tukey:

"small residuals $\simeq$ normal, larger have heavier tails."

Heterodistributionality

Heterodistributionality

Robustness:

Choose compromise $\bar{p}(\epsilon)$

good properties for others

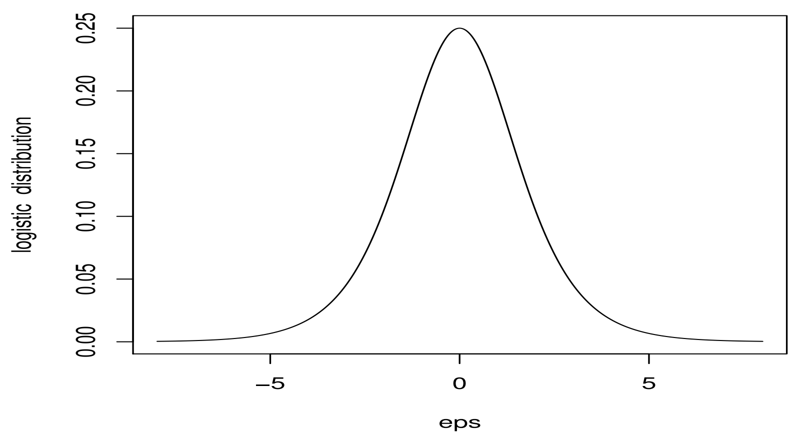
$\bar{p}(\epsilon) =$ normal, not good!

LOGISTIC DISTRIBUTION

$$\epsilon \,|\, \mathbf{x} = (y - f(\mathbf{x}))/s(\mathbf{x})$$

$$\bar{p}(\epsilon) = \frac{e^{-\epsilon}}{s(1+e^{-\epsilon})^2}$$

small $|\,\varepsilon\,| \sim$ normal, large $|\,\varepsilon\,| \sim$ exponential



Prediction: $\hat{y} = \hat{f}(\mathbf{x})$

$$\hat{f}(\mathbf{x}) = \arg\min_{f \in F} \sum_{i=1}^{N} [\varepsilon_i + 2\log(1 + e^{-\varepsilon_i})]$$

$$\varepsilon_i = (y_i - f(\mathbf{x}_i))/s(\mathbf{x}_i)$$

minimized at $f(\mathbf{x}_i) = y_i$ indep $s(\mathbf{x}_i)$

$1/s(\mathbf{x}_i) \sim$ "weight" for obs i

controls relative influence of i to fit

Using incorrect $s(\mathbf{x})$ to estimate $f(\mathbf{x})$

increases variance, not bias

assume $s(\mathbf{x}) =$ constant usually not too bad

ESTIMATE $\hat{s}(\mathbf{x})$

(1) Improve $\hat{f}(\mathbf{x})$ in high variance settings.

(2) Important inferential statistic:

(a) prediction interval $\sim$ accuracy of $\hat{y}$-prediction:

logistic: $IQR\,[\,y \mid f(\mathbf{x})] = 2\,s(\mathbf{x})/\log(3)$

(b) can affect decision

(3) Crucial with censoring

CENSORING (y-value partially known)

Data: $\{y_i, \mathbf{x}_i\}_1^N \rightarrow \{a_i, b_i, \mathbf{x}_i\}_1^N$

$$a_i \leq y_i \leq b_i$$

$a_i = b_i = y_i \Rightarrow y$-value known

$a_i = -\infty \Rightarrow$ censored below b_i

$b_i = \infty \Rightarrow$ censored above a_i

Otherwise: interval censored $[a_i, b_i]$

Special Case

$\{a_i, b_i\} \longrightarrow K$ disjoint intervals (bins):

$K = 2 \Rightarrow$ usual binary logistic regression

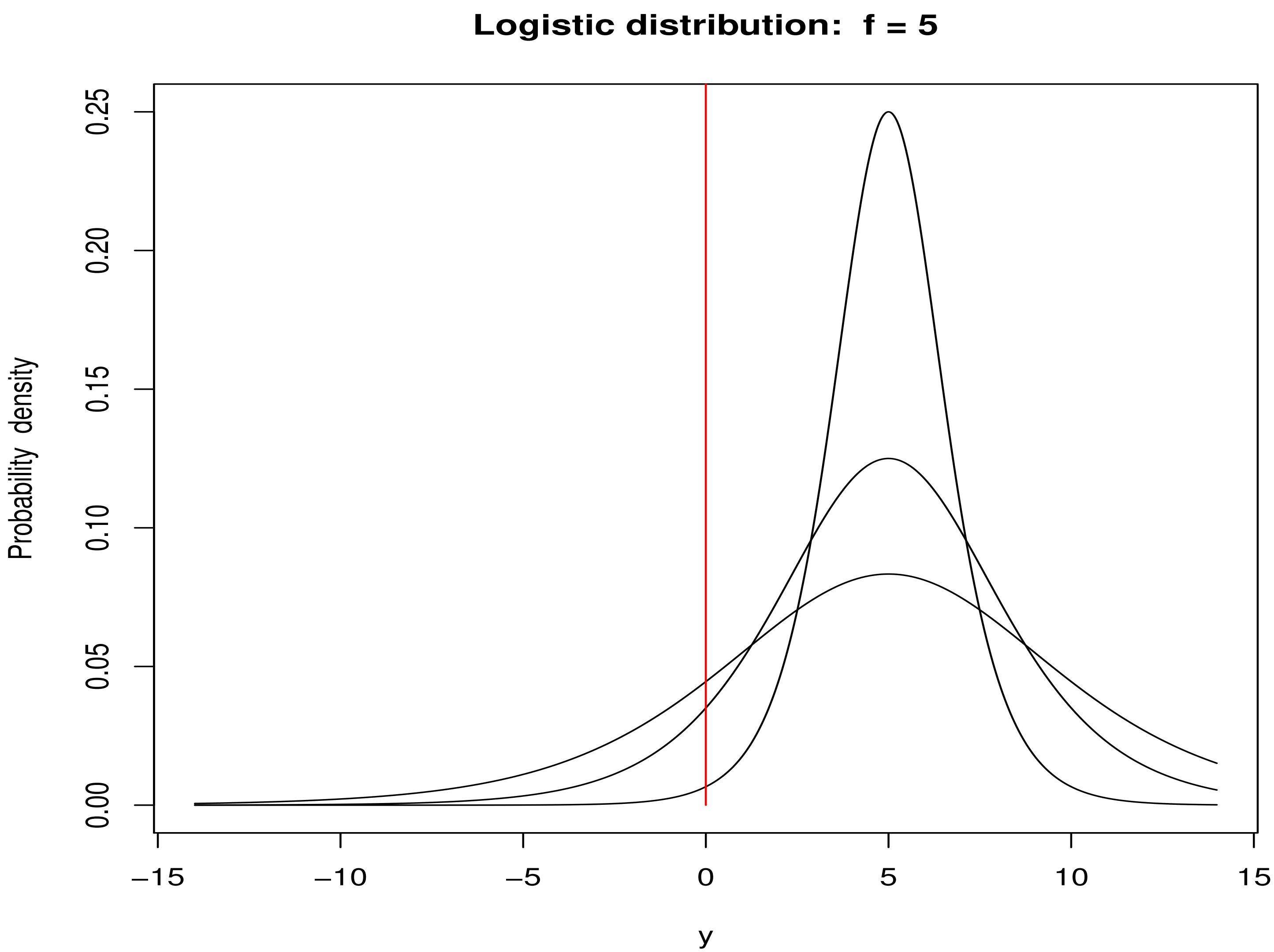
$K > 2 \Rightarrow$ ordered multiclass logistic regression

LIKELIHOOD

$$\Pr(a \leq y \leq b) = \frac{1}{1+e^{-(b-f)/s}} - \frac{1}{1+e^{-(a-f)/s}}$$

Depends strongly on *both* f *and* s

Need to estimate *both* $f(\mathbf{x})$ *and* $s(\mathbf{x})$



Logistic distribution: f = 5
Probability density
0.00
0.05
0.10
0.15
0.20
0.25
−15
−10
−5
0
5
10
15
y

EXERCISE

$$[\hat{f}(\mathbf{x}), \hat{s}(\mathbf{x})] = \arg \min_{(f,s)\in F} \sum_{i=1}^{N} L[a_i, b_i, f(\mathbf{x}_i), s(\mathbf{x}_i)]$$

$$L(a, b, f, s) = -\log\left(\frac{1}{1+e^{-(b-f)/s}} - \frac{1}{1+e^{-(a-f)/s}}\right)$$

PROBLEM

$L(a, b, f, s)$ NOT convex in s

IS convex in $t = 1/s$ $\Rightarrow$ solve for t

Constraint $t > 0$ $\Rightarrow$ solve for $\log(t) = -\log(s)$

GRADIENT BOOSTED TREE ENSEMBLES

Ann. Statist, **29**. 1189 – 1232 (2001)

$\hat{f}(\mathbf{x}) = \sum_{k=1}^{K_f} T_k^{(f)}(\mathbf{x})$

$\log(\hat{s}(\mathbf{x})) = \sum_{k=1}^{K_s} T_k^{(s)}(\mathbf{x})$

$T_k(\mathbf{x}) =$ CART–tree $(\mathbf{x})$

ITERATIVE GRADIENT BOOSTING

Start: $\hat{s}(\mathbf{x}) =$ constant

Loop {

$\hat{f}(\mathbf{x}) =$ tree–boost $f(\mathbf{x})$ given $\hat{s}(\mathbf{x})$

$\log(\hat{s}(\mathbf{x})) =$ tree–boost $\log(s(\mathbf{x}))$ given $\hat{f}(\mathbf{x})$

}

Until no change

DIAGNOSTICS

(1) $median\,[\,y \,|\, f(\mathbf{x})] = f(\mathbf{x})$

(2) $median\,[\,|\,y - f(\mathbf{x})\,|\;|\; s(\mathbf{x})\,] = s(\mathbf{x}) \cdot \log(3)$

(3) $\#\,(y_i \in [u,v] \,|\, f_i \in [g,h]) =$

$$\sum_{f_{i\in[g,h]}} \left(\frac{1}{1+e^{-(v-f_i)/s_i}} - \frac{1}{1+e^{-(u-f_i)/s_i}} \right)$$

$(f_i = \hat{f}(\mathbf{x}_i), \quad s_i = \hat{s}(\mathbf{x}_i))$

California Housing Price Data (STATLIB Repository)

