Robustness properties of support vector machines and related methods

Andreas Christmann
University of Dortmund, Dept. Statistics, Germany

Ingo Steinwart
Los Alamos National Laboratory, USA

Talk: PASCAL workshop, EURANDOM, The Netherlands, October 05, 2005

Contents:

1. Convex Risk Minimization

- $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} \subseteq \mathbb{R}$, closed or open, $\mathcal{X} \neq \varnothing$, $\mathcal{Y} \neq \varnothing$
- Classification: $\mathcal{Y} = \{-1, +1\}$
- Regression: $\mathcal{Y} \subseteq \mathbb{R}$
- Sample: $S = (z_1, \ldots, z_n)$, $z_i := (x_i, y_i) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, $1 \leq i \leq n$, $n \in \mathbb{N}$
- (X_i, Y_i) i.i.d. $\sim \mathrm{P} \in \mathcal{M}^1$, P unknown
- Loss function: $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$, $L(y_i, f(x_i))$, $L(y_i, f(x_i) + b)$
- Goals:
  - ⋆ estimation $\hat{f} = T(\mathrm{P}_n)$
  - ⋆ prediction $x_{new} \mapsto \hat{f}(x_{new})$

Kernels

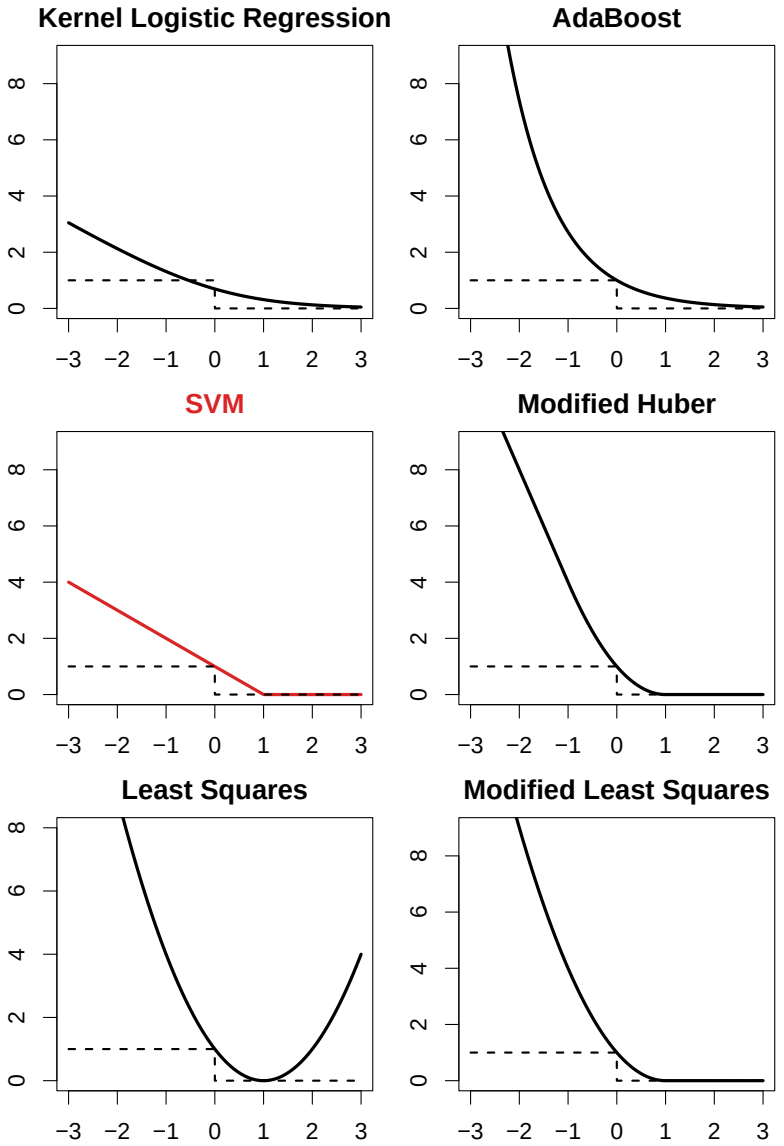
- kernel: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and ex. Hilbert space $\mathcal{H}$ and $\Phi : \mathcal{X} \to \mathcal{H}$: $k(x, x') = \langle \Phi(x), \Phi(x') \rangle, x, x' \in \mathcal{X}$
- reproducing kernel: $k(x, \cdot) = \Phi(x) \in \mathcal{H}$, $f(x) = \langle f, \Phi(x) \rangle$, $f \in \mathcal{H}$, $x \in \mathcal{X}$
- Reproducing Kernel Hilbert Space (RKHS): closure of $\{\sum_{i=1}^{n} \alpha_i k(x_i, \cdot); n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}\}$ w.r.t. $\left\langle \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \sum_{j=1}^{m} \beta_j k(x'_j, \cdot) \right\rangle := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j)$
- $k \Rightarrow$ RKHS unique. RKHS $\Rightarrow$ unique reprod. kernel k
- k continuous: $k(x, \cdot) : \mathcal{X} \to \mathbb{R}$ continuous, $x \in \mathcal{X}$
- k bounded: $||k||_\infty := \sqrt{\sup_{x \in \mathcal{X}} k(x, x)}] < \infty$
- k universal: k continuous (on compact metric space $\mathcal{X}$), $\mathcal{H}$ dense in $C(\mathcal{X})$
- RBF: $k(x, x') = \exp(-\gamma ||x - x'||_2^2)$, $\gamma > 0$, universal, if $\mathcal{X}$ compact

Convex Risk Minimization (Vapnik '98)

- risk: $\mathcal{R}_{L,\mathrm{P}}(f) := \mathrm{E}_{\mathrm{P}} L(Y, f(X))$
- reg. risk: $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) := \mathcal{R}_{L,\mathrm{P}}(f) + \lambda ||f||^2_{\mathcal{H}}$, $\lambda \in (0, \infty)$ fixed
- $\mathrm{P}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$
- reg. emp. risk: $\hat{f} = f_{\mathrm{P}_n,\lambda} = \arg\min_{f \in \mathcal{H}} \mathcal{R}^{reg}_{L,\mathrm{P}_n,\lambda}(f)$
  L convex, $\lambda > 0$, $\mathcal{H}$ Hilbert space (RKHS) with reprod. kernel k
- $f_{\mathrm{P}_n,\lambda}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$. If $\alpha_i \neq 0$: x_i is support vector.
- $T : \mathcal{M}^1(\mathcal{X} \times \mathcal{Y}) \mapsto \mathcal{H}$, $T(\mathrm{P}) := f_{\mathrm{P},\lambda} = \arg\min_{f \in \mathcal{H}} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)$ (1)
- analogously: $(f_{\mathrm{P},\lambda}\,,\, b_{\mathrm{P},\lambda})$ for $(f, b) \in \mathcal{H} \times \mathbb{R}$

Loss functions: classification

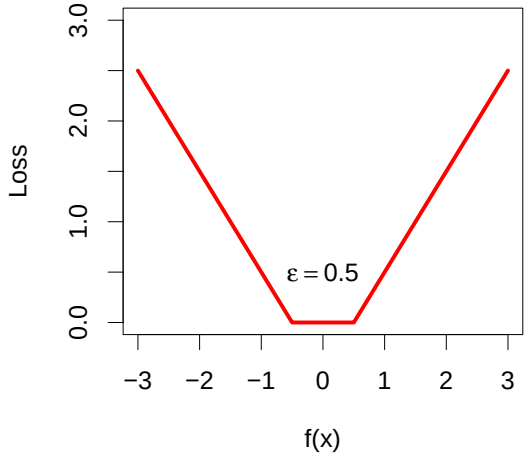
Method	L, $v = y(f(x) + b)$
Kernel Logistic Regression	$\ln(1 + \exp(-v))$
AdaBoost	$\exp(-v)$
Support Vector Machine	$\max(1 - v, 0)$
Modified Huber	$-4v$, if $v < -1$
	$\max(1 - v, 0)^2$, else
Least Squares	$(1 - v)^2$
Modified Least Squares	$\max(1 - v, 0)^2$

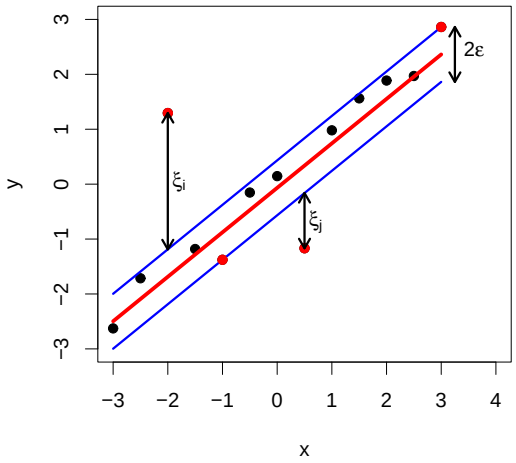


Vapnik '98,
Schölkopf & Smola '02,
Freund & Schapire '96,
Friedman, Hastie & Tibshirani '00,
Hastie, Tibshirani & Friedman '01,
Suykens et al. '02,
Zhang '04, ...

