

Christmann, Andreas; Steinwart, Ingo; Hubert, Mia

**Working Paper**

## Robust Learning from Bites for Data Mining

Technical Report, No. 2006,03

**Provided in Cooperation with:**

Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475),  
University of Dortmund

*Suggested Citation:* Christmann, Andreas; Steinwart, Ingo; Hubert, Mia (2006) : Robust Learning from Bites for Data Mining, Technical Report, No. 2006,03, Universität Dortmund, Sonderforschungsbereich 475 - Komplexitätsreduktion in Multivariaten Datenstrukturen, Dortmund

This Version is available at:

<https://hdl.handle.net/10419/22651>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# Robust Learning from Bites for Data Mining

BY ANDREAS CHRISTMANN<sup>1</sup>, INGO STEINWART<sup>2</sup>, AND MIA HUBERT<sup>3</sup>,  
*University of Dortmund*<sup>1</sup>, *Los Alamos National Laboratory*<sup>2</sup>, and *Katholieke  
Universiteit Leuven*<sup>3</sup>

Some methods from statistical machine learning and from robust statistics have two drawbacks. Firstly, they are computer-intensive such that they can hardly be used for massive data sets, say with millions of data points. Secondly, robust and non-parametric confidence intervals for the predictions according to the fitted models are often unknown. Here, we propose a simple but general method to overcome these problems in the context of huge data sets. The method is scalable to the memory of the computer, can be distributed on several processors if available, and can help to reduce the computation time substantially. Our main focus is on robust general support vector machines (SVM) based on minimizing regularized risks. The method offers distribution-free confidence intervals for the median of the predictions. The approach can also be helpful to fit robust estimators in parametric models for huge data sets.

## 1. Introduction

Data sets with millions of observations occur nowadays in many areas, *e.g.* insurance companies or banks collect many variables to develop tariffs and scoring methods for credit risk management, respectively. Other examples are large observational data sets in data mining projects and data from micro-arrays. Although such big data sets contain a lot of valuable information, the analysis of such data sets can not only be cumbersome due to computer memory or computational time problems. Classical parametric assumptions are often violated for such data sets which contain probably some outliers. We give only to three citations for these facts. J.W. Tukey, one of the pioneers of robust statistics, mentioned already in 1960 (citet from Hampel *et al.* (1986, p. 21)): *"A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians."* Le Cam (1980, p.478) concluded for data sets with  $n = 10^5$  to  $n = 10^8$  data points: *"Thus the asymptotics fail precisely when one would feel that they are applicable."* Hampel *et al.* (1986, p. 27f) made the following comment on data quality and gross errors. *"There are often no or virtually no gross errors in high-quality data, but 1% to 10% of gross errors in routine data seem to be more the rule than the exception"*. Hence, it is no surprise that the

- 
1. Address: University of Dortmund, Department of Statistics, D-44221 Dortmund, GERMANY.
  2. Address: Los Alamos National Laboratory, Los Alamos, NM 87545, CCS-3, Mail Stop B256, USA.
  3. Address: Katholieke Universiteit Leuven, Department of Mathematics, W. de Croylaan 54, B-3001 Leuven, BELGIUM.
  3. *AMS 2000 subject classification.* Primary 62G08, 62G35; secondary 68Q32, 62G20.
  4. *Keywords and Phrases.* Breakdown point, convex risk minimization, data mining, distributed computing, influence function, logistic regression, robustness, scalability, statistical machine learning, support vector machine.

data quality in large data mining problems is often far from being optimal, *cf.* Hand *et al.* (2001) or Hipp *et al.* (2001).

The application of robust statistical methods is therefore important in such situations. Unfortunately, many robust methods proposed in the literature have the following drawbacks which are serious limitations for their application. (a) They are computer-intensive such that they can hardly be used for massive data sets, say for several millions of observations with hundreds of explanatory variables. (b) Robust standard errors and robust confidence intervals for the estimated parameters or for robust predictions are often unknown. (c) Some statistical software packages like S-PLUS or R contain state-of-the-art algorithms for robust statistical methods, but the implemented numerical algorithms usually require that the whole data set fits into the memory of the computer.

In this paper a simple but quite general method for robust estimation in the context of huge data sets is proposed. The main goal of the proposal is to broaden in application of robust general SVM methods for massive data. The idea is to partition the huge data set  $S$  by random into disjoint subsets  $S_b$ ,  $b = 1, \dots, B$ . Then a robust method is applied to each subset, and the results are summarized in a robust manner. The proposal yields robust predictions. If the median is used to aggregate the  $B$  single predictions then we also get robust and distribution-free confidence intervals.

The rest of the paper is organized as follows. Section 2 gives the proposed method and Section 3 describes its properties. Section 4 given some numerical examples for the case of robust linear regression and kernel logistic regression. Section 5 contains a summary and compares RLB with competing methods. All proofs are given in the Appendix.

## 2. Method

In this section we describe a simple but rather general method for robust estimation for huge data sets. We restrict attention to classification and regression problems although the method can be used in other fields as well. The proposal has two goals: making robust general SVM methods usable for data sets which are too large for currently available algorithms due to memory or time limitations and offering robust and distribution-free confidence intervals based on the median for the predictions.

In classification and in regression problems one assumes an approximate functional relationship between an explanatory random variable  $X$  and a response random variable  $Y$  using  $n$  observations  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$  drawn independently from the same probability distribution  $P$  of the pair  $(X, Y)$ . In a non-parametric setting the distribution  $P$  is totally unknown. For technical reasons we assume throughout this work that  $\mathcal{X}$  and  $\mathcal{Y}$  are closed or open subsets of  $\mathbb{R}^d$  and  $\mathbb{R}$ , respectively. Hence we can split up  $P$  into the marginal distribution  $P_X$  and the regular conditional probability  $P(\cdot | x)$ ,  $x \in \mathcal{X}$ , on  $\mathcal{Y}$ . For the case of binary classification we have  $\mathcal{Y} = \{-1, +1\}$ .

For regression one often imposes the classical signal plus noise assumption, i.e. the assumption that  $Y_i | (X = x_i)$  is distributed as  $f(x_i) + \varepsilon_i$ , where  $f$  is an unknown function and  $\varepsilon_i$  are independent and identically distributed error terms,  $1 \leq i \leq n$ . In the linear (parametric) setup we additionally assume  $f(x) = f_\theta(x) = x'\theta$ ,  $\theta \in \Theta \subset \mathbb{R}^d$ , and in the general non-parametric setup  $f$  is simply a measurable function, *i.e.*  $f : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . A possible intermediate case is based on the assumption that  $f$  belongs to some Hilbert

space  $H$  of all measurable functions  $f : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . In our case  $H$  can be a (typically infinite dimensional) reproducing kernel Hilbert space (RKHS).

We assume that the statistical method of interest can be written as a function of the empirical distribution  $\frac{1}{n} \sum_{i=1}^n \delta_{\{(x_i, y_i)\}}$  based on the data set  $\mathcal{S}$  having data points  $(x_i, y_i)$ ,  $i = 1, \dots, n$ .<sup>1</sup> Here  $\delta_{\{z\}}$  denotes the Dirac-distribution in  $\{z\}$ . More general we assume that  $T_P$  is the quantity of interest for any distribution  $P$ , where

$$T : P \mapsto T_P \tag{1}$$

is a measurable function defined on the space of probability distributions  $\mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$ . Examples are the mean  $T_P = \mathbb{E}_P(X) = \int X dP$  (if existent) and the minimizer  $T_P = f_{P, \lambda}$ ,  $\lambda > 0$ , of the regularized theoretical risk defined in (8).