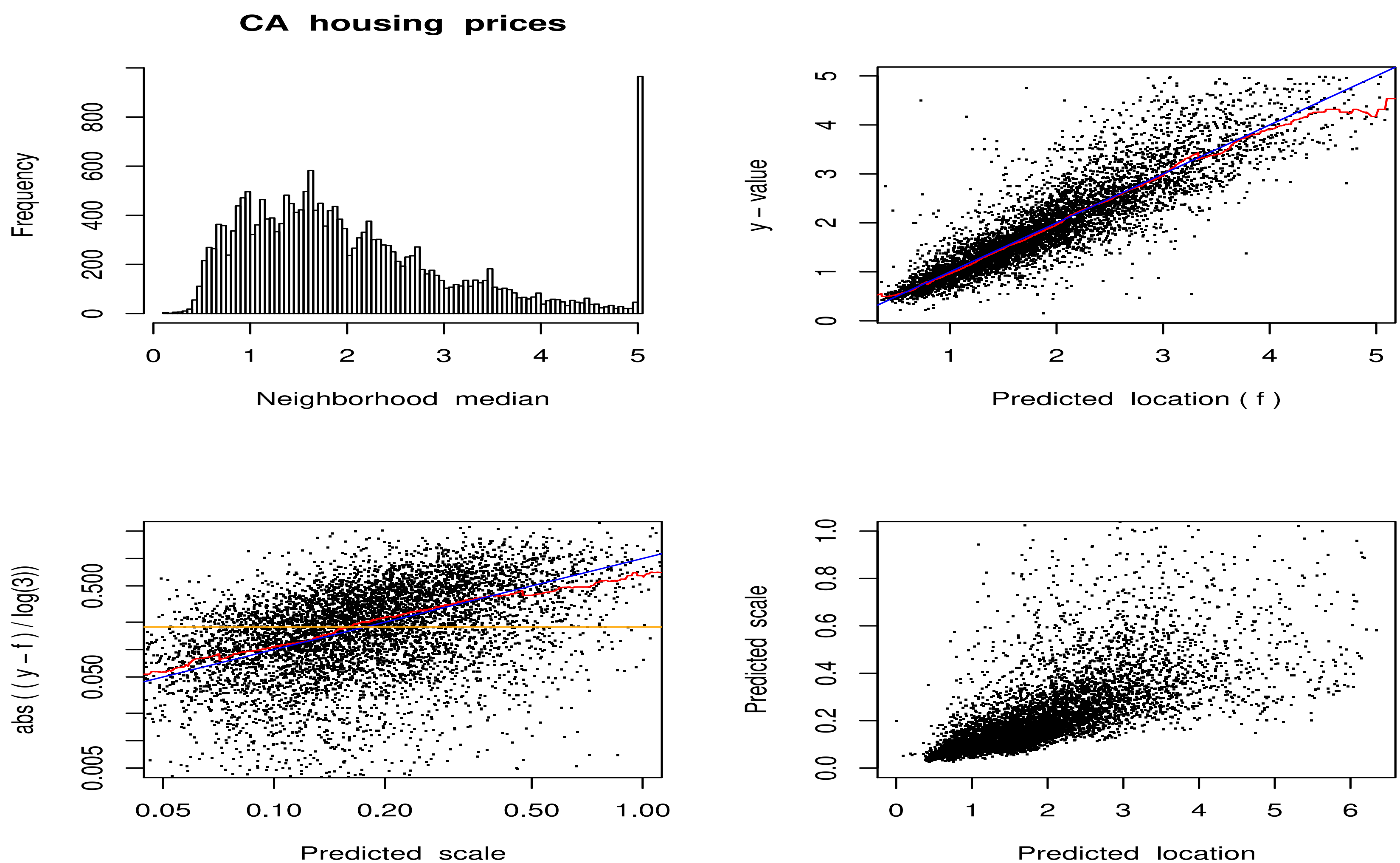
$N =$ 20460 CA neighborhoods (1990 census block groups)

$y =$ Median House Value

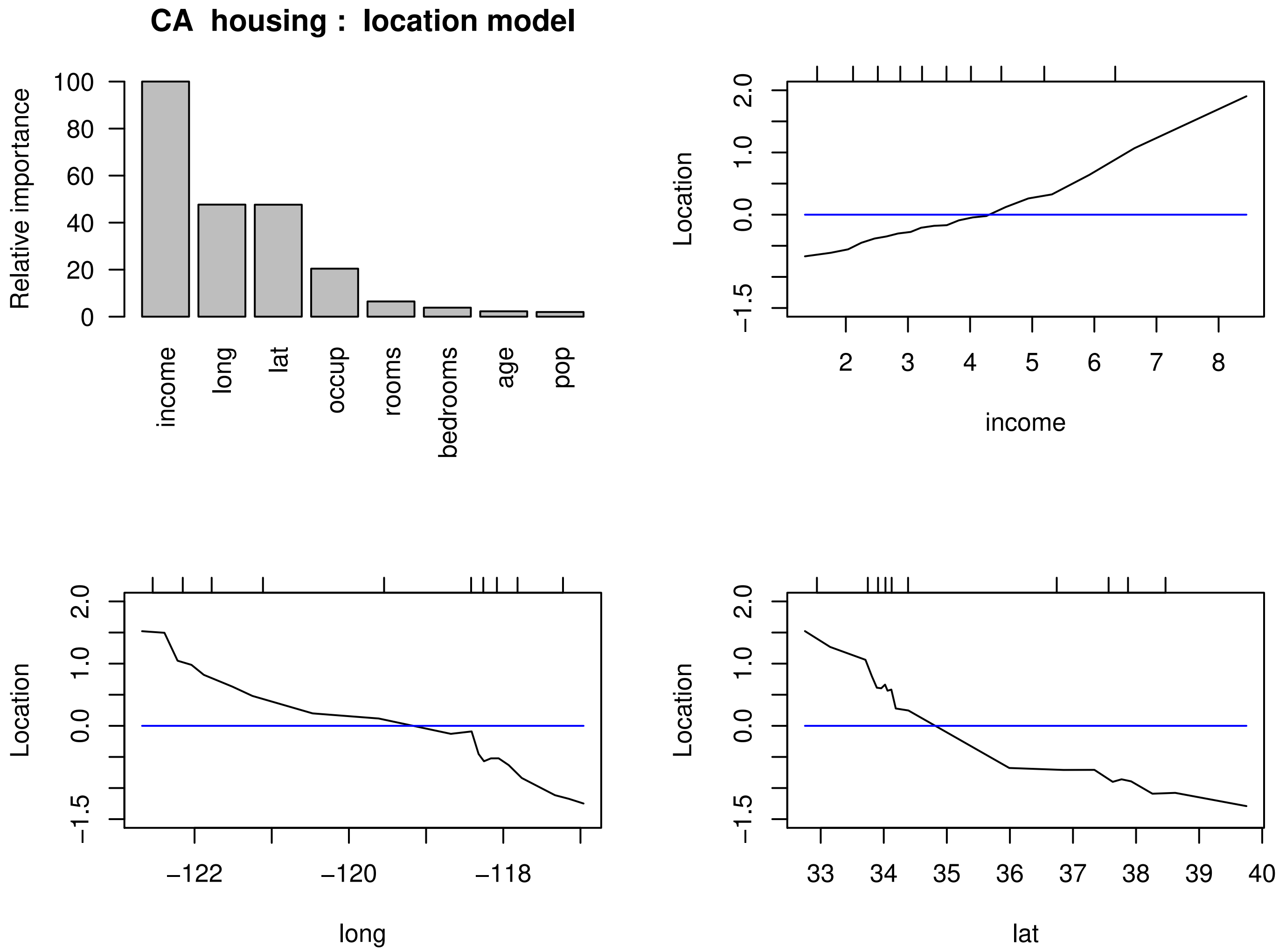
$\mathbf{x} =$ (Median Income, Housing Median Age,

Ave No Rooms, Ave No Bedrooms,

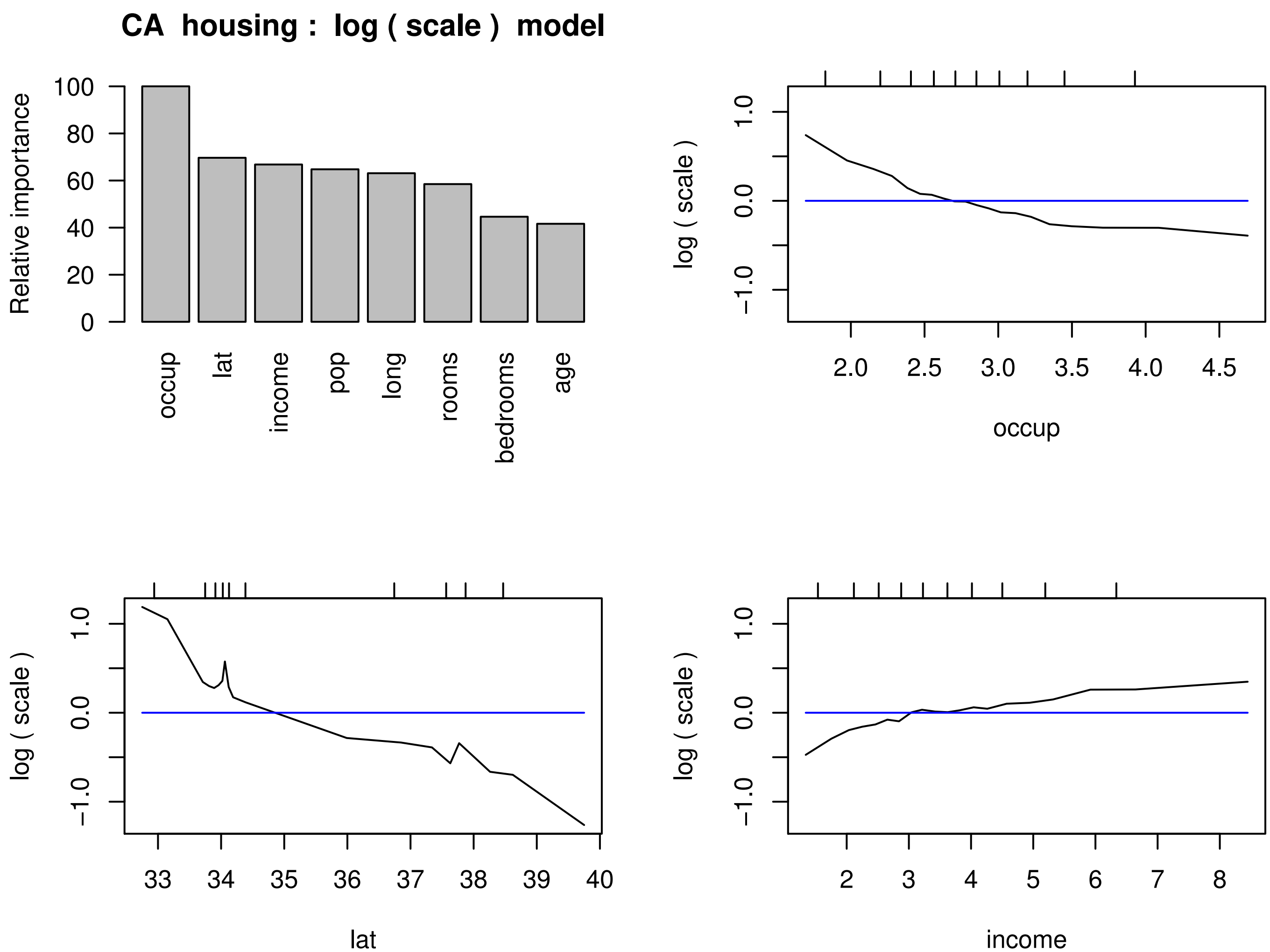
Population, Ave Occupancy, Latitude, Longitude)