Loss functions: regression

- $\varepsilon-$Support Vector Regression ($\varepsilon-$SVR)
  $L_\varepsilon(y, f(x)) = \max\{0, |y - [f(x) + b]| - \varepsilon\}$





- Least Squares SVR (LS-SVR)
  $L(y, f(x)) = (y - f(x))^2$
- Logistic SVR
  $L(y, f(x)) = -\log(4\Lambda(r)[1 - \Lambda(r)]),\ r = y - f(x),\ \Lambda(r) := 1/[1 + e^{-r}].$

2. Robustness of CRM

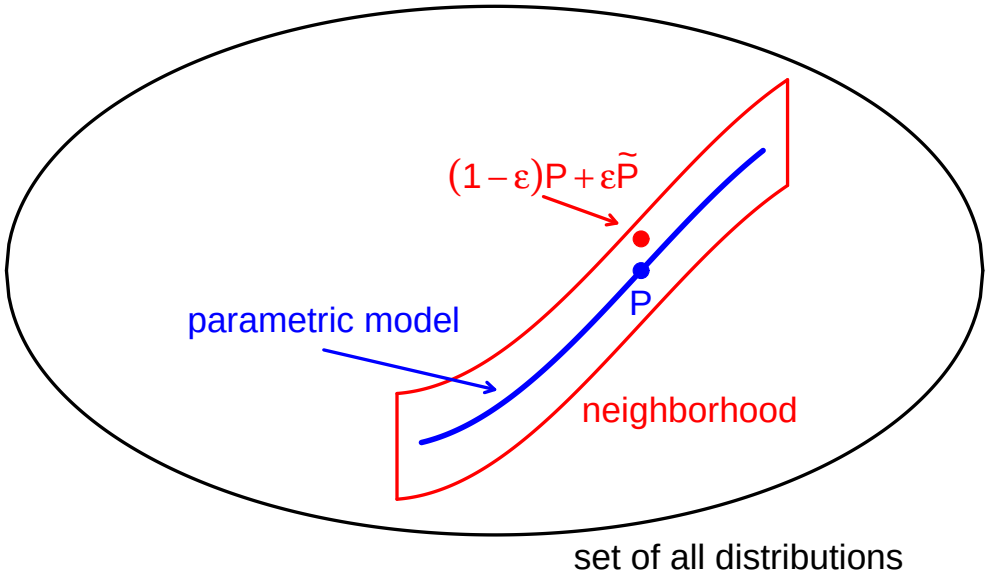
- (X_i, Y_i) i.i.d. $\sim$ P, P unknown
- What if *not* all $(X_i, Y_i) \sim$ P ?
- If $d(\mathrm{P}, \mathrm{Q}) < \delta \Rightarrow ||T(\mathrm{P}) - T(\mathrm{Q})||_{\mathcal{H}} < \varepsilon$?
- $T(\mathrm{P}_n) \approx T(\mathrm{P})$?
- Imprecise data ?
- Outliers ?
- Impact of L and k on $T(\mathrm{P}) = f_{\mathrm{P},\lambda}$?

Main question

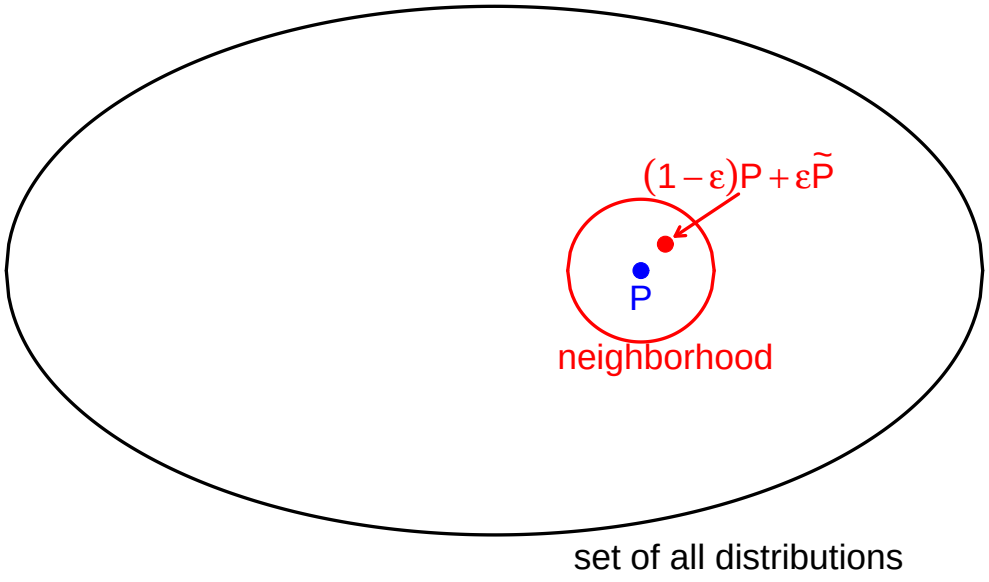
$$T((1-\varepsilon)\mathrm{P}+\varepsilon\tilde{\mathrm{P}}) \approx T(\mathrm{P}) \; ?$$

Here: $T(\mathrm{P}) = f_{\mathrm{P},\lambda} = \arg\min_{f\in\mathcal{H}} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)$ or $T(\mathrm{P}) = (f_{\mathrm{P},\lambda}, b_{\mathrm{P},\lambda})$

parametric / linear kernel



non-parametric / RBF kernel



Robustness concepts

⋆ Influence function (F. Hampel): should be bounded

$$IF(z;T,\mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{T\big((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z\big) - T(\mathrm{P})}{\varepsilon}$$

⋆ Sensitivity curve (J.W. Tukey)

$$SC_n(z;T_n) = n\big(T_n(z_1,\ldots,z_{n-1},z) - T_{n-1}(z_1,\ldots,z_{n-1})\big)$$

$$SC_n(z;T_n) = \frac{T\big((1-\varepsilon_n)\mathrm{P}_{n-1} + \varepsilon_n\delta_z\big) - T(\mathrm{P}_{n-1})}{\varepsilon_n}, \quad \varepsilon_n = \frac{1}{n}$$

⋆ Maxbias (P.J. Huber)

$$\mathrm{maxbias}(\varepsilon;T,\mathrm{P}) = \sup_{Q \in N_\varepsilon(\mathrm{P})} \|T(Q) - T(\mathrm{P})\|$$

$$N_\varepsilon(\mathrm{P}) = \{Q = (1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}};\ \tilde{\mathrm{P}} \in \mathcal{M}^1(\mathcal{X}\times\mathcal{Y}),\ 0 \le \varepsilon < \tfrac{1}{2}\}$$

Robustness concepts

⋆ Finite-sample breakdown point (Donoho & Huber)
measures the smallest contamination when the estimator is worthless

Let $S = \{(x_i, y_i), i = 1, \ldots, n\}$ be a data set with values in $\mathcal{X} \times \mathcal{Y}$. The finite-sample breakdown point of an estimator $T_n(S)$ is defined by

$$\varepsilon_n^*(T_n, S) = \min \left\{ \frac{m}{n} \, ; Bias(m; T_n, S) \quad \textit{is finite} \right\} ,$$

where

$$Bias(m; T_n, S) = \sup_{S'} \parallel T_n(S') - T_n(S) \parallel$$

and the supremum is over *all* possible samples S' that can be obtained by replacing *any* m of the original data points by *arbitrary* values in $\mathcal{X} \times \mathcal{Y}$.