In this paper we always assume that the sample size  $n$  is large. The whole data set is often partitioned by random into two or three disjoint parts for training, validation, and testing purposes. Instead of modelling the full training data set, we split the training data set by random into  $B \geq 1$  parts  $\mathcal{S}_b$  (called 'bites') of approximately the same sub-sample sizes  $n_b \approx n/B$ . Then we fit each bite with the robust method. Finally, we compute a robust location estimator of the estimators  $T_{\mathcal{S}_b}$  and summarize the predictions from the  $B$  fitted models.

**Definition 1** *Let  $\mathcal{S} = ((x_1, y_1), \dots, (x_n, y_n))$  be a sample of size  $n$  from a probability distribution  $P$  on  $(X \times Y, \mathcal{B}(X \times Y))$ . Let  $T_{\mathcal{S}}$  be the estimator of interest. Consider a random partition of  $\mathcal{S}$  into  $B$  non-empty subsets, i.e.*

$$\mathcal{S} = \mathcal{S}_1 \cup \dots \cup \mathcal{S}_B, \tag{2}$$

where  $\mathcal{S}_b \subset \mathcal{S}$ ,  $n_b := \#\mathcal{S}_b \approx \lfloor n/B \rfloor \in \mathbb{N}$ ,  $n = \sum_{b=1}^B n_b$ ,  $b = 1, \dots, B$ ,  $B \in \{1, \dots, n\}$ ,  $B \ll n$ . An **RLB estimator of type I** is defined by

$$T_{\mathcal{S}, B}^{RLB} = g(T_{\mathcal{S}_1}, \dots, T_{\mathcal{S}_B}), \tag{3}$$

where  $g : H^B \rightarrow H$  is a measurable map. An **RLB estimator of type II** is given by

$$T_{\mathcal{S}, B}^{RLB}(x) = g^*(T_{\mathcal{S}_1}(x), \dots, T_{\mathcal{S}_B}(x)), \quad \forall x \in X, \tag{4}$$

where  $g^* : \mathbb{R}^B \rightarrow \mathbb{R}$  is a measurable map.

**Remarks.** (i) An RLB estimator of type I can obviously be used to define an RLB estimator of type II. (ii) An RLB estimator of type II does not necessarily define an RLB estimator of type I, because the related function  $g^*$  does not necessarily correspond to a function  $g$  mapping onto the Hilbert space  $H$ . (iii) The class of RLB estimators of type I – and due to part (i) the class of RLB estimators of type II – is non-empty, because for  $g$  equal to the mean we obtain:  $g(T_{\mathcal{S}_1}, \dots, T_{\mathcal{S}_B}) := \frac{1}{B} \sum_{b=1}^B T_{\mathcal{S}_b} \in H$ .

---

1. If misunderstandings are improbable, we will use the symbol  $\mathcal{S}$  for the data set and for the corresponding empirical distribution.

We will mainly consider RLB estimators of type I which are *convex* combinations

$$T_{\mathcal{S},B}^{RLB} = \sum_{b=1}^B c_b T_{\mathcal{S}_b} \quad (5)$$

with weights  $c_b \in (0, 1)$  and  $\sum_{b=1}^B c_b = 1$  ( $c_b \equiv \frac{1}{B}$  gives the mean), and RLB estimators of type II based on the median. Of course, L-estimators such as  $\alpha$ -trimmed means, M-, S-, and R-estimators can also be used on the aggregation step.

If  $B$  is large enough, say above 15, precision estimates can additionally be obtained by computing standard deviations of the predictions  $T_{\mathcal{S},B}^{RLB}(x)$  using the central limit theorem. However, in general we favor a distribution-free method based on the median. If  $B$  is small or if the distribution or the variance of  $T_{\mathcal{S},B}^{RLB}(x)$  is unknown, one can construct distribution-free confidence intervals for the median of  $T_{\mathcal{S},B}^{RLB}(x)$  and distribution-free tolerance regions based on selected order statistics, see David and Nagaraja (2003, Chap. 7). Table 1 lists some values of  $B$ , the corresponding pair of order statistics determining the confidence interval, the lower bound of the actual confidence level which is  $0.5^B \sum_{j=r}^s \binom{B}{j}$ , and the finite sample breakdown point  $\varepsilon_B^* = \min\{r - 1, B - s\}/B$  of the confidence interval, see Definition 10. In Section 3 it will be shown that RLB inherits robustness properties from the original estimator and from the estimator used in the aggregation step. The actual confidence intervals based on the median can be conservative for small choices of  $B$ , see Table 1. If  $B$  is not too small, say  $B > 15$ , this breakdown point is high enough for many practical applications. E.g. fix  $B = 17$ . Then the 5<sup>th</sup> and the 13<sup>th</sup> order statistics give a confidence interval at the level 95% for the median which is valid for *all* distributions on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . The breakdown point of this confidence interval is  $4/17 = 0.235$  because the values of the four lowest and the four highest predictions are not used.

### 3. Properties of RLB

In this section properties of robust learning from bites are investigated. Computational time and memory space are considered in Section 3.1. RLB for general SVM estimators is investigated in Section 3.2, and robustness properties are proved in Section 3.3. In Section 3.4 some arguments are given how to choose the number of bites. All proofs are given in the appendix.

We will assume in this section that  $\min_{1 \leq b \leq B} n_b \rightarrow \infty, n \rightarrow \infty$ .

#### 3.1 General properties

Denote the estimator based on the whole data set by  $T_{\mathcal{S}}$  and denote the corresponding RLB estimator based on  $B$  bites,  $B$  fixed, with sub-sample sizes  $n_b$ , where  $n = \sum_{b=1}^B n_b$ , by  $T_{\mathcal{S},B}^{RLB}$ .

The estimators  $T_{\mathcal{S}_b}$ ,  $1 \leq b \leq B$ , from the bites are stochastically independent because they are computed from disjoint parts of the data set. The computation time and the memory space for RLB can be obviously approximated in the following way. Denote the number of available CPUs by  $k$  and let  $k_B$  be the smallest integer which is not smaller than  $B/k$ .

ROBUST LEARNING FROM BITES

$1 - \alpha$	$B$	$r$	$s$	lower bound of confidence level	finite sample breakdown point
0.90	8	2	7	0.930	0.125
	10	2	9	0.979	0.100
	13	4	10	0.908	0.231
	18	6	13	0.904	0.278
	30	11	20	0.901	0.333
	37	14	24	0.901	0.351
	44	17	28	0.904	0.364
	53	21	33	0.902	0.377
	62	25	38	0.902	0.387
	71	29	43	0.904	0.394
	82	34	49	0.903	0.402
	93	39	55	0.903	0.409
	104	44	61	0.905	0.413
0.95	9	2	8	0.961	0.111
	10	2	9	0.979	0.100
	17	5	13	0.951	0.235
	37	13	25	0.953	0.324
	51	19	33	0.951	0.353
	58	22	37	0.952	0.362
	67	26	42	0.950	0.373
	74	29	46	0.953	0.378
	83	33	51	0.952	0.386
	92	37	56	0.953	0.391
	101	41	61	0.954	0.396
0.99	10	1	10	0.998	0.000
	12	2	11	0.994	0.083
	26	7	20	0.991	0.231
	39	12	28	0.991	0.282
	49	16	34	0.991	0.306
	61	21	41	0.990	0.328
	73	26	48	0.990	0.342
	80	29	52	0.990	0.350
	94	35	60	0.990	0.362
	101	38	64	0.991	0.366

Table 1: Selected pairs  $(r, s)$  of order statistics for non-parametric confidence intervals at the  $(1 - \alpha)$ -level for the median.