CA housing prices
Frequency
0 200 400 600 800
0 1 2 3 4 5
Neighborhood median
y − value
0 1 2 3 4 5
1 2 3 4 5
Predicted location (f)
abs ((y − f) / log(3))
0.005 0.050 0.500
0.05 0.10 0.20 0.50 1.00
Predicted scale
Predicted scale
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3 4 5 6
Predicted location



CA housing : location model
Relative importance
100
80
60
40
20
0
income
long
lat
occup
rooms
bedrooms
age
pop
Location
2.0
1.0
0.0
−1.5
2
3
4
5
6
7
8
income
−122
−120
−118
long
33
34
35
36
37
38
39
40
lat



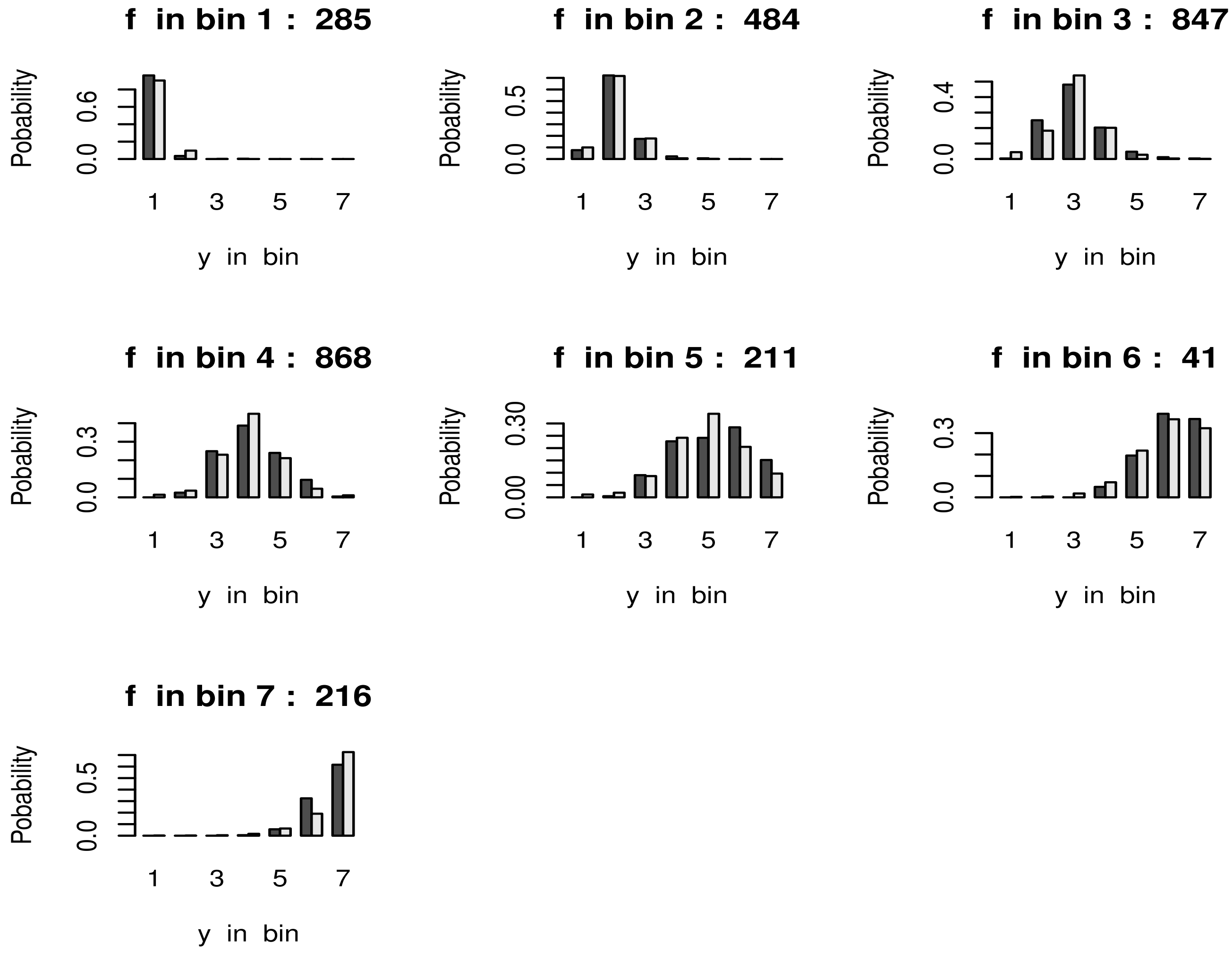
CA housing : log (scale) model
Relative importance
100
80
60
40
20
0
occup
lat
income
pop
long
rooms
bedrooms
age
log (scale)
1.0
0.0
−1.0
2.0
2.5
3.0
3.5
4.0
4.5
occup
33
34
35
36
37
38
39
40
lat
2
3
4
5
6
7
8
income

QUESTIONNAIRE DATA

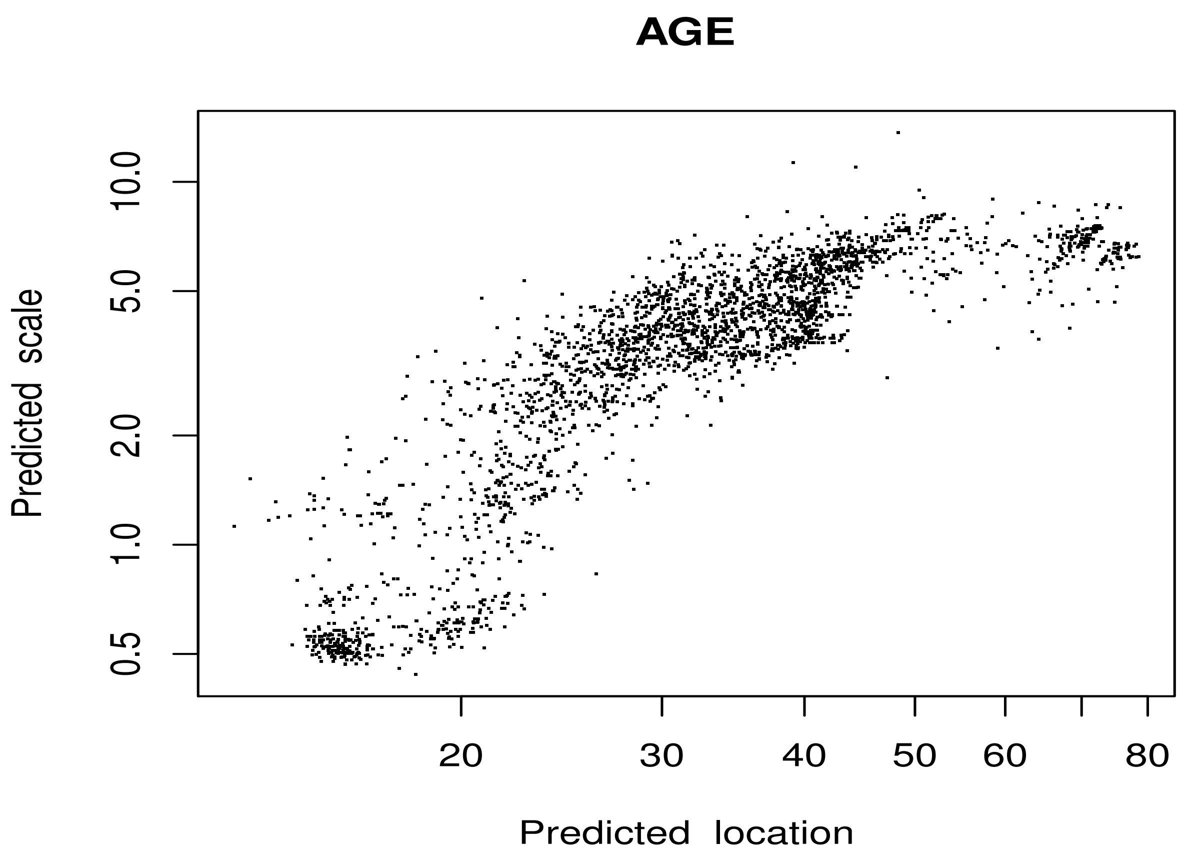
$$N = 8857, \quad p = 13$$

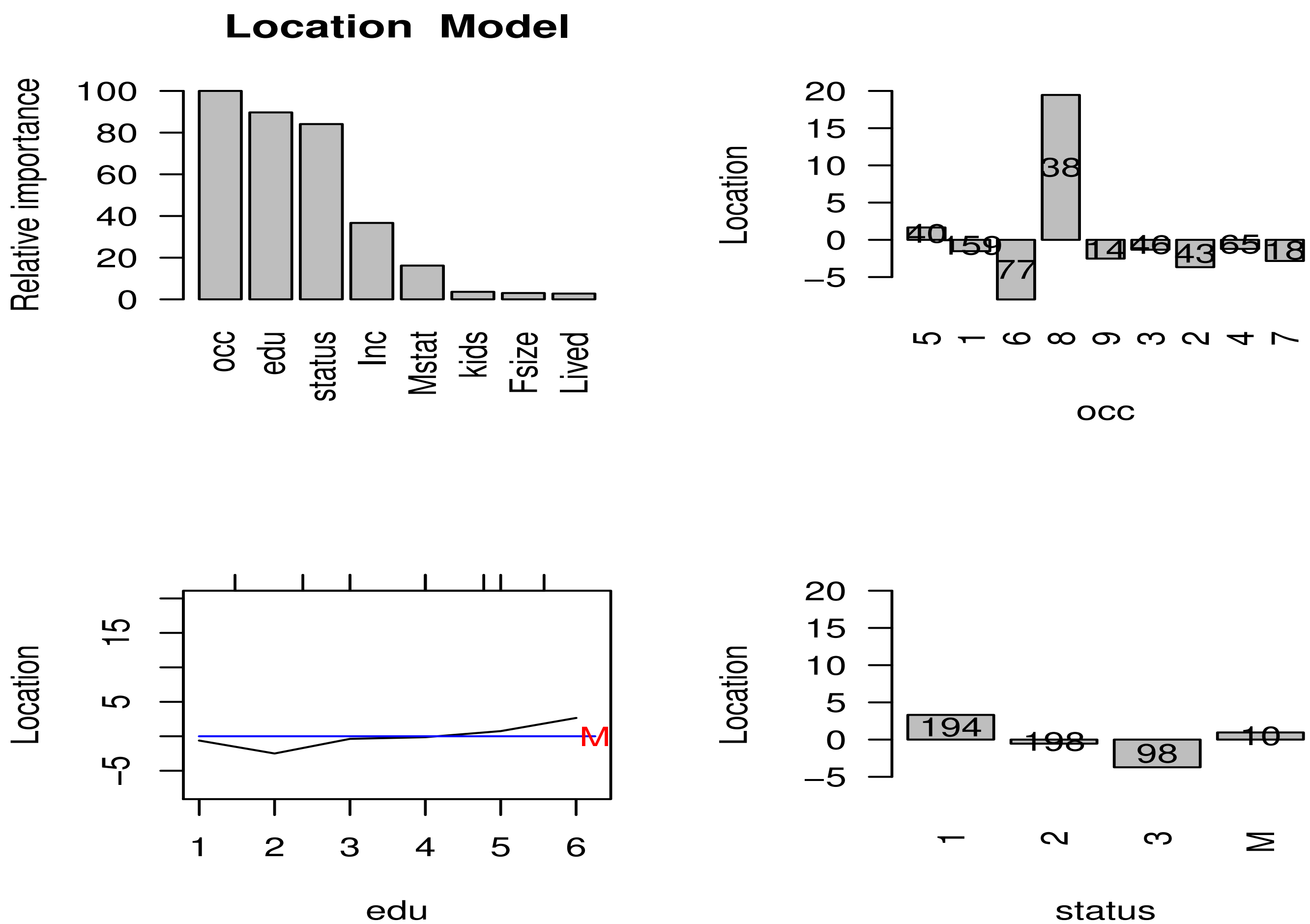
$$y = AGE \in \left\{ \begin{array}{cc|} 14 & 17 \\ 18 & 24 \\ 25 & 34 \\ 35 & 44 \\ 45 & 54 \\ 55 & 64 \\ 65 & \infty \end{array} \right.$$

$$
\begin{aligned}
\mathbf{x} = (&\text{Occupation, Type of Home, Sex,}\\
&\text{Marital Status, Education,Income,}\\
&\text{Lived in BA, Dual Incomes, Persons in}\\
&\text{Household, Persons in Household} < 18,\\
&\text{Householder Status, Ethnicity, Language})
\end{aligned}
$$