Comparison

	M-estimator	Kernel based estimator
Regularisation	–	$\lambda \|\|f\|\|^2_{\mathcal{H}}$
Parameter space	$\mathbb{R}^d$	$\mathcal{H}$
function	$f(x) = x'\theta$	$f(x) = \langle\, f, \Phi(x)\,\rangle$
$\mathrm{IF}(z; T, \mathrm{P})$, z *fixed*	vector in $\mathbb{R}^d$	function in $\mathcal{H}$

Robustness: classification (Chr & Steinwart, '04)

Let $\mathcal{Y} = \{-1, +1\}$, $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ convex with $L'' > 0$ continuous, $\mathcal{H}$ RKHS of continuous kernel k.

(a) Let [$\mathcal{X} \subset \mathbb{R}^d$ open or bounded and k bounded] or [$\mathcal{X}$ compact].
Then: influence function of $T(\mathrm{P}) = f_{\mathrm{P},\lambda}$ exists for all $z = (z_x, z_y) \in \mathcal{X} \times \mathcal{Y}$ and

$$IF(z; T, \mathrm{P}) = -S^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{\mathrm{P},\lambda}) \,,$$

where $\frac{\partial G}{\partial \varepsilon}(0, f_{\mathrm{P},\lambda}) = -\mathrm{E}_{\mathrm{P}}[L'(Y, f_{\mathrm{P},\lambda}(X))\Phi(X)] + L'(z_y, f_{\mathrm{P},\lambda}(z_x)) \cdot \Phi(z_x)$
and $S := \frac{\partial G}{\partial \mathcal{H}}(0, f_{\mathrm{P},\lambda}) = 2\lambda \,\mathrm{id}_{\mathcal{H}} + \mathrm{E}_{\mathrm{P}} L''(Y, f_{\mathrm{P},\lambda}(X))\langle \Phi(X), . \rangle \Phi(X)$.

(b) Similar result for the case (f, b).

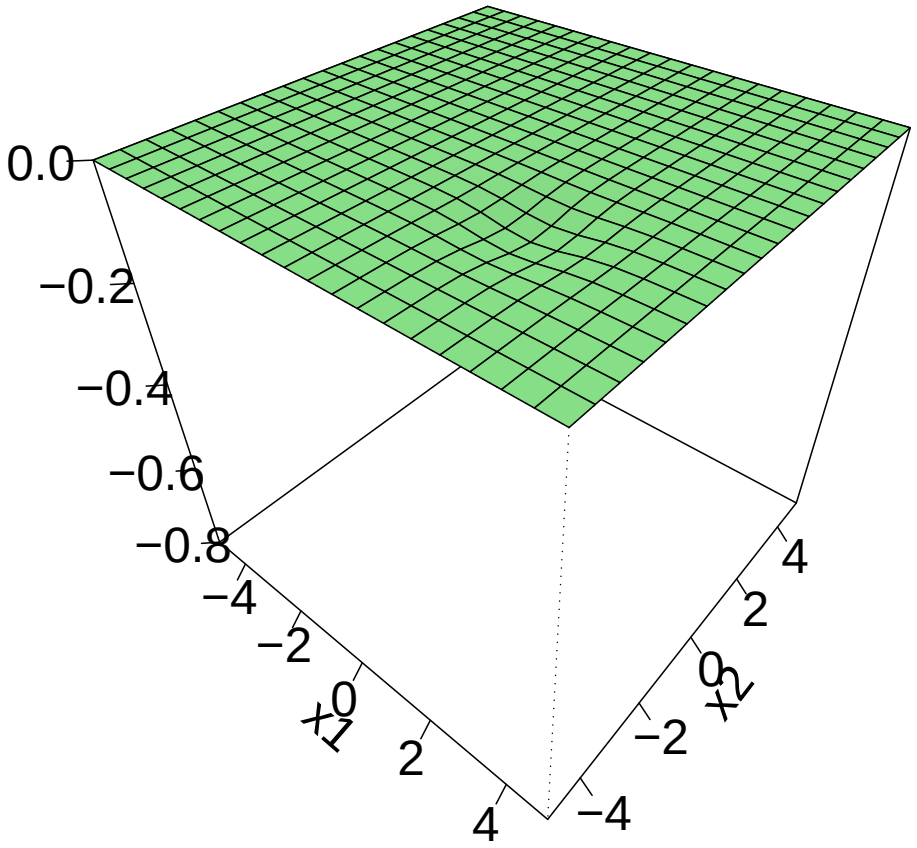
– – – – –

Mallows M-estimators for *linear* models: $\mathrm{IF}(z; T, \mathrm{P}) = M^{-1} \cdot \psi\left(z_y, z_x' T(\mathrm{P})\right) \cdot w(z_x) z_x$

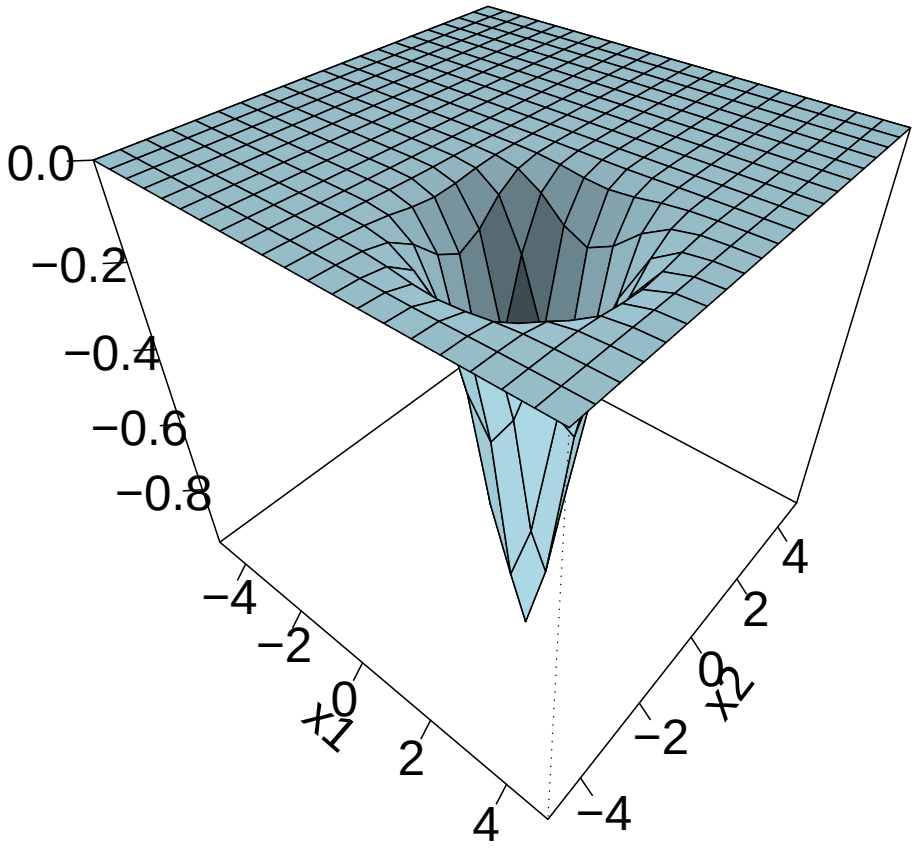
$\mathbf{L'(z_y, f_{P,\lambda}(z_x))\Phi(z_x)}$ for KLR with RBF-kernel