**Proposition 2 (Computation time,  $k$  CPUs)** *Assume that the computation time of  $T_S$  for a data set with  $n$  observations and  $d$  explanatory variables is of order  $O(h(n, d))$ , where*

$h$  is some positive function. Then the computation time of  $T_{\mathcal{S},B}^{RLB}$  with subsample sizes  $n_b \approx n/B$  is approximately of order  $O(k_B \cdot h(n/B, d))$ .

**Proposition 3 (Memory space,  $k$  CPUs)** Assume that the estimator  $T_{\mathcal{S}}$  for a data set with  $n$  observations and  $d$  explanatory variables needs memory space and hard disk space of order  $O(h_1(n, d))$  and  $O(h_2(n, d))$ , respectively, where  $h_1$  and  $h_2$  are positive functions. Then the computation of  $T_{\mathcal{S},B}^{RLB}$  for subsample sizes  $n_b \approx n/B$  needs approximately memory space and hard disk space of order  $O(k \cdot h_1(n/B, d))$  and  $O(k \cdot h_2(n/B, d))$ , respectively.

**Proposition 4 (Consistency)** Consider an RLB estimator  $T_{\mathcal{S},B}^{RLB}$  of type I based on a convex combination with  $c_b \in (0, 1)$  and  $\sum_{b=1}^B c_b = 1$ .

(i) If  $\mathbb{E}(T_{\mathcal{S}_b}) = \mathbb{E}(T_{\mathcal{S}})$  for all  $b \in \{1, \dots, B\}$ , then  $\mathbb{E}(T_{\mathcal{S},B}^{RLB}) = \mathbb{E}(T_{\mathcal{S}})$ .

(ii) If  $T_{\mathcal{S}}$  converges in probability (or almost sure) to  $T_{\mathcal{P}}$  for  $n \rightarrow \infty$  and if  $(n/n_b) \rightarrow B$ ,  $B$  fixed, then  $T_{\mathcal{S},B}^{RLB}$  converges in probability (or almost sure) to  $T_{\mathcal{P}}$ .

(iii) Let  $c_b \equiv \frac{1}{B}$ . Assume that  $n_b^{1/2}(T_{\mathcal{S}_b} - T_{\mathcal{P}})$  converges weakly to a multivariate normal distribution  $N(0, \Sigma)$ , where  $\Sigma \in \mathbb{R}^{d \times d}$  is positive definite, and that  $(n/n_b) \rightarrow B$ ,  $1 \leq b \leq B$ ,  $B$  fixed. Then  $n^{1/2}(T_{\mathcal{S},B}^{RLB} - T_{\mathcal{P}})$  converges weakly to a multivariate normal distribution  $N(0, \Sigma)$ ,  $n \rightarrow \infty$ .

**Proposition 5 (Consistency)** Consider an RLB estimator  $T_{\mathcal{S},B}^{RLB}$  of type II where the **median** is used in the aggregation step. If  $T_{\mathcal{S}}(x)$  converges in probability (or almost sure) to  $T_{\mathcal{P}}(x)$ ,  $x \in \mathcal{X}$ , and if  $\lim_{n \rightarrow \infty} (n/n_b) \equiv B$ ,  $B$  fixed, then  $T_{\mathcal{S},B}^{RLB}(x)$  converges in probability (or almost sure) to  $T_{\mathcal{P}}(x)$ .

### 3.2 Properties of RLB using the mean for general SVM methods

Now we consider general SVM estimators

$$f_{\mathcal{S},\lambda} := \arg \min_{f \in H} \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)) + \lambda \|f\|_H^2, \quad (6)$$

where  $L : \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty)$  is a convex loss function, i.e. a measurable function which is convex in its second argument,  $H$  is the reproducing kernel Hilbert space defined via the kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , and regularizing parameter  $\lambda > 0$ , see Vapnik (1998) and Schölkopf and Smola (2002). Special cases of such general SVM methods are the *support vector machine*:  $L(y, t) = \max\{0, 1 - yt\}$ ,  $y \in \{-1, +1\}$ ,  $t \in \mathbb{R}$ , *kernel logistic regression*:  $L(y, t) = \ln(1 + \exp[-yt])$ ,  $y \in \{-1, +1\}$ ,  $t \in \mathbb{R}$ , and  $\varepsilon$ -*support vector regression*:  $L(y, t) = \max\{|y - t| - \varepsilon, 0\}$ ,  $y, t \in \mathbb{R}$ , where  $\varepsilon > 0$  is fixed. The general SVM estimator  $f_{\mathcal{S}_b, \lambda_{n_b}}(x)$ ,  $x \in \mathcal{X}$ , defined as the solution of (6) for bite  $\mathcal{S}_b$  is a kernel based estimator and can be written as

$$f_{\mathcal{S}_b, \lambda_{n_b}}(x) = \sum_{i=1}^{n_b} \alpha_{i,b} k(x, x_i), \quad i \in \mathcal{S}_b, x \in \mathcal{X}, \quad (7)$$

where  $\alpha_{i,b} \in \mathbb{R}$ . If  $\alpha_{i,b} \neq 0$ , then  $(x_i, y_i)$  is called a support vector (SV). Obviously, the minimization problem (6) can be interpreted as a stochastic approximation of the minimization of the theoretical regularized risk

$$f_{\mathcal{P},\lambda} := \arg \min_{f \in H} \mathbb{E}_{\mathcal{P}} L(Y, f(X)) + \lambda \|f\|_H^2 \in H, \quad (8)$$

where the minimizer  $f_{P,\lambda}$  exists under rather general assumptions on  $H$  and  $L$ , see e.g. Steinwart (2005a) and Christmann and Steinwart (2005).

**Theorem 6 (RLB for general SVMs)** *Assume that the estimator  $f_{S,\lambda}$  is a general SVM estimator defined by (6) for the whole data set with  $n = \sum_{b=1}^B n_b$  observations,  $B$  fixed. Consider an RLB estimator of type I based on a convex combination with  $c_b \in (0, 1)$  and  $\sum_{b=1}^B c_b = 1$ . Then the RLB estimator is itself a kernel based estimator and can be written as*

$$f_{S,B,(\lambda_{n_b})}^{RLB}(x) = \sum_{i=1}^n \alpha_{i,RLB} k(x, x_i) \quad (9)$$

$$= \sum_{i \in SV(S_1) \cup \dots \cup SV(S_B)} \alpha_{i,RLB} k(x, x_i), \quad x \in \mathcal{X}, \quad (10)$$

where  $\alpha_{i,RLB} = \sum_{b=1}^B c_b \alpha_{i,b}$ ,  $i \in \mathcal{S}$ .