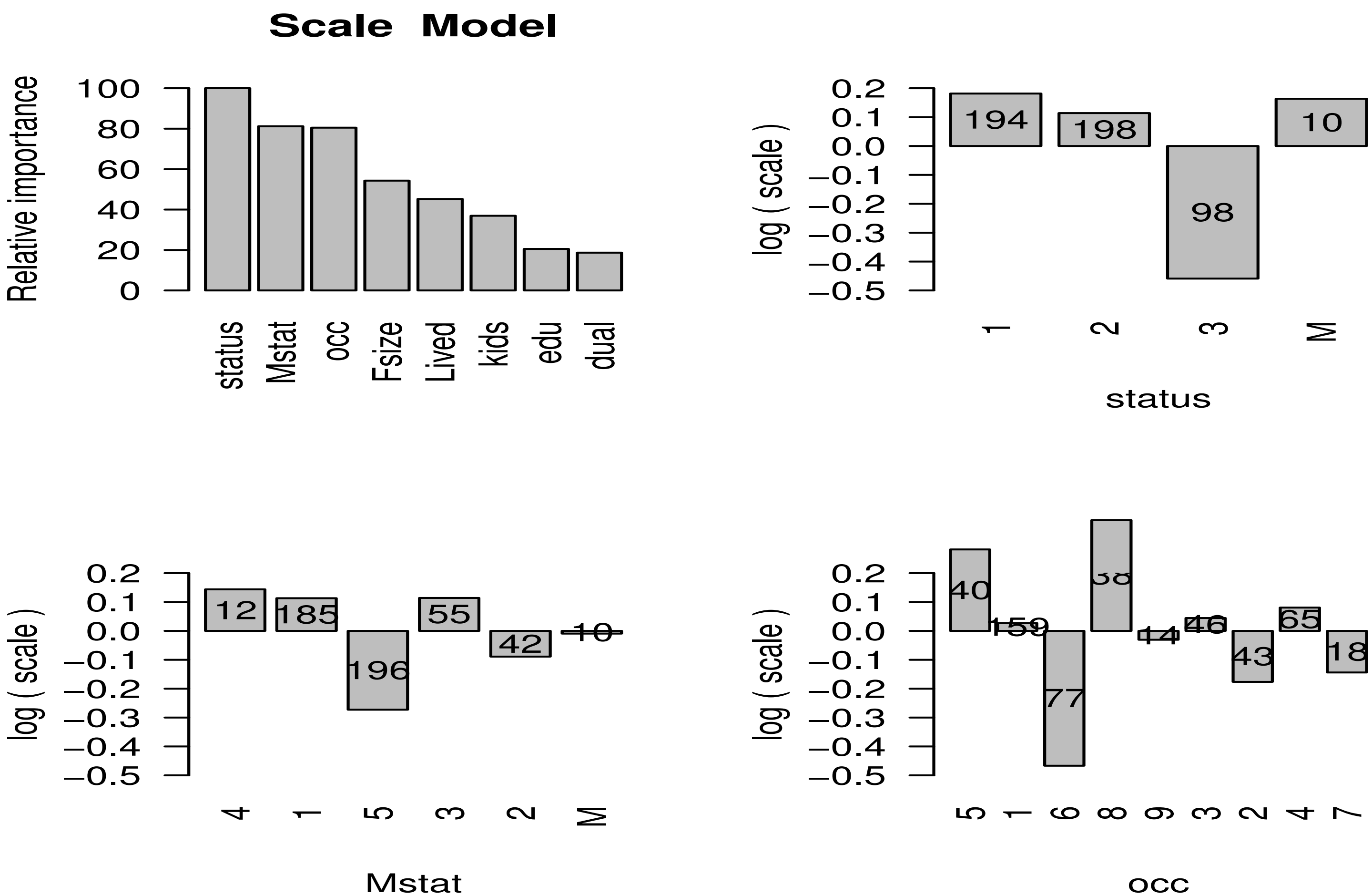


AGE predictions


AGE
Predicted scale
10.0
5.0
2.0
1.0
0.5
20
30
40
50
60
80
Predicted location



Location Model
Relative importance
100
80
60
40
20
0
occ
edu
status
Inc
Mstat
kids
Fsize
Lived
Location
20
15
10
5
0
−5
40
159
77
38
14
46
43
65
18
5
1
6
8
9
3
2
4
7
occ
Location
15
5
−5
M
1
2
3
4
5
6
edu
Location
194
198
98
10
1
2
3
M
status


Scale Model
Relative importance
100
80
60
40
20
0
status
Mstat
occ
Fsize
Lived
kids
edu
dual
log (scale)
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
194
198
98
10
1
2
3
M
status
log (scale)
12
185
196
55
42
10
4
1
5
3
2
M
Mstat
log (scale)
40
159
77
14
46
43
65
18
5
1
6
8
9
3
2
4
7
occ

Wine Quality Data (Irvine Repository)

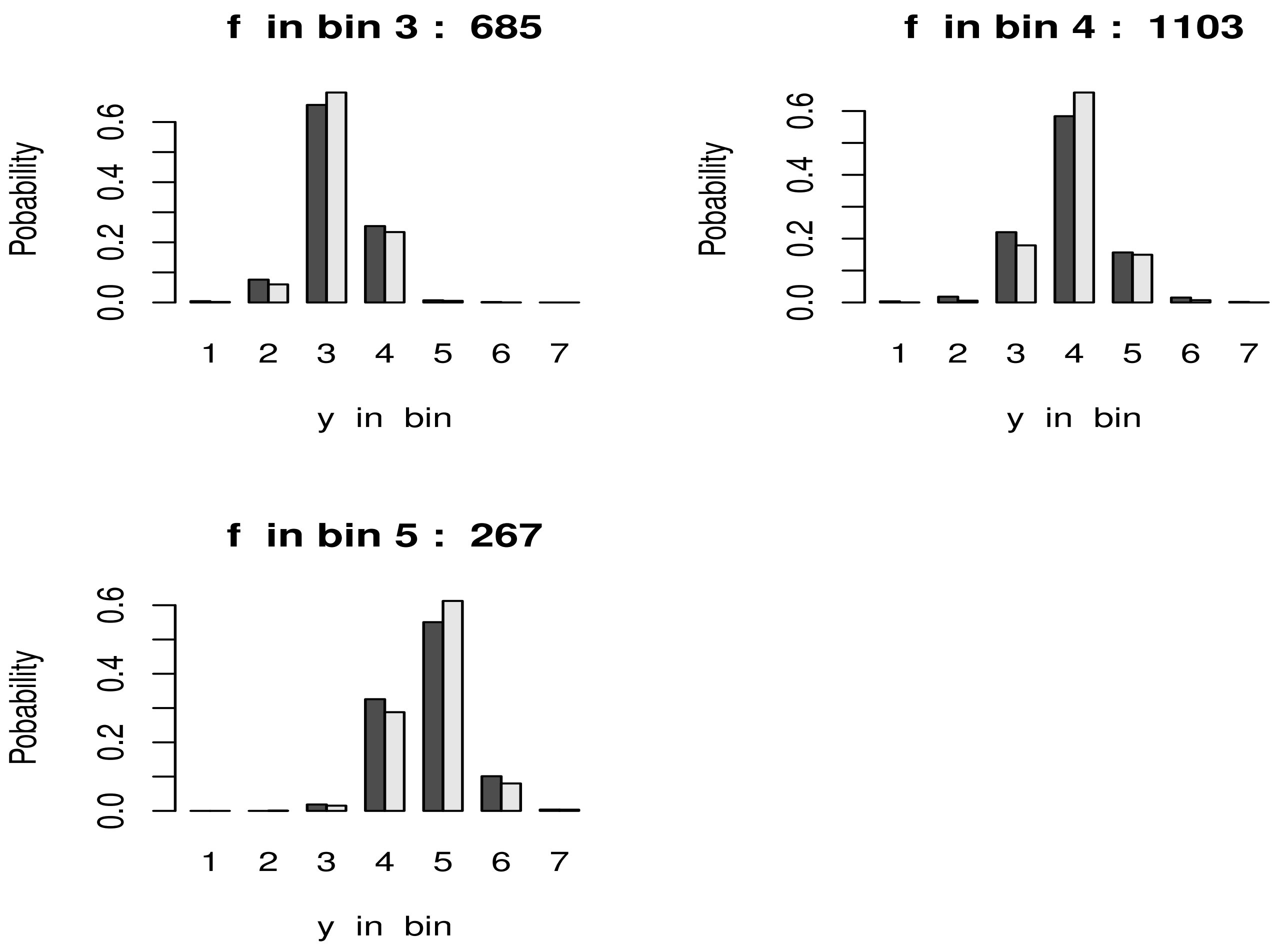
$N = 6497$ samples of Portuguese "Vinho Verde"