Contamination in $z_x = (2, -2)$, $\mathrm{P}(Y = +1 \mid X = z_x) = 0.982$

'probable' obs.: $z_y = +1$



'improbable' obs.: $z_y = -1$



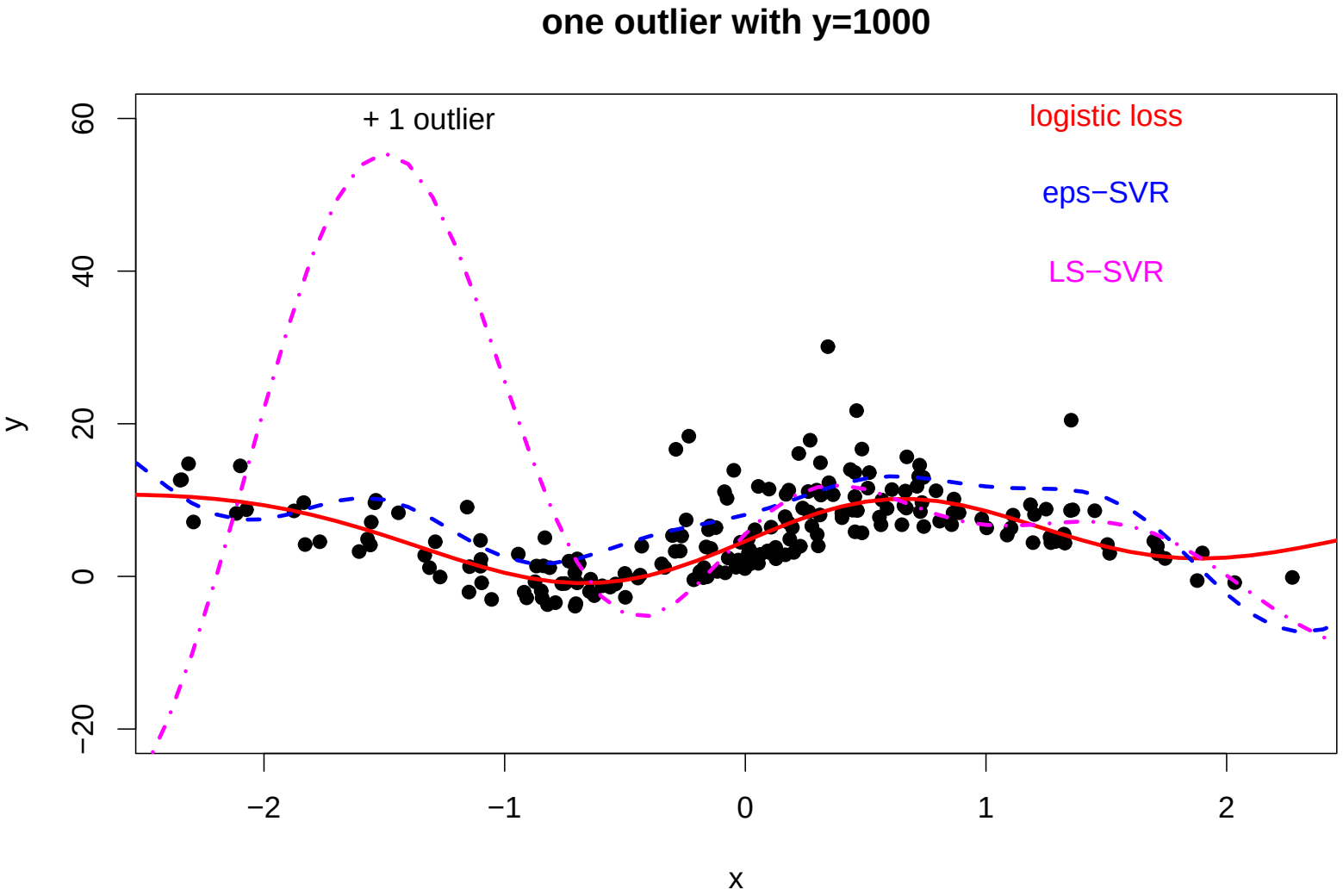
bounded & local impact

Further results (Chr & Steinwart, '04, '05)

- for classification:
  - difference quotient in definition of IF is uniformly bounded
  - special cases: SVM, KLR, AdaBoost, LS, Mod.LS, Mod.Huber
  - norm of SC_n and maxbias are uniformly bounded
- for regression:
  - $T(\mathrm{P}) = \arg\min_{f \in \mathcal{H}} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)$:
    - ⋆ existence, uniqueness and consistency (under tail-assumptions for P)
  - influence function of $T(\mathrm{P})$
    - ⋆ existence (under tail-assumptions for P)
    - ⋆ bounded, if L' and k bounded
    - ⋆ (universal) consistency $\leftrightsquigarrow$ robustness
    - ⋆ connections to stability

      → Bousquet & Elisseeff '02, Elisseeff, Evgeniou & Pontil '05, Mukherjee et al. '04, Poggio et al. '04

Regression examples: n=200 data points, skewness



Problem: SVM / CRM are 'non-robust posed problems'

- $f_{\mathrm{P},\lambda} = \arg\min_{f \in \mathcal{H}} \mathrm{E}_{\mathrm{P}} L(Y, f(X)) + \lambda ||f||_{\mathcal{H}}^2$
- Assume: $supp(\mathrm{P}_{Y|X=x})$ and $supp(\mathrm{P}_{Y-f(x)})$ are unbounded
- $\forall \varepsilon > 0 \quad \exists\, \mathrm{Q} \in \mathcal{M}_1$ with $d(\mathrm{P}, \mathrm{Q}) < \varepsilon$, but:
$$0 \leq \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) < \infty \text{ and } \mathcal{R}_{L,\mathrm{Q}}(f_{\mathrm{Q},\lambda}) = \infty \ .$$
  - inherent instability/non-robustness of the CRM problem itself due to E!
    $\Rightarrow$ other models: *Davies & Gather '05, Ann. Statist.*
  - unbounded, convex L: can not circumvent the problem (for *all* P)
  - bounded, non-convex L: numerical problems, multiple solutions
  - need for robust alternative of the CRM problem!
- $\hat{f} = f_{\mathrm{P}_n,\lambda} = \arg\min_{f \in \mathcal{H}} \mathrm{RobLoc}(\mathrm{P}_{L(Y,f(X))}) + \lambda ||f||_{\mathcal{H}}^2$
  First ideas: median ("regularized kernel LMS"), Hodges-Lehmann, ...
- $\hat{f} = f_{\mathrm{P}_n,\lambda} = \arg\min_{f \in \mathcal{H}} \mathrm{RobScale}(\mathrm{P}_{Y-f(X)}) + \lambda ||f||_{\mathcal{H}}^2$

3. Robust Learning from Bites (Chr '05)

Problem: algorithms for KLR are relatively slow

[and for many other CRM methods, robust estimators, S, MM, ...]

Simulated data: $\mathrm{P}(Y_i = +1 | X_i = x_i) = 1/[1 + \exp(-f(x_i))]$ and

$$f(x_i) = \sum_{j=1}^{8} x_{i,j} - x_{i,1}x_{i,2} - x_{i,2}x_{i,3} - x_{i,4}x_{i,5} - x_{i,1}x_{i,6}x_{i,7} \, .$$

n	CPU time	Cache [MB]	available Cache [MB]
2,000	4 sec	33	1000
10,000	1 min, 33 sec	787	1000
100,000	9 h, 56 min, 46 sec	1000	1000

myKLR: Keerthi et al. (2002, 2004), Rüping (2003)

Insurance data: $n \approx 4.6$ millions, would need several months . . .

RLB: Robust Learning from Bites

Step 1: Partition in bites.

Divide data set S by random in B disjoint subsets S_b of size $n_b \approx n/B$

Step 2: Model the bites.

Compute robust estimates T_{n_b} based on S_b, $b = 1, \ldots, B$.

Step 3: Aggregation and prediction based on median.