We have  $\alpha_{i,RLB} = c_b \alpha_{i,b}$  in (10) if all support vectors are different.

Let us now investigate the *number of support vectors* in more detail for the case of pattern recognition, *i.e.*  $\mathcal{Y} = \{-1, +1\}$ . For part (ii) of our next result we need the following quantities. Denote the marginal distribution of  $X$  by  $P_X$ . Let  $X_0 := \{x \in \mathcal{X}; P(1|X = x) = 1/2\}$ ,  $\mathcal{X}_{cont} := \{x \in \mathcal{X}; P_X(\{x\}) = 0\}$ , and let  $\partial_2 L$  denote the subdifferential operator of the loss function  $L$  with respect to the second variable. Further, define the set-valued function

$$F_L^*(\alpha) := \{t \in \overline{\mathbb{R}}; [\alpha L(1, t) + (1 - \alpha)L(-1, t)] = \min_{s \in \overline{\mathbb{R}}} [\alpha L(1, s) + (1 - \alpha)L(-1, s)]\}, \quad \alpha \in [0, 1],$$

the set

$$S = \{(x, y) \in \mathcal{X}_{cont} \times \mathcal{Y}; 0 \notin \partial_2 L(y, F_L^*(P(1|X = x))) \cap \mathbb{R}\},$$

and the quantity

$$S_{L,P} = \begin{cases} P(S) & \text{if } 0 \notin \partial_2 L(1, F_L^*(0.5)) \cap \partial_2 L(-1, F_L^*(0.5)) \\ P(S) + \frac{1}{2} P_X(X_0 \cap \mathcal{X}_{cont}) & \text{else,} \end{cases}$$

see Steinwart (2003, p.1082). We also need the notion of a classification calibrated loss function. Such loss functions were called admissible by Steinwart (2003), but we think that the notion of classification calibrated is more precise. A loss function is called classification calibrated if for every  $\alpha \in [0, 1]$  we have

$$\begin{aligned} F_L^*(\alpha) &\subset [-\infty, 0) & \text{if } \alpha < 1/2 \\ F_L^*(\alpha) &\subset (0, \infty] & \text{if } \alpha > 1/2. \end{aligned}$$

For more information on this and related concepts we refer to Bartlett *et al.* (2006) and Steinwart (2005b).

Finally, we need a way to describe the richness of the reproducing kernel Hilbert space  $H$ . In this work we use the following definition taken from Steinwart (2001):



**Definition 7** Let  $\mathcal{X} \subset \mathbb{R}^d$  be compact and  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a continuous kernel with reproducing kernel Hilbert space  $H$ . Then  $k$  is **universal** if  $H$  is dense in the space of continuous functions  $C(\mathcal{X})$  equipped with  $\|\cdot\|_\infty$ .

Now we can formulate the following theorem on the number of support vectors:

**Theorem 8 (Number of support vectors)** Consider an RLB estimator of type I based on a convex combination with  $c_b \in (0, 1)$  and  $\sum_{b=1}^B c_b = 1$ . Under the assumptions of Theorem 6 the RLB estimator has the following properties.

(i) The number of support vectors, i.e.  $\alpha_{i,RLB} \neq 0$ , of the RLB estimator is given by

$$\#\{SV(\mathcal{S}_1) \cup \dots \cup SV(\mathcal{S}_B)\}. \quad (11)$$

(ii) Consider a binary classification problem, i.e.  $\mathcal{Y} = \{-1, +1\}$ . Let  $B$  be fixed, and consider  $n := B \cdot n_b \rightarrow \infty$ . Let  $L$  be a classification calibrated and convex loss function,  $k$  be a universal kernel and  $\lambda_{n_b} > 0$  be a sequence of regularization parameters with  $\lambda_{n_b} \rightarrow 0$  and  $n_b \lambda_{n_b}^2 / |L_{\lambda_{n_b}}|_1^2 \rightarrow \infty$ . Then for all Borel probability measures  $P$  on  $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$  the RLB-classifier based on (6) with respect to  $k$ ,  $L$  and  $(\lambda_{n_b})$  satisfies

$$\Pr^{*n} \left( \mathcal{S}_1 \cup \dots \cup \mathcal{S}_B \in (\mathcal{X} \times \mathcal{Y})^n; \#SV(f_{\mathcal{S},B,(\lambda_{n_b})}^{RLB}) \geq (S_{L,P} - \varepsilon)n \right) \rightarrow 1. \quad (12)$$

Here  $\Pr^{*n}$  denotes the outer probability measure of  $P^n$  in order to avoid measurability considerations.

Note that part (ii) of the above result gives for the RLB estimator the *same* asymptotical bound for the number of support vectors as the one for  $B = 1$  derived by Steinwart (2003). The result given in (12) has the following interpretation: with probability tending to 1 if the total sample size  $n = Bn_b$  converges to  $\infty$ , but  $B$  is fixed, the fraction of support vectors of the kernel based RLB estimator  $f_{\mathcal{S},B,(\lambda_{n_b})}^{RLB}(x)$  in a binary pattern recognition problem is essentially greater than the Bayes risk.

Now we investigate conditions to guarantee that RLB estimators using general SVM estimators are  $L$ -risk consistent, i.e. that they are able to learn. If  $P$  is a probability distribution on  $\mathcal{X} \times \mathcal{Y}$ , the  $L$ -risk of a measurable map  $f : \mathcal{X} \rightarrow \mathbb{R}$  with respect to  $P$  is defined by

$$\mathcal{R}_{L,P}(f) := \int L(Y, f(X)) dP = \int L(y, f(x)) P(dy|x) P_X(dx).$$

The above integral is always defined since  $L$  is non-negative and continuous, although it may be infinite. Consider a general SVM estimator  $f_{\mathcal{S},\lambda}$  defined by (6) for the whole data set  $\mathcal{S}$ . The estimator  $f_{\mathcal{S},\lambda_n}$  is called  $L$ -risk consistent, if

$$\mathcal{R}_{L,P}(f_{\mathcal{S},\lambda_n}) \rightarrow \mathcal{R}_{L,P}^* := \inf\{\mathcal{R}_{L,P}(f); f : X \rightarrow \mathbb{R} \text{ measurable}\} \quad (13)$$

holds in probability for  $n \rightarrow \infty$  for suitable chosen regularization sequences  $(\lambda_n)_{n \in \mathbb{N}}$ . Several authors have given conditions to guarantee that general SVM estimators are  $L$ -risk consistent, cf. Steinwart (2002, 2005a), Zhang (2004), and Christmann and Steinwart (2005).

If  $f_{\mathcal{S},\lambda_n}$  is  $L$ -risk consistent,  $B \geq 1$  fixed, and  $\lim_{n \rightarrow \infty} \min_{1 \leq b \leq B} n_b = \infty$ , we obtain by Slutsky's theorem for an RLB estimator of type I based on a convex combination with weights  $c_b \in (0, 1)$  and  $\sum_{b=1}^B c_b = 1$  that

$$\sum_{b=1}^B c_b \mathcal{R}_{L,P}(f_{\mathcal{S}_b,\lambda_{n_b}}) \rightarrow \mathcal{R}_{L,P}^* \quad (14)$$

in probability for  $n \rightarrow \infty$ . The next result gives  $L$ -risk consistency of RLB estimators of type I using a convex combination.