$\tilde{y} =$ Quality: integer $(1, 2, \cdots, 10)$

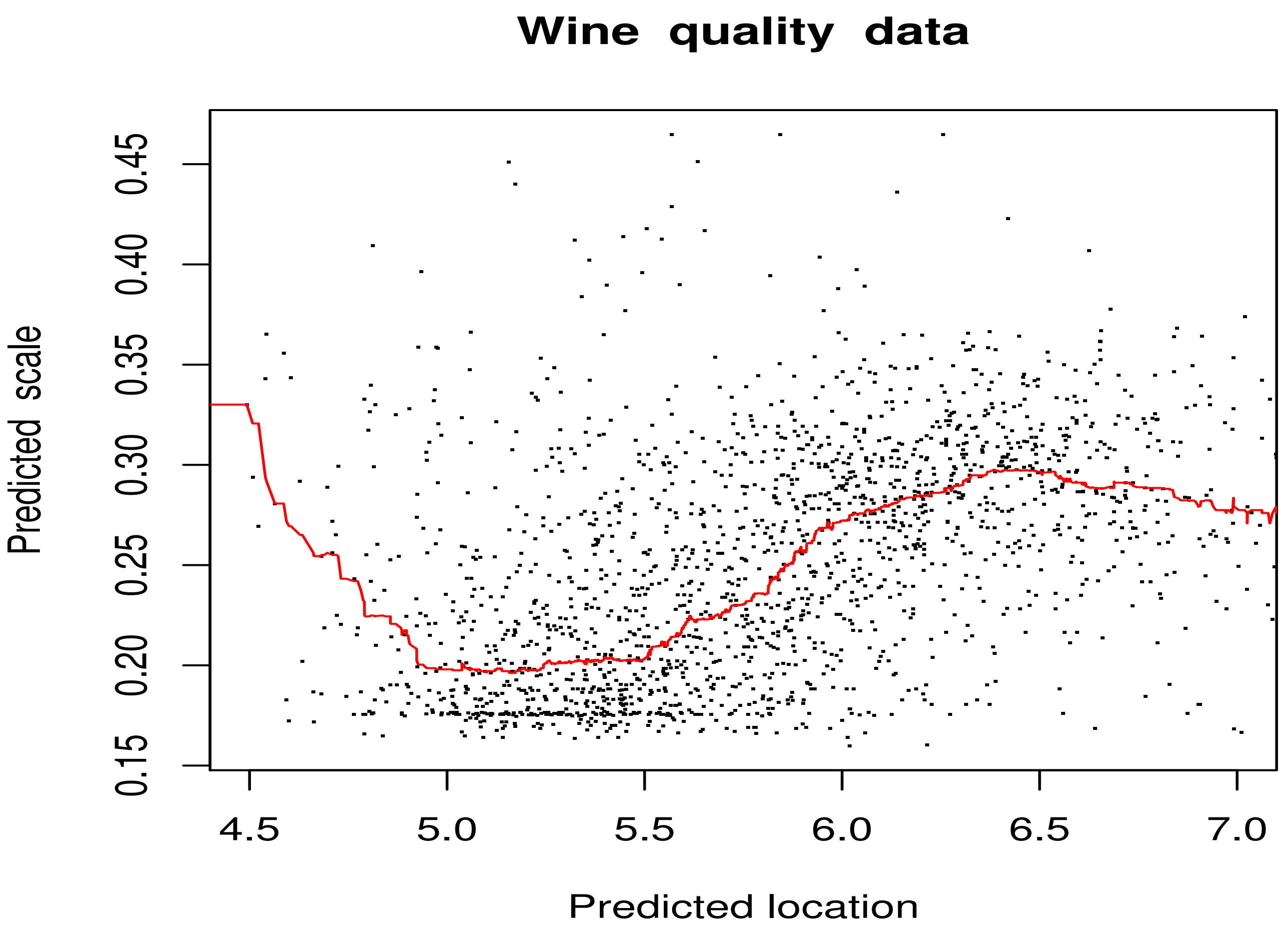
median of at least 3 expert evaluations

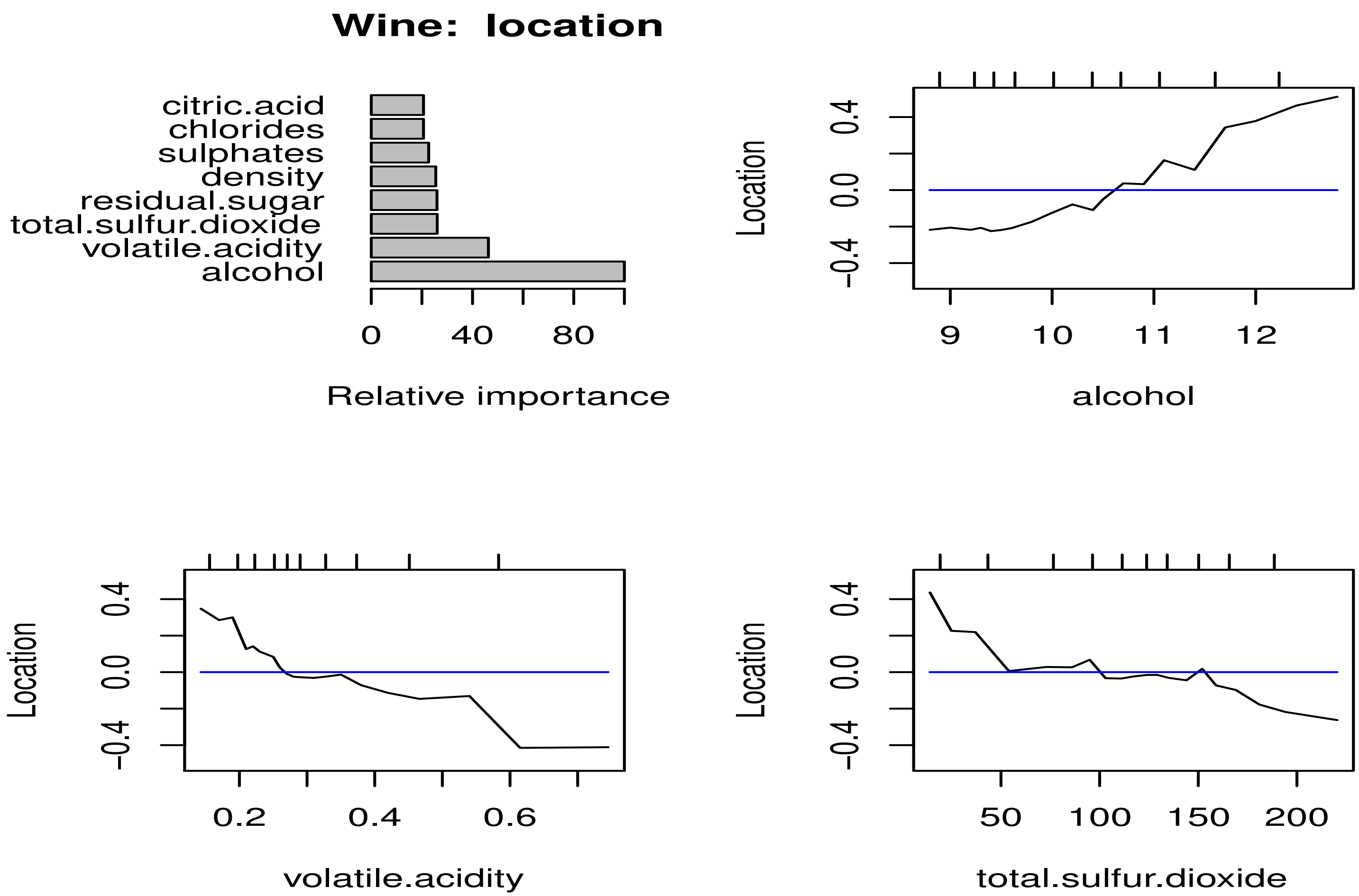
$\tilde{y} = k \Rightarrow y \in [k - 1/2,\ k + 1/2]$

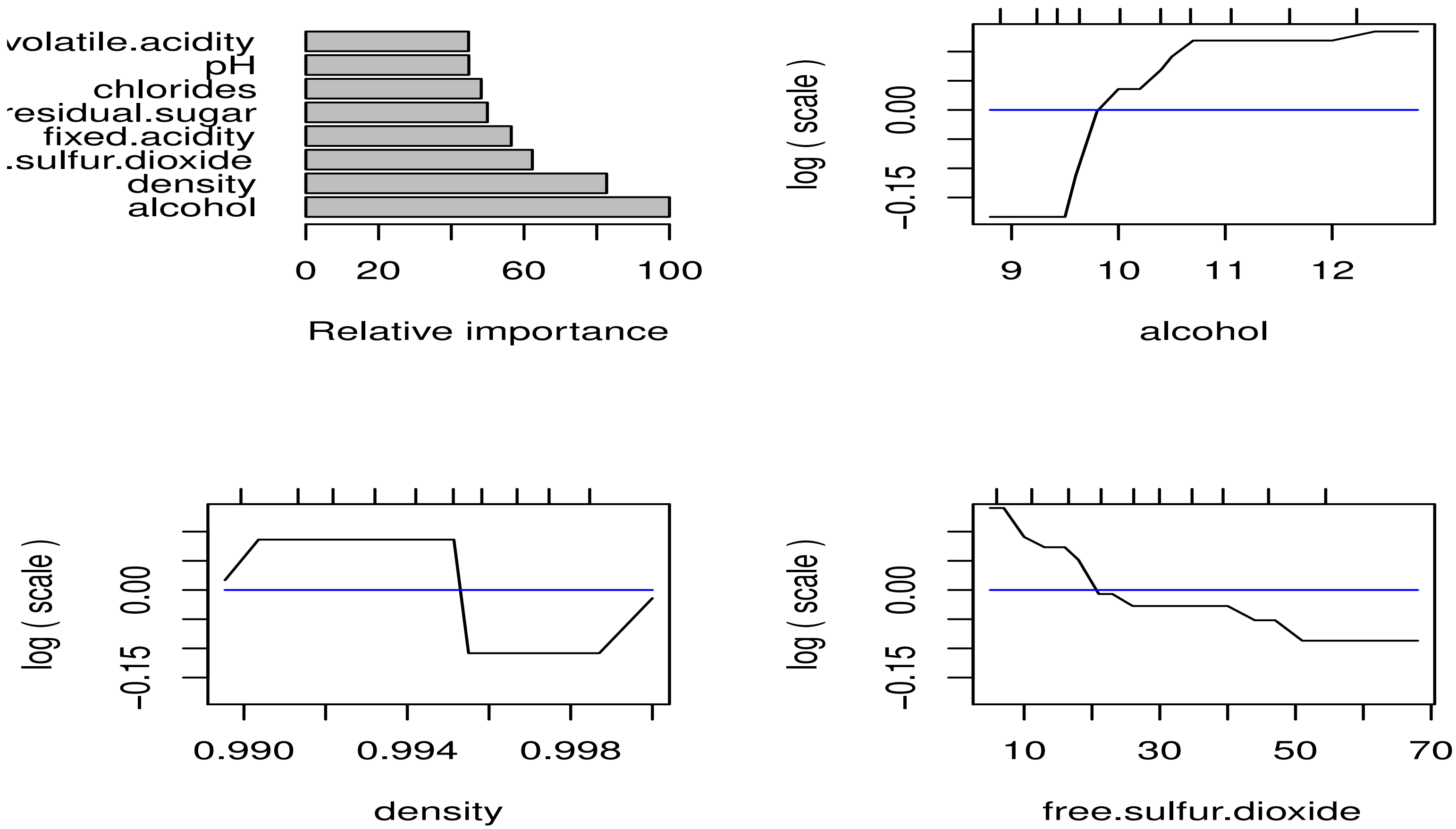
$$\mathbf{x} = (\text{Fixed acidity, Volatile acidity, Citric acid,}$$
$$\text{Residual sugar, Chlorides, Free sulfur dioxide}$$
$$\text{Total sulfur dioxide, Density, pH, Sulfates,}$$
$$\text{Alcohol})$$