$T_{RLB,n,B} = \text{median}_{1 \le b \le B} T_{n_b}$ (componentwise or multivariate M-estimation)

$T_{RLB,n,B}(x_i) = \text{median}_{1 \le b \le B} T_{n_b}(x_i)$, $x_i \in S$

distribution-free $(1 - \alpha)$ confidence intervals for $T_{RLB,n,B}(x)$

based on order statistics, i.e. $[T_{RLB,n,(r:B)}(x), T_{RLB,n,(s:B)}(x)]$.

Step 3': ... based on mean . $T_{RLB,n,B} = \sum_1^B \frac{n_b}{n} T_{n_b}$.

Connections:

- remedian (Rousseeuw & Bassett '90)
- stochastic blockwise subsampling (Politis, Romano, Wolf '99)
- bagging (Breiman '96): sample B sets of n elements from S with replacement
- learning ensembles for bites (Chawla et al. '04): non-robust

RLB: computational aspects

performance, distributed computing, scalability

criterion	**original algorithm**	**RLB with B bites**
computation time, k CPUs	$O(g_1(n,d))$	$O(\llbracket B/k \rrbracket \cdot g_1(n/B,d))$
memory space, k CPUs	$O(g_2(n,d))$	$O(k \cdot g_2(n/B,d))$
hard disk space, k CPUs	$O(g_3(n,d))$	$O(k \cdot g_3(n/B,d))$

g_1, g_2, g_3 are positive functions

Insurance data: $n \approx 4.6$ millions, $B = 17$, PC with 2 CPUs: ≈ 1 week
still long, but without RLB: several months

RLB for kernel methods

Assume

- original estimator $\hat{f}_n$ is a kernel estimator defined in (1)
- B is fixed.

Then the RLB estimator based on the mean is itself a kernel estimator:

$$\begin{aligned}\hat{f}_{RLB,n,B}(x) &= \sum_{i=1}^{n} \alpha_{i,RLB}\, k(x, x_i) \\ &= \sum_{i \in SV(\mathcal{S}_1) \cup \ldots \cup SV(\mathcal{S}_B)} \alpha_{i,RLB}\, k(x, x_i), \quad x \in \mathcal{X},\end{aligned}$$

where $\alpha_{i,RLB} = \frac{1}{B} \sum_{b=1}^{B} \alpha_{i,b}\,,\ i \in \mathcal{S}$.
If all support vectors in $\mathcal{S}_1, \ldots, \mathcal{S}_B$ are different, we have $\alpha_{i,RLB} = \frac{\alpha_{i,b}}{B}$.

RLB for kernel methods: number of support vectors

RLB estimator based on mean:

- The number of support vectors, i.e. $\alpha_{i,RLB} \neq 0$, of the RLB estimator is given by
$$\#\left\{SV(\mathcal{S}_1) \cup \ldots \cup SV(\mathcal{S}_B)\right\} .$$
- Assume binary classification problems, i.e. $\mathcal{Y} = \{-1, +1\}$, B fixed, $n/B \equiv n_b$, and $[\ldots]$. Then
$$\Pr^{*n}\left(\mathcal{S}_1 \cup \ldots \cup \mathcal{S}_B \in (\mathcal{X} \times \mathcal{Y})^n; \#\mathrm{SV}(\hat{f}_{RLB,n,B}) \geq \sum_{b=1}^{B} (S_{L,\mathrm{P}} - \varepsilon) n_b\right) \to 1 .$$

"With probability tending to 1 the fraction of support vectors of RLB estimator is essentially greater than the average of the Bayes risks for the bites."

RLB for kernel methods: $L-$risk consistency

Assume $B \geq 1$ is fixed with $n/n_b \to B$, if $n \to \infty$,

- loss function L convex
- kernel k is universal, e.g. RBF kernel
- $L-$risk consistent kernel estimator $\hat{f}_{n,\lambda_n}$

$\Rightarrow$ L-risk consistency of RLB estimator $\hat{f}_{RLB,n,\lambda_n,B} := \sum_{b=1}^{B} \frac{n_b}{n} \hat{f}_{n_b,\lambda_{n_b}}$:

$$\mathcal{R}_{L,\mathrm{P}}(\hat{f}_{RLB,n,\lambda_n,B}) \xrightarrow{\mathrm{P}} \mathcal{R}_{L,P} \ .$$

"The RLB estimator is able to learn !"

Proof of $L-$risk consistency

$\hat{f}_{RLB,n,\lambda_n,B}(x)$ based on mean is convex combination of $\hat{f}_{n_b,\lambda_{n_b}}(x)$, $b = 1,\ldots,B$, because $n_b/n \in [0,1]$ and $\sum_{b=1}^{B}(n_b/n) = 1$.

$$
\begin{aligned}
0 &\leq \mathcal{R}_{L,\mathrm{P}}(\hat{f}_{RLB,n,\lambda_n,B}) - \mathcal{R}_{L,\mathrm{P}} \\
&= \int L\left(y, \sum_{b=1}^{B} \frac{n_b}{n} \hat{f}_{n_b,\lambda_{n_b}}(x)\right) \, d\mathrm{P}(x,y) - \mathcal{R}_{L,\mathrm{P}} \\
&\leq \int \sum_{b=1}^{B} \frac{n_b}{n} L\left(y, \hat{f}_{n_b,\lambda_{n_b}}(x)\right) \, d\mathrm{P}(x,y) - \mathcal{R}_{L,\mathrm{P}} \qquad (1.1) \\
&= \sum_{b=1}^{B} \frac{n_b}{n} \left[\int L\left(y, \hat{f}_{n_b,\lambda_{n_b}}(x)\right) \, d\mathrm{P}(x,y) - \mathcal{R}_{L,\mathrm{P}}\right] \quad \xrightarrow{\mathrm{P}} \quad 0, \qquad (1.2)
\end{aligned}
$$

if $\min_{1\leq b\leq B} n_b \to \infty$.

Consistency of RLB

RLB based on median: (B fixed)

- If $T_n \xrightarrow{\mathrm{P}} T(\mathrm{P})$, $n \to \infty$, and if $(n/n_b) \to B$, then

$$T_{RLB,n,B} \xrightarrow{\mathrm{P}} T(\mathrm{P}) \qquad (\text{same for } \xrightarrow{\text{wp1}}).$$

RLB based on mean: (B fixed)

- If $\mathrm{E}(T_b) = \mathrm{E}(T_n)$ for all $b \in \{1, \ldots, B\}$, then

$$\mathrm{E}(T_n) = \mathrm{E}(T_{RLB,n,B}).$$

- If $T_n \xrightarrow{\mathrm{P}} T(\mathrm{P})$, $n \to \infty$, and if $(n/n_b) \to B$, then

$$T_{RLB,n,B} \xrightarrow{\mathrm{P}} T(\mathrm{P}) \qquad (\text{same for } \xrightarrow{\text{wp1}}).$$

- If $n_b^{1/2}(T_{n_b} - T(\mathrm{P})) \xRightarrow{\mathcal{D}} N(0, \Sigma)$, $\Sigma \in \mathbb{R}_{p.d.}^{d \times d}$, and $(n/n_b) \to B$. Then

$$n^{1/2}(T_{RLB,n,B} - T(\mathrm{P})) \xRightarrow{\mathcal{D}} N(0, \Sigma),\ n \to \infty.$$

Influence function of RLB based on the mean

Assume

- original estimator $T_n(S)$ has representation $T(\mathrm{P}_n)$, where P_n is the empirical distribution of the sample S,
- influence function of the map $T(\mathrm{P})$ exists for P.

Then:

$$\mathrm{IF}(z; T_{RLB,n,B}, \mathrm{P}) = \mathrm{IF}(z; T(\mathrm{P}), \mathrm{P}) \ .$$

True for many kernel based methods and classical robust methods.