**Theorem 9 (L-risk consistency)** *Let  $f_{\mathcal{S},\lambda_n}$  be an  $L$ -risk consistent general SVM estimator based on (6) with a convex loss function. Then the RLB estimator of type I defined by  $f_{\mathcal{S},B,(\lambda_{n_b})}^{RLB} = \sum_{b=1}^B c_b f_{\mathcal{S}_b,\lambda_{n_b}}$  with  $c_b \in (0, 1)$  and  $\sum_{b=1}^B c_b = 1$  is  $L$ -risk consistent, i.e.*

$$\mathcal{R}_{L,P} \left( \sum_{b=1}^B c_b f_{\mathcal{S}_b,\lambda_{n_b}} \right) \xrightarrow{P} \mathcal{R}_{L,P}^*, \quad n \rightarrow \infty. \quad (15)$$

From the no-free-lunch theorem by Devroye (Devroye *et al.*, 1997) our proof given in the appendix can in general not be modified in a simple way to cover the case that the number  $B$  of bites depends on the sample size, because we have *no uniform rate* of consistency without restricting the class of probability distributions, see also Tsybakov (2004).

### 3.3 Robustness properties of RLB

Now we derive results which show that certain robustness properties are inherited from the original estimator  $T_{\mathcal{S}}$  to the RLB estimator. We will restrict attention to two robustness approaches. The finite sample breakdown point proposed by Donoho and Huber (1983) measures the worst case behavior of a statistical estimator. Then influence function proposed by Hampel (1968, 1974), measures the impact on the estimation due to an infinitesimal small contamination of the distribution  $P$  in direction of a Dirac-distribution.

**Definition 10 (Finite-sample breakdown point)** *Let  $\mathcal{S}_n = \{(x_i, y_i), i = 1, \dots, n\}$  be a data set with values in  $\mathcal{X} \times \mathcal{Y}$ . The finite-sample breakdown point of an estimator  $T_{\mathcal{S}_n}$  is defined by*

$$\varepsilon_n^*(T_{\mathcal{S}_n}) = \min \left\{ \frac{m}{n}; \text{Bias}(m; T_{\mathcal{S}_n}) \text{ is finite} \right\}, \quad (16)$$

where

$$\text{Bias}(m; T_{\mathcal{S}_n}) = \sup_{\mathcal{S}'_n} \| T_{\mathcal{S}'_n} - T_{\mathcal{S}_n} \| \quad (17)$$

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

**Theorem 11 (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  $T_{\mathcal{S}_b}$  for bite  $b$  by  $\varepsilon_{n_b}^*(T_{\mathcal{S}_b})$  and denote the finite sample breakdown point of the estimator  $\hat{\mu} = \hat{\mu}(T_{\mathcal{S}_1}, \dots, T_{\mathcal{S}_B})$*

in the aggregation step by  $\varepsilon_B^*(\hat{\mu})$ . Then the finite sample breakdown point of the RLB estimator is given by

$$\varepsilon_{RLB,S,B}^* = \frac{1}{n} \left( \sum_{b=1}^k (n_b \varepsilon_{n_b}^*(T_{S_b}) + 1)_{(b:B)} - 1 \right), \quad (18)$$

where  $k$  is the smallest integer not less than  $B\varepsilon_B^*(\hat{\mu}) + 1$  and  $z_{(1:B)} \leq \dots \leq z_{(B:B)}$  denote the ordered values of  $\{z_1, \dots, z_B\}$ .

**Remark.** If all values  $n_b \varepsilon_{n_b}^*(T_{S_b})$  are equal, we obtain

$$\varepsilon_{RLB,S,B}^* = \frac{(n_b \varepsilon_{n_b}^*(T_{S_b}) + 1) \lceil B\varepsilon_B^*(\hat{\mu}) + 1 \rceil - 1}{n} \geq \varepsilon_{n_b}^*(T_{S_b}) \varepsilon_B^*(\hat{\mu}). \quad (19)$$

If the mean or any other estimator with  $\varepsilon_B^*(\hat{\mu}) = 0$  is used in this situation, then the RLB has a finite sample breakdown point of  $\varepsilon_{n_b}^*(T_{S_b})/B \rightarrow 0$ , if  $B \rightarrow \infty$ . Hence  $B$  should not be too large.

**Example 12 (Univariate location model)** Consider the univariate location problem, where  $x_i \equiv 1$  and  $y_i \in \mathbb{R}$ ,  $i = 1, \dots, n$ ,  $n = 55$ . Assume that  $y_i \neq y_j$  for  $i \neq j$ . The finite sample breakdown point of the median is  $\lfloor n/2 \rfloor / n = 0.49$ . The mean has a finite sample breakdown point of 0. Let us investigate the robustness of RLB with  $B = 5$  and  $n_b = 11$ ,  $b = 1, \dots, B$ . (a) If the median is used as the location estimator in each bite and if the median is used in the aggregation step, then  $\varepsilon_{RLB,S_n,B}^* = 0.309$ . This value is reasonably high, but lower than the finite sample breakdown point of the median for the whole data set. Note that in a *fortunate* situation the impact of up to  $(2 \times 11 + 5 \times 3)/55 = 0.672$  extreme large data points (say with values equal to  $+\infty$ ) is still bounded for the RLB estimator in this setup: modify all data points in  $B\varepsilon_B^*(\hat{\mu}) = 2$  bites and up to  $n_b \varepsilon_{n_b}^*(T_{S_b}) = 5$  data points in the remaining  $B(1 - \varepsilon_B^*(\hat{\mu})) = 3$  bites. This is no contradiction because the breakdown point measures the *worst case* behavior. (b) If the median is used as the location estimator in each bite and if the *mean* is used in the aggregation step, then we obtain  $\varepsilon_{RLB,S_n,B}^* = (1/B)\varepsilon_{n_b}^*(T_{S_b}) = 0.09$ . (c) If the mean is used as the location estimator in each bite and also in the aggregation step we have  $\varepsilon_{RLB,S_n,B}^* = 0$ .  $\triangleleft$

Now we investigate the influence function of an RLB estimator  $T_{S,B}^{RLB} = \sum_{b=1}^B c_b T_{S_b}$  of type I with weights  $c_b \in (0, 1)$  and  $\sum_{b=1}^B c_b = 1$ .

To this end we assume the existence of a map  $T$  which assigns to every distribution  $P$  on a given set  $Z$  an element  $T(P)$  of a given Banach space  $E$  such that our RLB estimator for a data set  $S = S_1 \cup \dots \cup S_B$  has the representation

$$T_{S,B}^{RLB} = \sum_{b=1}^B c_b T(P_{S_b}). \quad (20)$$

Here  $P_{S_b}$  denotes the empirical distribution of the data points in bite  $S_b$ ,  $b = 1, \dots, B$ . We have  $T(P) = \theta \in E = \mathbb{R}^d$  for parametric models and  $E = H$  and  $T(P) = f_{P,\lambda}$  for general SVM methods defined by (6).

**Definition 13 (Influence function)** *The influence function of  $T$  at a point  $z$  for a distribution  $P$  is the special Gâteaux derivative (if it exists)*

$$IF(z; T, P) = \lim_{\varepsilon \downarrow 0} \frac{T((1 - \varepsilon)P + \varepsilon\delta_z) - T(P)}{\varepsilon}, \quad (21)$$

where  $\delta_z$  is the Dirac distribution at the point  $z$  such that  $\delta_z(\{z\}) = 1$ .