Wine quality data


Wine quality data
Predicted scale
0.15
0.20
0.25
0.30
0.35
0.40
0.45
4.5
5.0
5.5
6.0
6.5
7.0
Predicted location


Wine: location
citric.acid
chlorides
sulphates
density
residual.sugar
total.sulfur.dioxide
volatile.acidity
alcohol
0
40
80
Relative importance
Location
−0.4
0.0
0.4
9
10
11
12
alcohol
Location
−0.4
0.0
0.4
0.2
0.4
0.6
volatile.acidity
Location
−0.4
0.0
0.4
50
100
150
200
total.sulfur.dioxide


volatile.acidity
pH
chlorides
residual.sugar
fixed.acidity
.sulfur.dioxide
density
alcohol
0
20
60
100
Relative importance
log (scale)
0.00
−0.15
9
10
11
12
alcohol
0.990
0.994
0.998
density
10
30
50
70
free.sulfur.dioxide

ORDERED MULTICLASS LOGISTIC REGRESSION

$$y_i \in \{C_1 < C_2, \cdots, C_{K-1} < C_K\}$$

Interval censored:

$\{a_i, b_i\} \rightarrow K$ disjoint intervals (bins):

$$\{b_0, b_1, \cdots b_K\} \;\; b_0 = -\infty, \;\; b_K = \infty$$

bins $\sim$ classes with separating boundaries

$\mathbf{b} = \{b_1, b_2, \cdots, b_{K-1}\}$ unknown

(overall location & scale arbitrary)

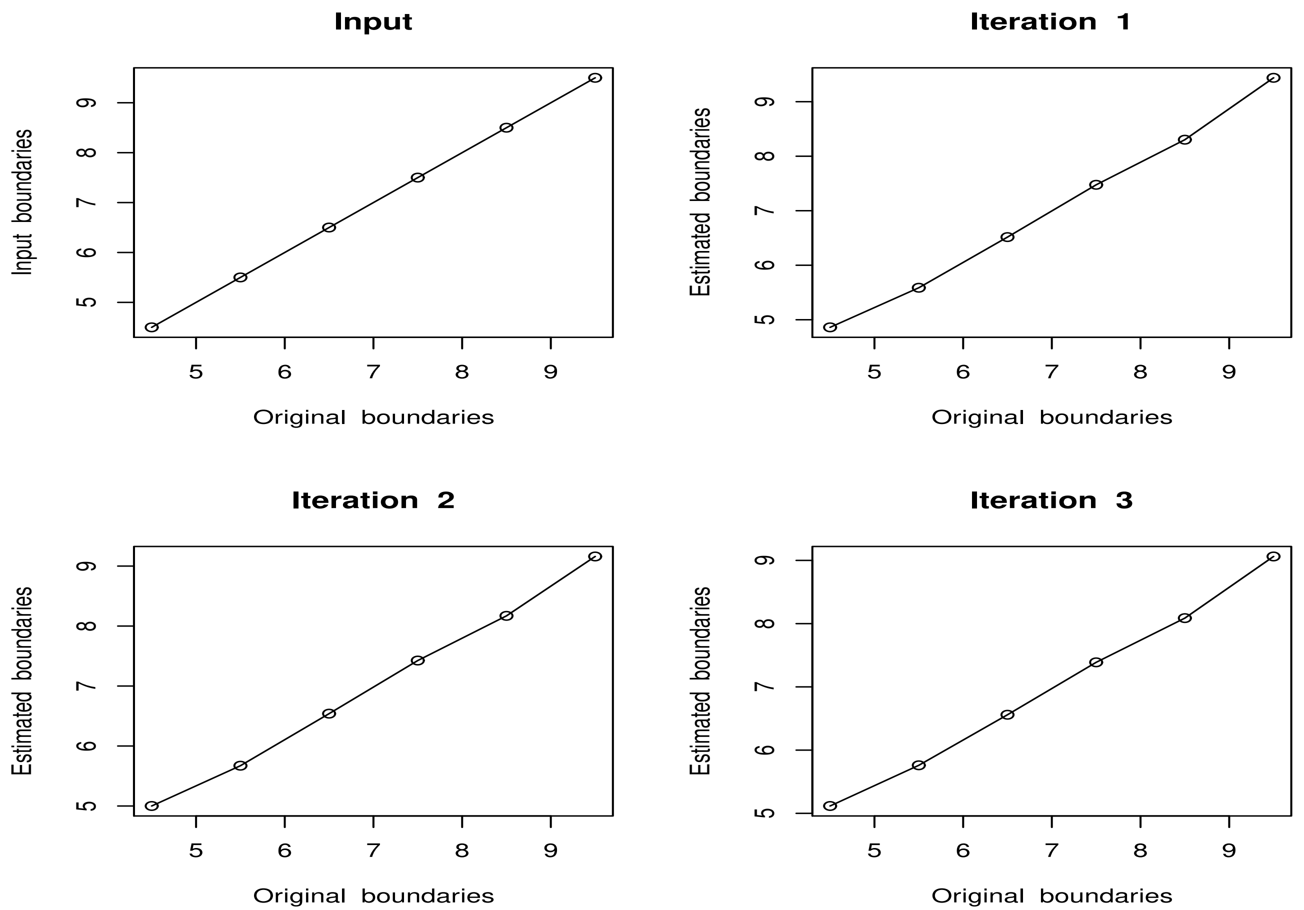
OPTIMAL SCALING (aka ACE)

$$[\hat{\mathbf{b}}, \hat{f}(\mathbf{x}), \hat{s}(\mathbf{x})] = \arg \min_{\mathbf{b},\,(f,s)\in F} \sum_{i=1}^{N} L[b_{k(i)-1}, b_{k(i)},\ f(\mathbf{x}_i), s(\mathbf{x}_i)]$$

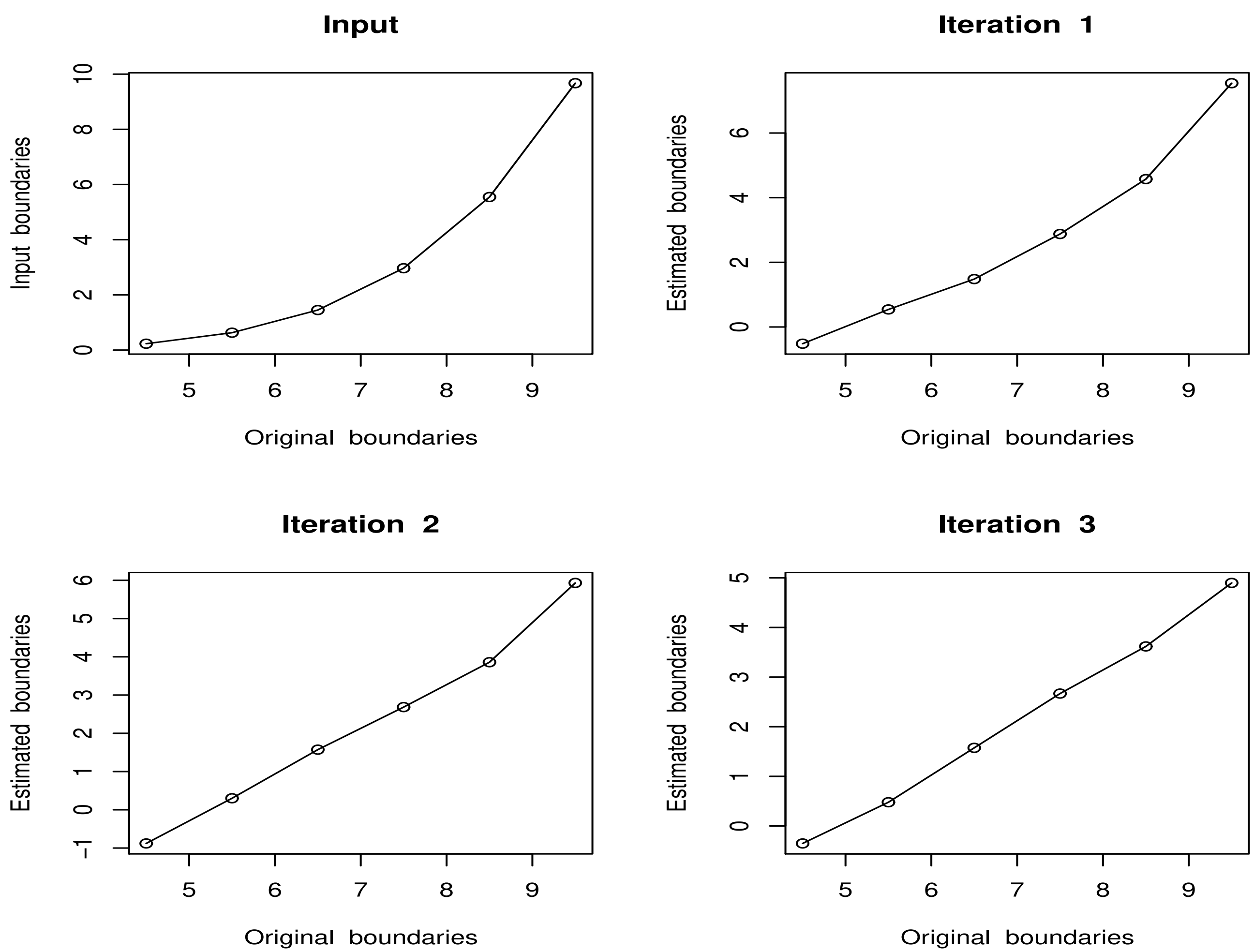
$$L(u, v,\ f, s) = -\log\left(\frac{1}{1+e^{(f-v)/s}} - \frac{1}{1+e^{(f-u)/s}}\right)$$

Alternating optimization:

$$\mathbf{b} : \sum_{i=1}^{N} \left(\frac{1}{1+e^{(\,f(\mathbf{x}_i)-b_k)/s(\mathbf{x}_i)}} - \sum_{j=1}^{k} I(y_i = c_j)\right) = 0$$