Finite-sample breakdown point of RLB

Consider RLB with B bites where $n_b \equiv n/B$. Denote the finite sample breakdown point of the estimator $T_b(S_b)$ for bite b by $\varepsilon^*_{n_b}(T_b; \mathcal{S}_b)$ and denote the finite sample breakdown point of the estimator $\hat{\mu} = \hat{\mu}(T_1(S_1), \ldots, T_B(S_B))$ in the aggregation step by $\varepsilon^*_B(\hat{\mu})$. Then

$$\varepsilon^*_{RLB,n,B} = \varepsilon^*_{n_b}(T_b; \mathcal{S}_b) \cdot \left(\varepsilon^*_B(\hat{\mu}) + \frac{1}{B} \right) + \frac{B}{n} \cdot \varepsilon^*_B(\hat{\mu}) \,.$$

Remarks:

$$\varepsilon^*_{RLB,n,B} \geq \varepsilon^*_{n_b}(T_b; \mathcal{S}_b) \cdot \varepsilon^*_B(\hat{\mu}) \,.$$

$$\text{If } \; \varepsilon^*_B(\hat{\mu}) = 0 : \quad \varepsilon^*_{n_b}(T_b; \mathcal{S}_b)/B \to 0, \quad \text{if } B \to \infty \,.$$

$$\text{For Median: } \; \varepsilon^*_B(\hat{\mu}) \approx 1/2 \,.$$

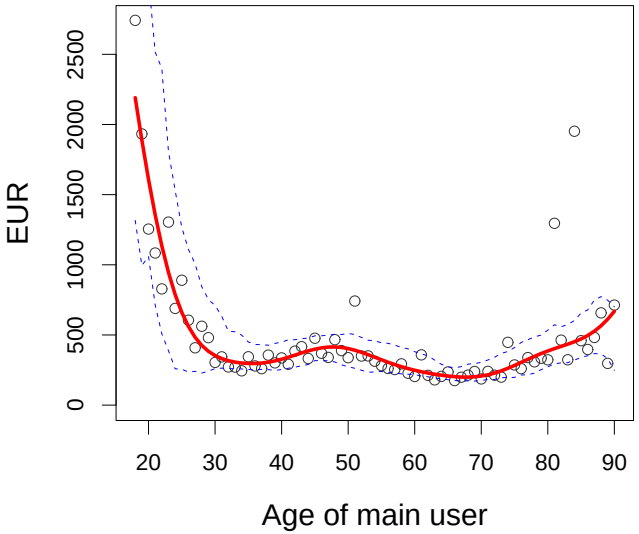
4. Application: Motor Vehicle Insurance

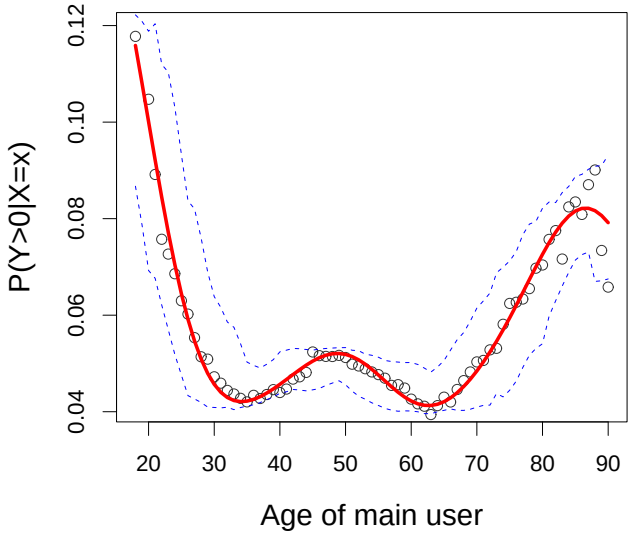
- 'risk differentiation in high-dimensional data structures' in SFB 475 (joint with A. Kovac, Bristol)
- Verband öffentlicher Versicherer, Düsseldorf
- Data of 15 insurance companies
- 3×3 GB compressed SAS-file
- $>$ 4.6 millions customers
- $>$ 70 explanatory variables, many discrete

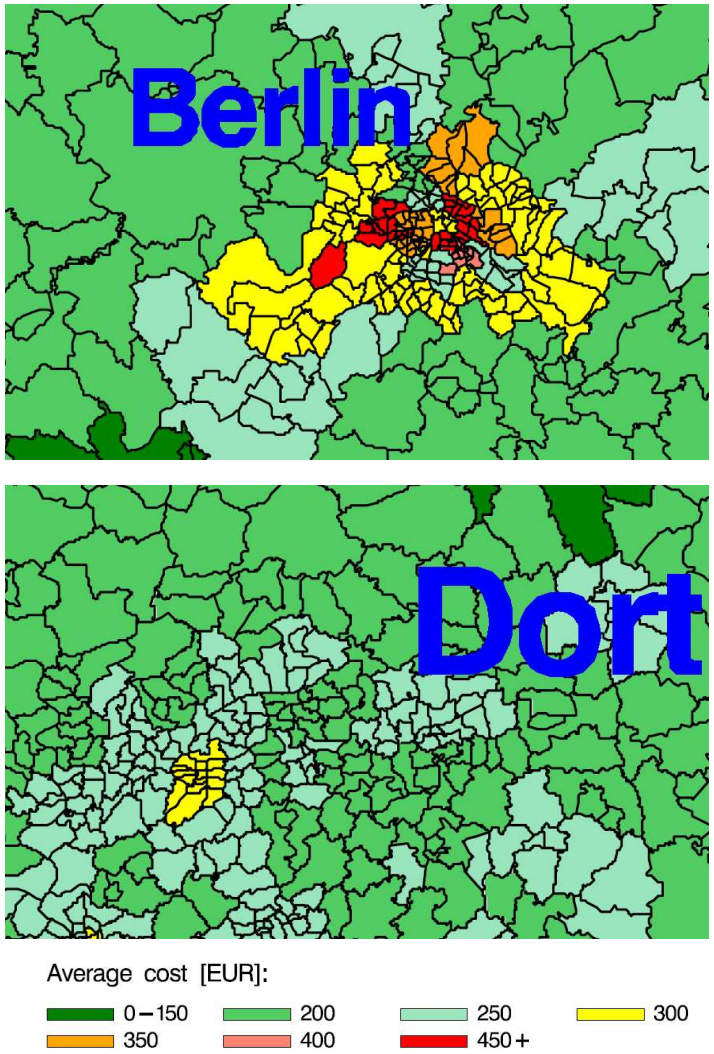
Statistical objectives

- Y claim amount [EUR]
- $x \in \mathbb{R}^d$ explanatory variables
- Actual premium charged to the customer:
  pure premium + safety loading + administrative costs + desired profit
- Primary response: pure premium $\mathrm{E}(Y|X = x)$
- Secondary response: prob. of claim $\mathrm{P}(Y > 0|X = x)$
- (individual) claim amount [year]: $y_i = \dfrac{\sum_{j=1}^{n_i} y_{i,j}}{t_i/360}$

Complex dependencies







Statistical model

- Y: pure premium [EUR]
- X: 8 explanatory variables: *Age of main user*,
  gender, garage, driving distance, geographical region,
  population density, number of years without claims, strength of engine
- $C := 0$, if $Y = 0$, $C := 1$, if $Y \in (0, 2000]$,
  $C := 2$, if $Y \in (2000, 10000]$, $C := 3$, if $Y \in (10000, 50000]$
  $C := 4$, if $Y > 50000$
- Estimation by (KLR , $\varepsilon-$SVR) with RBF-kernel:

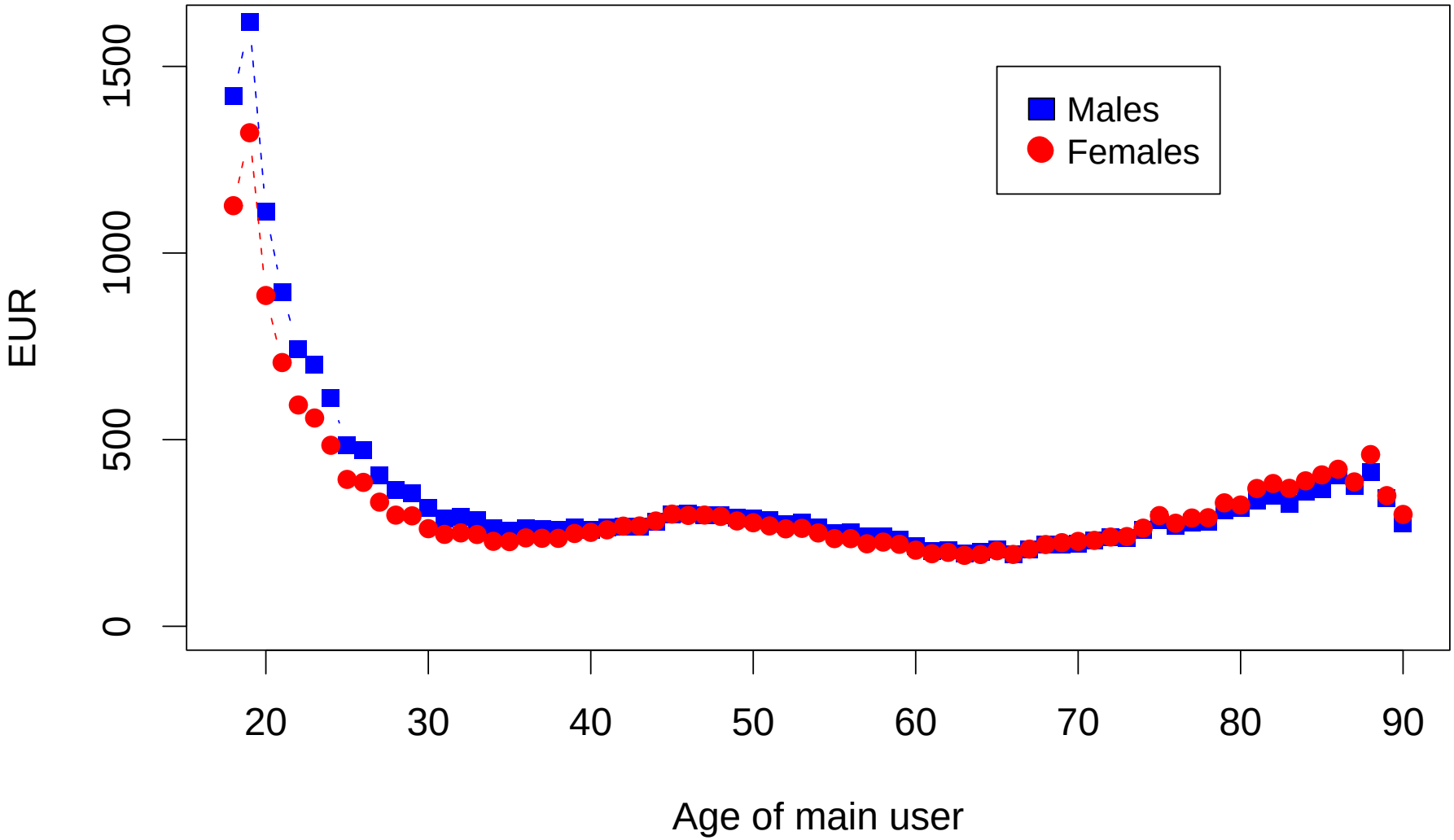
$$\mathrm{E}(Y|X = x) = \mathrm{P}(C > 0|X = x) \cdot \sum_{c=1}^{4} \mathrm{P}(C = c|C > 0, X = x) \cdot \mathrm{E}(Y|C = c, X = x)$$

- Splitting: 50% training, 25% validation, 25% test

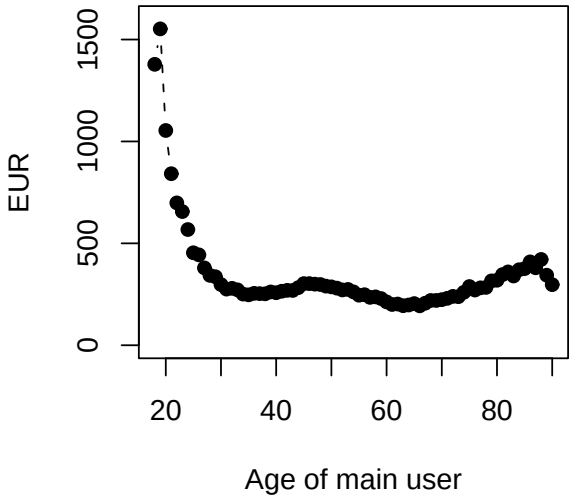
Problem of parametric models:

201 main effects, 15604 interaction terms of order 1, $>$ 600.000 order 2, $>$ 14 millions of order 3

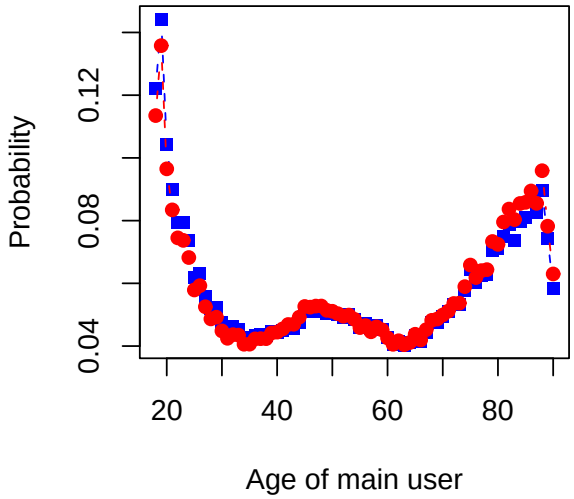
Estimation of pure premium E(Y|X=x)



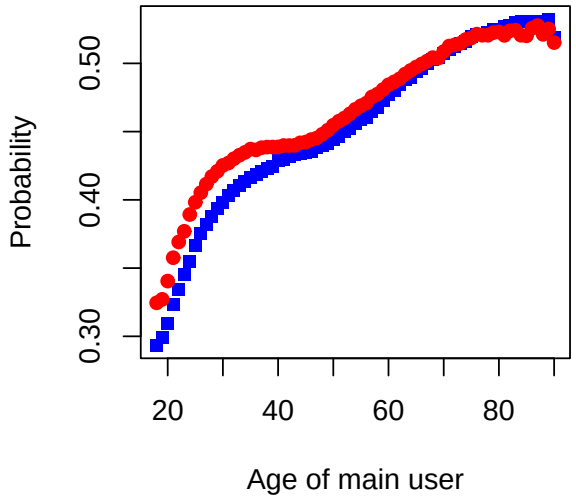
E(Y|X=x)



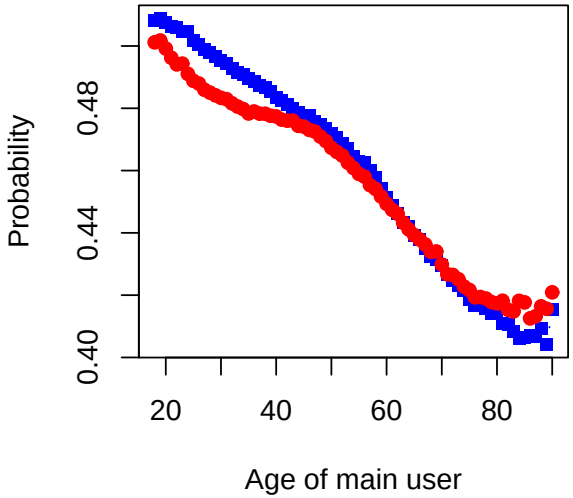
P(Y>0 | X=x)