The influence function has the interpretation, that it measures the impact of an (infinitesimal) small amount of contamination of the probability distribution  $P$  in direction of a Dirac distribution located in the point  $\{z\}$  on the theoretical quantity of interest  $T(P)$ . Therefore, in the robustness approach based on influence functions it is desirable that a statistical method which can be written as  $T(P)$  has a *bounded* influence function.

**Theorem 14 (Influence function of RLB)** *Assume that the original estimator  $T_{\mathcal{S}}$  has the representation  $T(P_n)$ , where  $P_n$  is the empirical distribution of the sample  $\mathcal{S}$ , and that the influence function of the map  $T(P)$  exists for the probability distribution  $P$ . Then the RLB estimator of type I using the weights  $c_b \in (0, 1)$  and  $\sum_{b=1}^B c_b = 1$  with a fixed number  $B$  of bites and  $n/n_b \equiv B$  exists and equals the influence function of  $T(P)$ .*

Hence, if  $T(P)$  has a bounded influence function, the same is true for RLB. The influence function is one of the cornerstones of robust statistics. Many robust estimators in parametric statistical methods have a bounded influence function, see e.g. Hampel *et al.* (1986) for M-estimators and GM-estimators and Davies (1990, 1993) for  $S$ -estimators. Recently, Christmann and Steinwart (2004, 2005) showed that the influence function of many general SVM methods exists for the case of classification and regression. Further the influence function of such methods is bounded if a loss function  $L$  with bounded first derivative and a bounded and universal kernel  $k$  are used. An example is kernel logistic regression in combination with a Gaussian radial basis function kernel  $k(x, x') = \exp(-\gamma\|x - x'\|^2)$ ,  $x, x' \in \mathcal{X}$ , where  $\gamma \in (0, \infty)$ .

### 3.4 Determination of the number $B$ of bites

From the results given in Sections 3.1 to 3.3 it is obvious, that the number of bites has some impact on the statistical behavior of the RLB estimator and also on the computation time and the necessary computer memory. An optimal choice of the number  $B$  of bites will in general depend on the unknown distribution  $P$ . But some informal arguments are given how to determine  $B$  in an appropriate manner.

One should take the sample size  $n$ , the computer resources (number of CPUs, RAM) and the acceptable computation time into account. The quantity  $B$  should be much lower than  $n$ , because otherwise there is not much hope to obtain useful estimators from the bites and because the finite sample breakdown point is generally decreasing with increasing values of  $B$ . Further,  $B$  should depend on the dimensionality  $d$  of the explanatory vectors  $x_i \in \mathcal{X}$ . *E.g.* a rule of thumb for linear regression is that  $n/d$  should be at least 5. Because the function  $f$  is completely unknown in nonparametric regression assumptions on the complexity of  $f$  are crucial. The sample size  $n_b$  for each bite should converge to infinity, if  $n \rightarrow \infty$ , to obtain consistency of RLB. The results from some numerical experiments not

given here can be summarized as follows. (i) If  $B$  is too large, the computational overhead and the danger of bad fits increase because  $n_b$  is too small to provide reasonable estimators. (ii) A major decrease in computation time and memory saving is often already present, if  $B$  is chosen in a way such that each bite fits nicely into the computer (CPU, RAM). Nowadays robust estimators can often be computed for sample sizes up to  $n_b = 10^4$  or  $n_b = 10^5$ . In this case  $B = \lceil n/n_b \rceil$  can be a reasonable choice. (iii) If distribution-free confidence intervals at the  $(1 - \alpha)$  level for the median of the predictions, i.e.  $T_{S,B}^{RLB}(x) = \text{median}_{1 \leq b \leq B} T_{S_b}(x)$ ,  $x \in \mathcal{X}$ , are needed, one should take into account that the actual confidence level of such confidence intervals based on order statistics can be conservative, *i.e.* higher than the specified level, for some pairs  $(r, s)$  of order statistics due to the discreteness of order statistics. (iv) In our examples  $B = 17$  gave good results.

#### 4. Examples

In this section we give a few numerical results for RLB. We apply our proposal for a parametric and for a non-parametric method, namely robust linear regression by MM-estimation (Yohai, 1987) and kernel logistic regression (Wahba, 1999). The computations are done on a PC with a 2.8 GHz processor.

Let us begin with robust estimation in linear regression. We simulated data sets with  $n = Bn_b$  independent observations  $(x_i, y_i)$ . The explanatory variables where  $x_i = (x_{i,1}, x_{i,2}, x_{i,3})$  were independent and identically simulated from a Student distribution with 3 degrees of freedom. The responses were taken independently from the mixture model  $P = 0.8P_1 + 0.2\delta_{\{(x,y)\}}$ , where  $P_1$  denotes a Student distribution with 3 degrees of freedom and location parameter  $f(x_i) = \sum_{j=1}^3 x_{i,j}$  and  $\delta_{\{(x,y)\}}$  is a Dirac distribution in the point  $x = (50, 50, 50)$  and  $y = 1000$ . Obviously the distribution  $P$  produces approximately 20% bad leverage points in  $(x, y)$  with respect to a linear regression model with parameter vector  $\theta = (0, 1, 1, 1)$ . Here the first component of  $\theta$  is zero because the intercept term was set to zero. Further, this model contains extreme values in  $y$ -direction due to the use of a Student distribution.

Table 2 shows the computation times in seconds, the bias of an MM-estimator computed for the whole data set and of the RLB estimator for  $B = 17$ , and the width of the componentwise distribution-free confidence intervals based on the median at the 95%-level for different sub-sample sizes  $n_b$ . The MM-estimates were computed with the function `rlm` from the R-library `MASS` (Venables and Ripley, 2002). This function first computes an S-estimate as a starting point which has an approximate finite sample breakdown point of 0.5. Then an M-estimator with Tukey's biweight and fixed scale is iteratively computed using this starting value that will inherit this breakdown point from the S-estimator. The time-consuming phase of MM-estimators is the computation of the highly robust starting value. The confidence intervals for the original MM-estimator were computed due to the asymptotical normality assumption. The distribution-free confidence intervals for the RLB estimator were based on the 5<sup>th</sup> and the 13<sup>th</sup> order statistics. Because the bias terms and the width of the confidence intervals are very small due to the large sample size, the values in Table 2 are multiplied by  $10^3$ .