Wine quality data – optimal scaling



Wine quality data – optimal scaling

ASYMMETRIC ERRORS

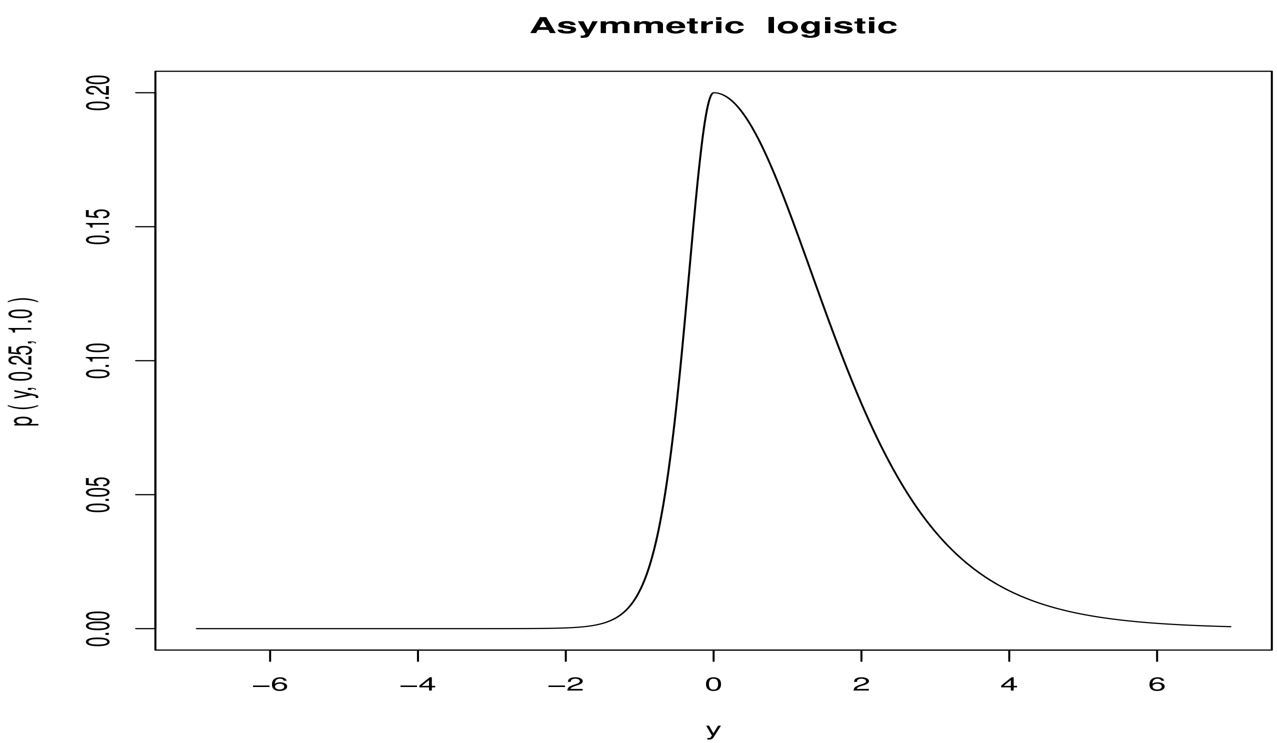
$$y \mid \mathbf{x} = f(\mathbf{x}) + \begin{cases} s_l(\mathbf{x}) \cdot \varepsilon & \varepsilon \leq 0 \\ s_u(\mathbf{x}) \cdot \varepsilon & \varepsilon > 0 \end{cases}$$

$f(\mathbf{x}) =$ mode at $\mathbf{x}$

$s_l(\mathbf{x}) =$ lower scale at $\mathbf{x}$

$s_u(\mathbf{x}) =$ upper scale at $\mathbf{x}$

$\varepsilon \sim$ standard logistic distribution



$$f = 0, \;\; s_l = 1/4, \;\; s_u = 1$$

EXERCISE (no censoring)

$$[\hat{f}(\mathbf{x}), \hat{s}_l(\mathbf{x}), \hat{s}_u(\mathbf{x})]$$

$$= \text{arg min}_{(f, s_l, s_u) \in F} \sum_{i=1}^{N} L[y_i, f(\mathbf{x}_i), s_l(\mathbf{x}_i), s_u(\mathbf{x}_i)]$$

$$L[y, f, s_l, s_u] =$$

$$L[y, f, s_l] \cdot I\,[\,y - f\, \leq 0] + L[y, f, s_u] \cdot I\,[\,y - f\, > 0]$$

$$L(y, f, s) = \log(s) + (y - f)/s + 2\log(1 + e^{-(y-f)/s})$$

Iterative gradient boosting

ASYMMETRIC DIAGNOSTICS

(1) $median\,[\,y \mid f(\mathbf{x}), s_l(\mathbf{x}), s_u(\mathbf{x})] = f(\mathbf{x})$

$$+ \begin{cases} s_u(\mathbf{x}) \log\left(\frac{3s_u(\mathbf{x}) - s_l(\mathbf{x})}{s_u(\mathbf{x}) + s_l(\mathbf{x})}\right) & s_l(\mathbf{x}) \le s_u(\mathbf{x}) \\ -s_l(\mathbf{x}) \log\left(\frac{3s_l(\mathbf{x}) - s_u(\mathbf{x})}{s_u(\mathbf{x}) + s_l(\mathbf{x})}\right) & s_l(\mathbf{x}) > s_u(\mathbf{x}) \end{cases}$$

(2) $median_{\,y \le f(\mathbf{x})}[\,|\,y - f(\mathbf{x})\,|\;|\,s_l(\mathbf{x})]/\log(3) = s_l(\mathbf{x})$

(3) $median_{\,y > f(\mathbf{x})}[\,|\,y - f(\mathbf{x})\,|\;|\,s_u(\mathbf{x})]/\log(3) = s_u(\mathbf{x})$

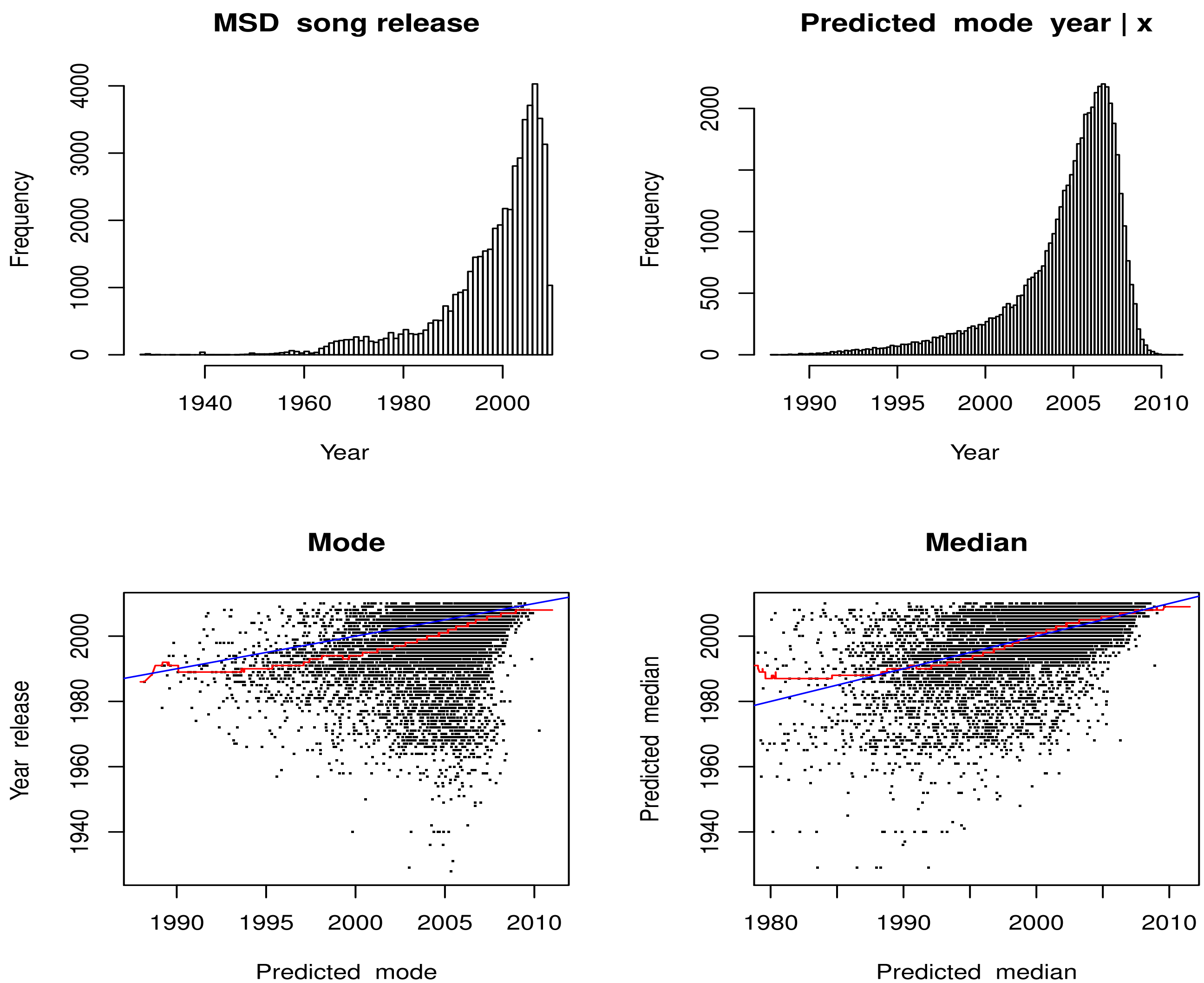
(1/2) – Million Song Dataset (Irvine Repository)

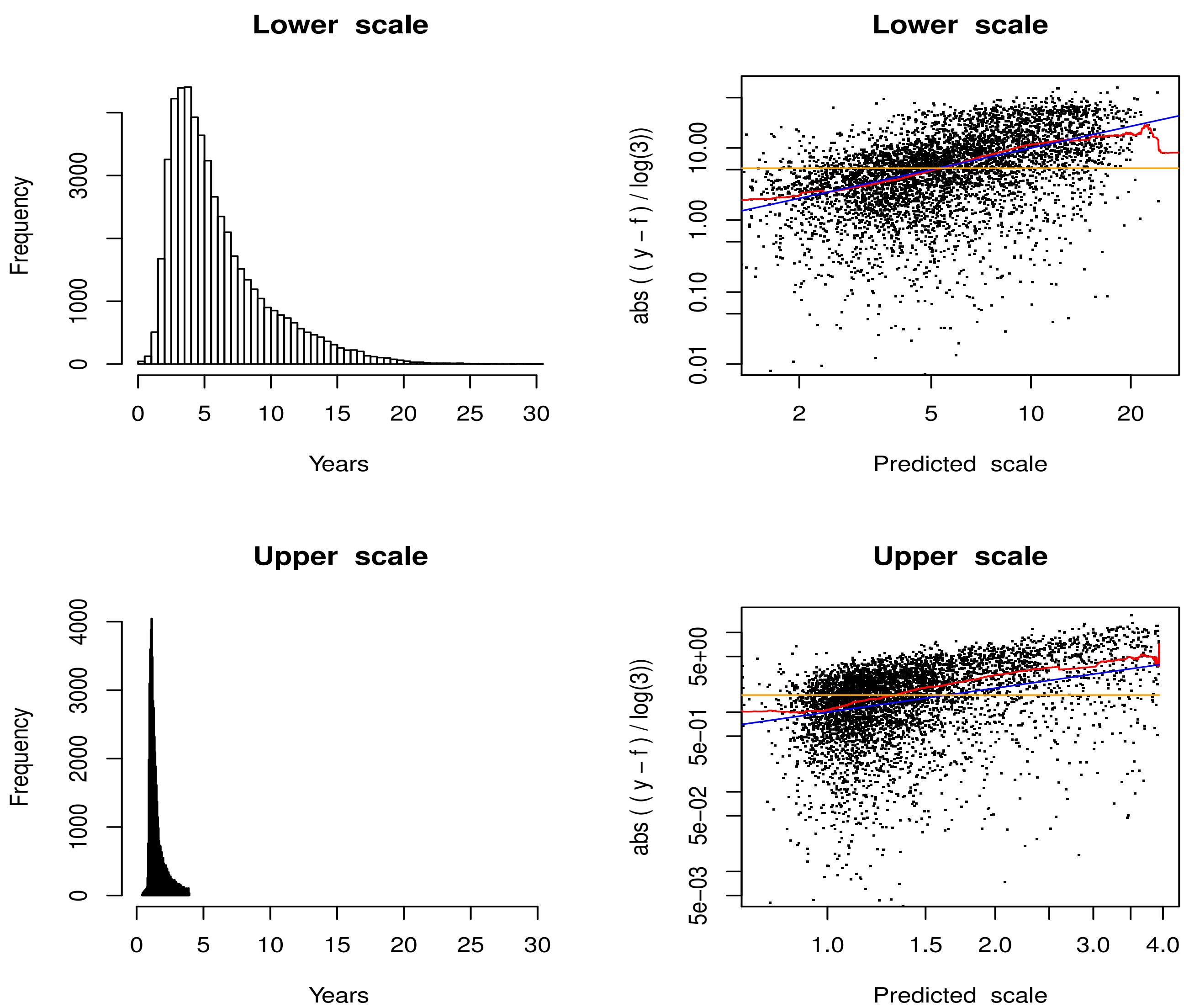
$N = 515345$ songs (463715 train, 51630 test)