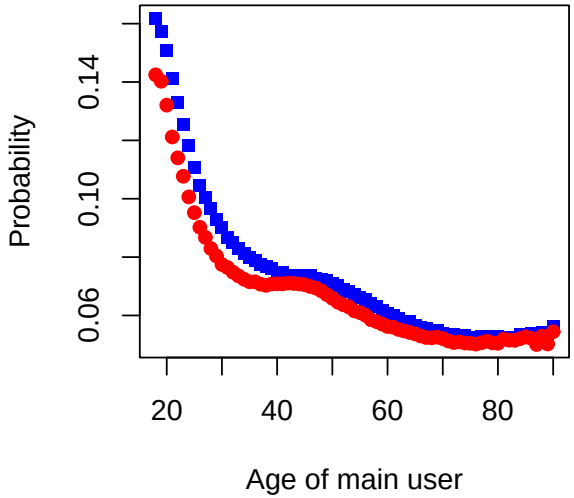
P(Yε(0,2000] | Y>0, X=x)



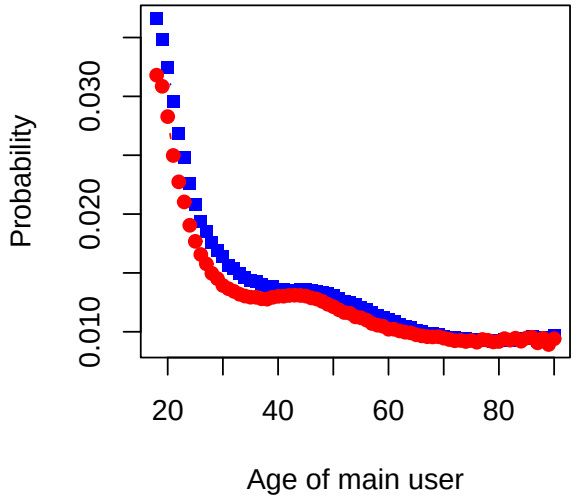
P(Yε(2000,10000] | Y>0, X=x)



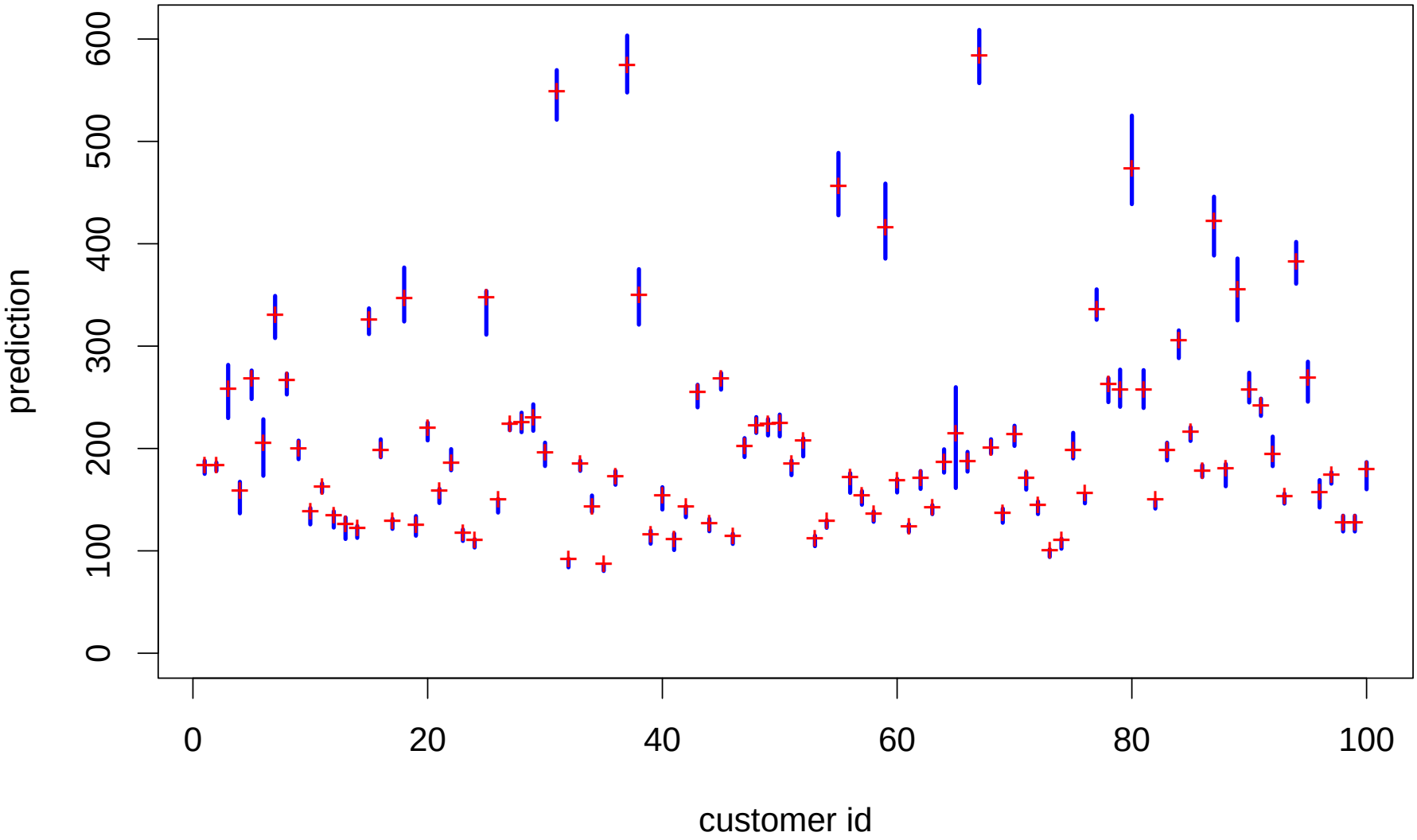
P(Yε(10000,50000] | Y>0, X=x)



P(Y>50000 | Y>0, X=x)



RLB based on median: predictions $\hat{y}_i$ for 100 customers



95% non-parametric confidence intervals: asymmetric + width increasing !

4. Summary

convex risk minimization:

- can model complex high-dimensional dependency structures
- robustness for classification and regression, if L' and k bounded
- L-risk consistency $\leftrightsquigarrow$ robustness

References

Chr (2005). RLB: Robust Learning from Bites. SFB-475, TR-07/05.
Chr & Steinwart (2005). Consistency and robustness of kernel based regression. SFB-475, TR-01/05.
Chr & Steinwart (2004). On robust properties of convex risk minimization methods for pattern recognition. *J. Machine Learning Research*, 5, 1007-1034.
Chr (2004). An approach to model complex high-dimensional insurance data. *Allg. Statist. Archiv*, 88, 375–397.
Chr (2004). *On Properties of Support Vector Machines for Pattern Recognition in Finite Samples.* In: Theory and Applications of Recent Robust Methods. Eds.: M. Hubert et al. Birkhäuser, Basel.
Chr & Fischer, Joachims (2002). *Computational Statistics*, 17, 273-287.
– – – – –
Bartlett, Tewari (2004). Berkeley, Preprint.
Hampel et al. (1986). *Robust statistics: The Approach Based on Influence Functions.* Wiley.
Höffgen, Simon, van Horn (1995). *J. Computer and System Sciences*, 50, 114-125.
Huber (1981). *Robust statistics.* Wiley.
Mukherjee, Niyogi, Poggio, Rifkin (2004). MIT. CBCL 223 Preprint.
Poggio, Rifkin, Mukherjee, Niyogi (2004). *Nature*, 428, 419-422.
Schölkopf, Smola (2002). *Learning with Kernels.* MIT Press.
Vapnik (1998). *Statistical Learning Theory.* Wiley.
Zhang (2004). *Ann. Statist.*, 32, 56–134.

Further references

Breiman (1996). Bagging predictors. *Machine Learning*, 24, 123–140.
Breiman (1999). Pasting bites together for prediction in large data sets. *Machine Learning*, 36, 85–103.
Chawla et al. (2004). Learning ensembles for bites: a scalable and accurate approach. *J. Machine Learning Research*, 5, 421–451.
Davies & Gather (2005). Breakdown and groups (with discussion). *Ann.Statist.*
Grandvalet (2004). Bagging equalizes influence. *Machine Learning*, 55, 251-270.
Politis, Romano, Wolf (1999). *Subsampling*, Springer.
Rousseeuw, Bassett (1990). The remedian: ... *JASA*, 85, 97-104.