In the considered situations RLB gave good results: the bias values are small, which shows that the RLB method indeed gave robust estimates, and the width of the confidence

	$n_b = 10000$		$n_b = 100000$		$n_b = 200000$	
	RLB	MM	RLB	MM	RLB	MM
seconds	33.89	44.64	348.78	460.95	684.61	–
Bias( $\hat{\theta}_0$ ) ( $\times 1000$ )	2.32	0.35	0.17	0.17	0.31	–
width of c.i. ( $\times 1000$ )	17.42	15.36	5.15	4.87	5.27	–
Bias( $\hat{\theta}_1$ ) ( $\times 1000$ )	1.21	1.18	-2.02	-1.44	0.46	–
width of c.i. ( $\times 1000$ )	8.78	7.39	3.29	2.31	1.39	–
Bias( $\hat{\theta}_2$ ) ( $\times 1000$ )	0.62	0.23	0.09	-0.32	0.90	–
width of c.i. ( $\times 1000$ )	8.06	7.38	2.32	2.30	2.82	–
Bias( $\hat{\theta}_3$ ) ( $\times 1000$ )	-1.60	-2.22	0.31	-0.16	-0.54	–
width of c.i. ( $\times 1000$ )	8.72	7.36	5.19	2.28	1.86	–

Table 2: Results for robust linear regression with MM-estimator and RLB with  $B = 17$ . The computation of the MM-estimates for the whole data set with  $n = 17 \cdot 200000 = 3.4 \cdot 10^6$  data points was not possible due to memory problems.

intervals is of similar size than for the original MM-estimator. It is not surprising that the distribution-free confidence intervals for the RLB estimator are somewhat larger (often by a factor between 1.1 and 1.2) than the confidence intervals of the MM-estimator based on the assumption of asymptotic normality. If the total sample size  $n$  is not too big, such that the MM-estimates can be computed with `r1m` only using the RAM of the computer, RLB only saves a little bit of computation time. However, one can fit much larger data sets using RLB for which the algorithm used by `r1m` would need much more RAM than the available PC has (2 GB), such that the computation of the MM-estimates for the whole data set was impossible. In contrast to that, the computation time of RLB increased only approximately linearly in  $n_b$ , and the used RAM was low in contrast to the used RAM to compute the MM-estimates for the whole data set. No memory problems occurred for RLB with  $n = 3.4 \cdot 10^6$  and  $B = 17$ .

Now we apply the RLB approach to kernel logistic regression (KLR), see (Wahba, 1999). KLR is a flexible method for classification problems and provides also estimates for the conditional probabilities  $P(Y = 1|X = x)$ ,  $x \in \mathcal{X}$ , which is not true for the support vector machine, see Bartlett and Tewari (2004). Christmann and Steinwart (2004) showed KLR has good robustness properties, e.g. a bounded influence function. All computations are done with the program `myKLR` (Rüping, 2003) which is an implementation in C++ of the algorithm proposed by Keerthi *et al.* (2004) to solve the dual problem. We choose KLR for two reasons. Firstly, the computation of KLR needs much more time than for the support vector machine because the latter solves a quadratic instead of a convex program in dual space. Therefore, the need for computational improvements is greater for KLR than say for the SVM, and the potential gains of RLB can be more important. Secondly, the number of support vectors of KLR is approximately equal to  $n$  which slows down the computation of predictions.

sample size $n$	CPU time	used cache in MB	available cache in MB
2000	4 sec	33	200
5000	25 sec	198	200
10000	5 min, 21 sec	200	200
10000	1 min, 33 sec	787	1000
20000	24 min, 11 sec	1000	1000
20000	14 min, 35 sec	1000	1000
100000	9 h, 56 min, 46 sec	1000	1000

Table 3: Computation times for kernel logistic regression using myKLR.

The simulated data sets contain  $n$  data points  $(x_i, y_i) \in \mathbb{R}^8 \times \{-1, +1\}$  simulated in the following way. All 8 components of  $x_i = (x_{i,1}, \dots, x_{i,8})$  are simulated independently from a uniform distribution on  $(0, 1)$ . The responses  $y_i$  are simulated independently from a logistic regression model according to  $P(Y_i = +1 | X_i = x_i) = 1/(1 + \exp[-f(x_i)])$ . We define

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

such that there are 8 main effects and 4 interaction terms. The data points are saved as ASCII files where  $x_{i,j}$  is stored with four decimal places. The numerical results of fitting kernel logistic regression to such data sets is given in Table 3. It is obvious that in this situation RLB can save a lot of computation time. If the whole data set has  $n = 10^5$  observations, approximately 10 hours were needed to compute KLR. If RLB with  $B = 10$  bits are used each with a sub-sample size of  $n_b = 10^4$ , one needs approximately  $10 \times 93.3$  seconds, *i.e.* 16 minutes, 1 GB of kernel cache available. This is a reduction by a factor of 38. If there are 5 CPUs available and each processor can use up to 200 MB kernel cache, RLB with  $B = 10$  will need approximately 11 minutes which is a reduction by a factor of 55.

Christmann (2005) describes a strategy combining  $\varepsilon$ -support vector regression and kernel logistic regression to construct insurance tariffs. The whole data from 15 German motor vehicle insurance companies contains data from around 4.6 million customers with dozens of explanatory variables. A direct use of kernel logistic regression with myKLR is unfeasible due to the computation time, see Table 3. Although a strategy was used to reduce the computational effort by exploiting certain characteristic features of such data sets, RLB offers an additional substantial reduction of the computation time. Fitting the model to the whole data set would need more than six months on a PC with 2.8 GHz CPU, whereas RLB with  $B = 17$  using 2 CPUs was able to provide a good fit within 4 days: this is a reduction by a factor of 45. If RLB is allowed to use 8 CPUs the computation can be done in approximately one day and the reduction factor is around 180.

Concluding, RLB proved to be quite useful for kernel logistic regression for large data sets.

## 5. Discussion

In this paper robust learning from bites (RLB) was proposed to broaden the usability of computer-intensive robust estimators in the case of large data sets which occur nowadays often in data mining projects. RLB is especially designed for situations under which the original robust method cannot be used due to excessive computation time or due to memory space problems. In these situations RLB offers robust estimates and additionally robust confidence intervals. RLB estimators will in general not fulfill certain optimality criteria, but the method has four nice properties. *Scalability*: the number  $B$  of bites can be chosen such that the algorithm used to fit the bites needs less memory than the computer offers. *Performance*: the computational steps for different bites can easily be distributed on several processors because they are independent and use disjoint parts of the data set. *Robustness*: we considered the finite sample breakdown point and the influence function. These properties are inherited from the original robust estimator computed for each bite and from the location estimator used to aggregate the results from the bites. *Confidence intervals*: no complex formulae are needed to obtain distribution-free (componentwise) confidence intervals for the estimates or for the predictions if the median is used in the aggregation step because the estimators computed from the  $B$  bites are independent and identically distributed. Such confidence intervals for the predictions are especially interesting for general SVM methods (*e.g.* support vector machines and kernel logistic regression), because such methods have nice properties but finite sample confidence intervals for the predictions based on applying such methods once for the whole data set are typically unknown.

The subsampling approach used by RLB has connections to the remedian proposed by Rousseeuw and Bassett (1990) for univariate location estimation. The remedian with base  $B$  computes medians of groups of  $n_b$  observations, and then the medians of these medians etc., until only a single estimate remains. The remedian needs only  $O(\log(n))$  total storage for fixed  $B$  which makes it especially useful for robust estimation in large data bases, for real-time engineering applications in which the data are not present at the same time and perhaps not stored, and for resistant aggregation of curves. RLB has also similarities to Rvote proposed by Breiman (1999) and DRvote with classification trees using majority voting proposed by Chawla *et al.* (2004). Bootstrapping computer-intensive robust methods for huge data sets is often impossible due to computation time and memory limitations of the computer. The focus of the present paper was on robustness aspects and the computation of robust distribution-free confidence intervals for the median of the predictions even for very large data sets. Such confidence intervals are often a problem for robust estimators and general SVM methods based on Vapnik's convex risk minimization principle. These topics were not covered in the papers mentioned above. RLB has also some similarity to the algorithms FAST-LTS and FAST-MCD developed by Rousseeuw and Driessen (1999, 2000) for robust estimation in linear regression or multivariate location and scatter models for large data sets. FAST-LTS and FAST-MCD split the data set into sub-samples, optimize the objective function in each sub-sample, and use these solutions as starting values to optimize the objective function for the whole data set. This is in contrast to RLB which aggregates estimation results from the bites to obtain robust confidence intervals.