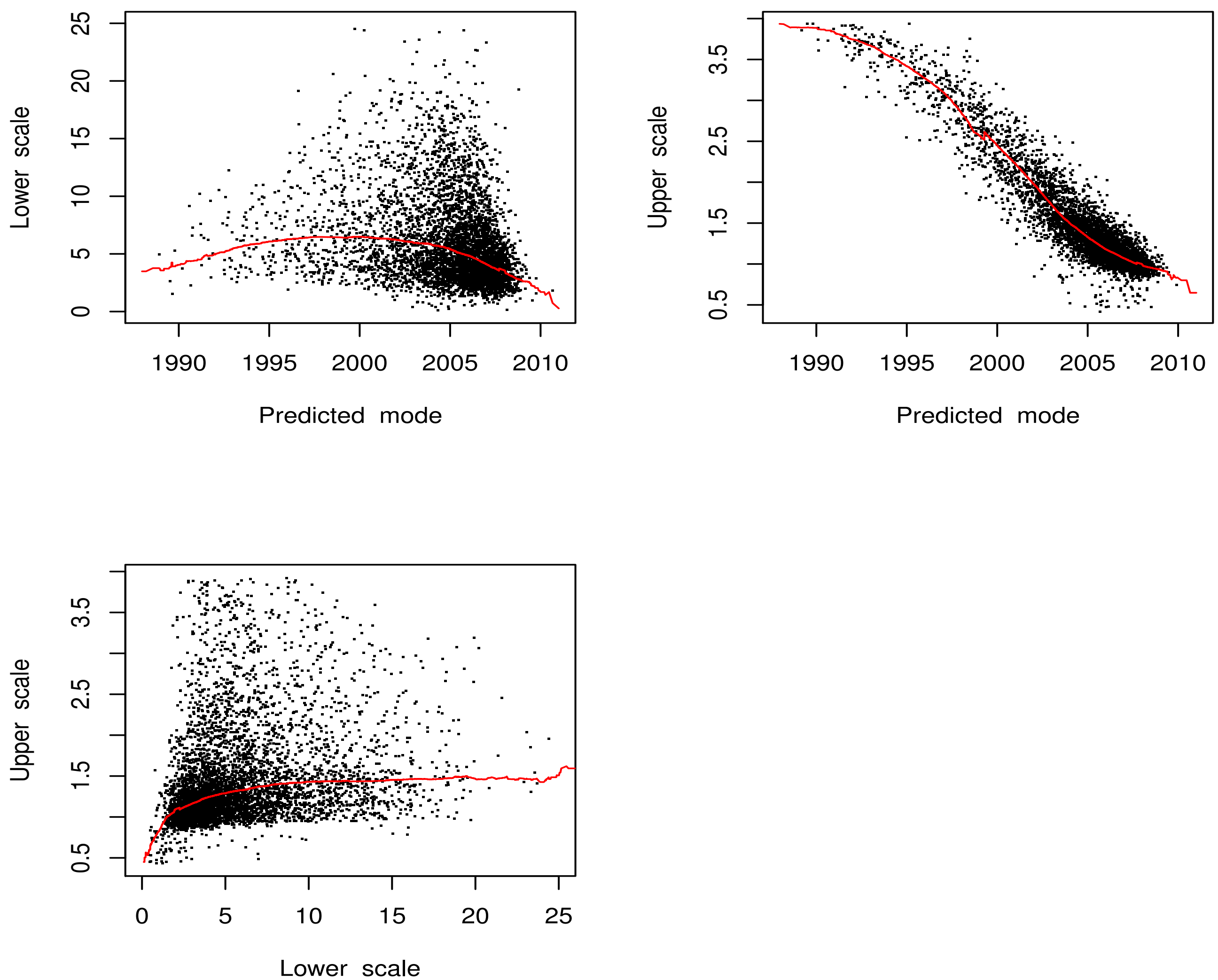
$y =$ year released (1922 – 2011)

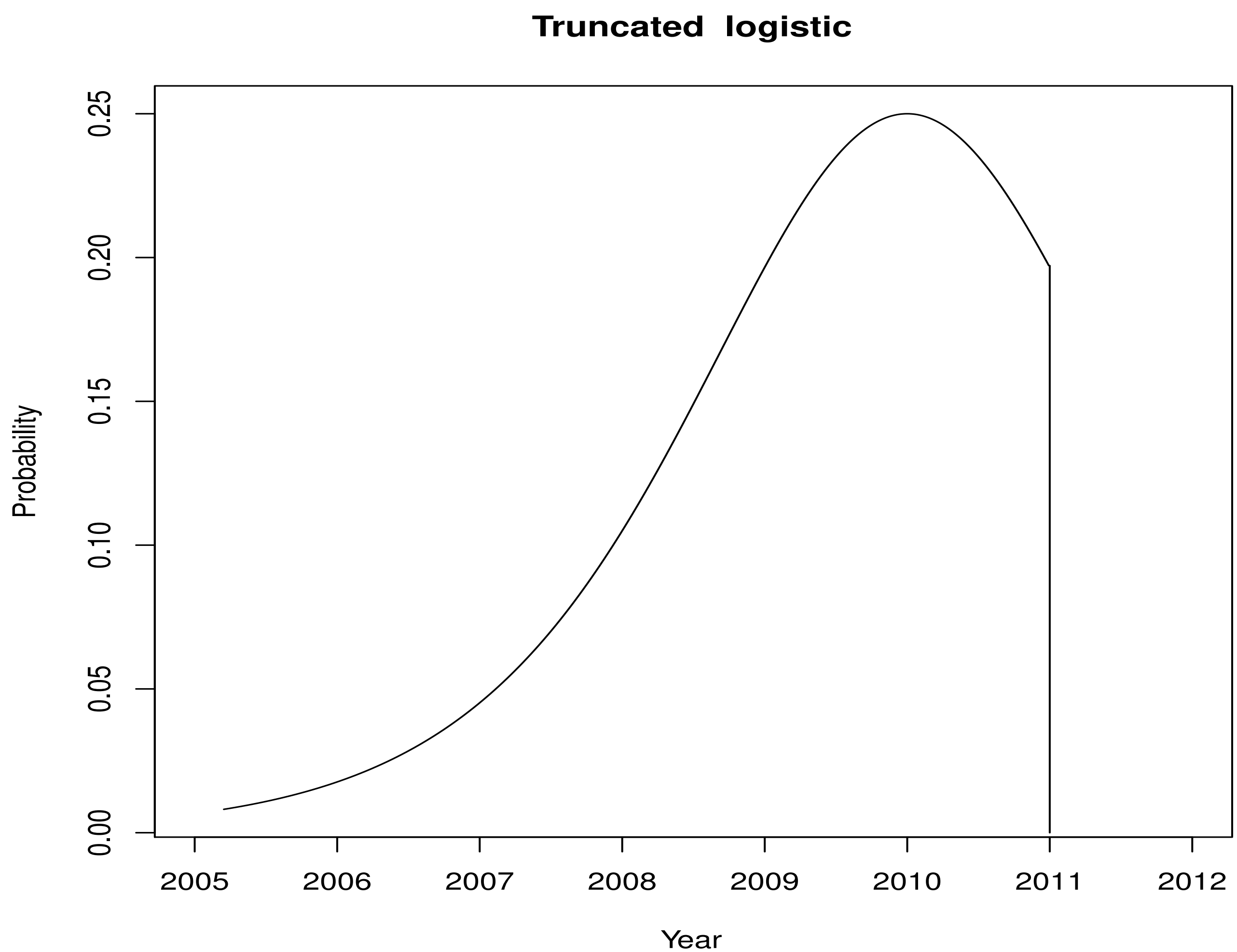
$\mathbf{x} = 90$ attributes (Echo Nest API):

12 timbre average, $78 =$ timbre covariance


MSD song release
Frequency
Year
Predicted mode year | x
Frequency
Year
Mode
Year release
Predicted mode
Median
Predicted median
Predicted median


Lower scale
Frequency
0
1000
3000
0
5
10
15
20
25
30
Years
Lower scale
abs ((y − f) / log(3))
0.01
0.10
1.00
10.00
2
5
10
20
Predicted scale
Upper scale
Frequency
0
1000
2000
3000
4000
0
5
10
15
20
25
30
Years
Upper scale
abs ((y − f) / log(3))
5e−03
5e−02
5e−01
5e+00
1.0
1.5
2.0
3.0
4.0
Predicted scale


Lower scale
Predicted mode
Upper scale
Predicted mode
Upper scale
Lower scale



Truncated logistic
Probability
0.00
0.05
0.10
0.15
0.20
0.25
2005
2006
2007
2008
2009
2010
2011
2012
Year

Predict $y \mid \mathbf{x}$, settle for $p(y \mid \mathbf{x})$

1. ROBUST (**log**istic) estimate: loc $f(\mathbf{x})$ & scale $s(\mathbf{x})$
2. General censoring: $a_i \leq y_i \leq b_i$ (ubiquitous)
3. Graphical diagnostics
4. Ordered multi–class classification
5. Asymmetric $p(y \mid \mathbf{x})$: $f(\mathbf{x})$, $s_l(\mathbf{x})$, $s_u(\mathbf{x})$

REFERENCES

Gradient boosting: Ann. Statist, **29**. 1189 – 1232 (2001)

Optimal scaling:

ACE: J. Amer. Statist. Assoc. **80**. 580 – 598 (1985)

GIFI: *Nonlinear Multivariate Analysis*. Wiley, N.Y. (1990)

Slides: http://statweb.stanford.edu/~jhf/talks/kdd.pdf