Some good robust estimators are not  $n^{-1/2}$ -consistent having a complicated non-normal limiting distribution or are hard to compute for large data sets, see *e.g.* Rousseeuw (1984),



Davies (1990), Kim and Pollard (1990), Rousseeuw and Hubert (1999), Bai and He (1999), Aelst *et al.* (2002), Croux *et al.* (2003), and Zuo and Cui (2005). Then RLB can be useful if distribution-free confidence intervals for the median of the predictions are needed for large data sets.

## Acknowledgments

We thank Claudio Agostinello and Matias Salibian-Barrera for helpful discussions during ICORS-2005. The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") and of the Forschungsband Do-MuS (University of Dortmund) is gratefully acknowledged.

## Appendix

The appendix contains the proofs for the results given in Section 3.

**Proof of Proposition 2.** Obvious.  $\square$

**Proof of Proposition 3.** Obvious.  $\square$

**Proof of Proposition 4.** (i) follows from the linearity of the expectation operator. (ii) and (iii) follow from Slutsky's theorem.  $\square$

**Proof of Proposition 5.** By construction of RLB the bites are disjoint and the estimators from the bites are independent. Assume that the original estimator  $T_{\mathcal{S}}$  is consistent in probability for  $T_{\mathbb{P}}$ . Then we have for all  $\varepsilon > 0$  that

$$\begin{aligned} & \mathbb{P}(|\text{median}_{b=1,\dots,B} T_{\mathcal{S}_b}(x) - T_{\mathbb{P}}(x)| < \varepsilon) \\ & \geq \mathbb{P}(|T_{\mathcal{S}_b}(x) - T_{\mathbb{P}}(x)| < \varepsilon \text{ for all } b = 1, \dots, B) \\ & = \prod_{b=1}^B \mathbb{P}(|T_{\mathcal{S}_b}(x) - T_{\mathbb{P}}(x)| < \varepsilon) \rightarrow 1, \quad n \rightarrow \infty, \quad x \in \mathcal{X}, \end{aligned}$$

because  $B$  is fixed and  $\lim_{n \rightarrow \infty} (n/n_b) = B$ . Now, assume that the original estimator  $T_{\mathcal{S}}$  is strongly consistent to  $T_{\mathbb{P}}$ . Then we obtain analogously

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \text{median}_{b=1,\dots,B} T_{\mathcal{S}_b}(x) = T_{\mathbb{P}}(x)\right) \geq \prod_{b=1}^B \mathbb{P}\left(\lim_{n_b \rightarrow \infty} T_{\mathcal{S}_b}(x) = T_{\mathbb{P}}(x)\right) = 1, \quad x \in \mathcal{X},$$

because  $B$  is fixed and  $\lim_{n \rightarrow \infty} (n/n_b) = B$ .  $\square$

**Proof of Theorem 6.** By assumption each bite  $\mathcal{S}_b$  is fitted with a general SVM estimator having the representation

$$f_{\mathcal{S}_b, \lambda_{n_b}}(x) = \sum_{i=1}^{n_b} \alpha_{i,b} k(x, x_i), \quad i \in \mathcal{S}_b, \quad b = 1, \dots, B, \quad x_i \in \mathcal{X}. \quad (22)$$

Because the bites  $\mathcal{S}_b$ ,  $b = 1, \dots, B$ , are disjoint, the RLB estimator of type I using a convex combination in the aggregation step is given by

$$f_{\mathcal{S},B,(\lambda_{n_b})}^{RLB}(x) = \sum_{b=1}^B c_b \sum_{i=1}^{n_b} \alpha_{i,b} k(x, x_i) \quad (23)$$

$$= \sum_{i=1}^n \sum_{b=1}^B c_b \alpha_{i,b} k(x, x_i), \quad x \in \mathcal{X}. \quad (24)$$

The formula (10) follows immediately.  $\square$

**Proof of Proposition 8.** (i) This follows immediately from (10).

(ii) Steinwart (2003, Th.9) proved that the general SVM estimator evaluated for the whole data set  $\mathcal{S}$  has the property

$$\Pr^{*n}(\mathcal{S} \in (\mathcal{X} \times \mathcal{Y})^n; \#SV(f_{\mathcal{S},\lambda_n}) \geq (S_{L,P} - \varepsilon)n) \rightarrow 1, \quad n \rightarrow \infty, \quad (25)$$

if the conditions of the proposition are satisfied. Denote the outer probability measure of the product measure  $\mathbb{P}^{b,n_b}$  by  $\Pr^{*b,n_b}$ . The pairs  $(X_i, Y_i)$  in the bites  $\mathcal{S}_b$ ,  $b = 1, \dots, B$ , are independent and identically distributed by construction of RLB. Using (25) and  $n \equiv Bn_b$  we obtain

$$\Pr^{*n}(\mathcal{S} = (\mathcal{S}_1, \dots, \mathcal{S}_B) \in (\mathcal{X} \times \mathcal{Y})^n; \#SV(f_{\mathcal{S},B,(\lambda_{n_b})}^{RLB}) \geq (S_{L,P} - \varepsilon)n) \quad (26)$$

$$= \Pr^{*n}(\mathcal{S} = (\mathcal{S}_1, \dots, \mathcal{S}_B) \in (\mathcal{X} \times \mathcal{Y})^n; \#SV(f_{\mathcal{S}_b, \lambda_{n_b}}) \geq \sum_{b=1}^B (S_{L,P} - \varepsilon)n_b) \quad (27)$$

$$\geq \Pr^{*n}(\text{for all } \mathcal{S}_b \in (\mathcal{X} \times \mathcal{Y})^{n_b}, b = 1, \dots, B; \#SV(f_{\mathcal{S}_b, \lambda_{n_b}}) \geq (S_{L,P} - \varepsilon)n_b) \quad (28)$$

$$= \prod_{b=1}^B \Pr^{*b,n_b}(\mathcal{S}_b \in (\mathcal{X} \times \mathcal{Y})^{n_b}; \#SV(f_{\mathcal{S}_b, \lambda_{n_b}}) \geq (S_{L,P} - \varepsilon)n_b) \rightarrow 1, \quad n \rightarrow \infty, \quad (29)$$

because  $c_b \in (0, 1)$ ,  $b = 1, \dots, B$ ,  $B$  is fixed, and  $n_b \rightarrow \infty$ .  $\square$

**Proof of Theorem 9.** The RLB estimator  $f_{\mathcal{S},B,(\lambda_{n_b})}^{RLB}$  of type I is a convex combination of  $f_{\mathcal{S}_b, \lambda_{n_b}}$ ,  $b = 1, \dots, B$ , because  $c_b \in (0, 1)$  and  $\sum_{b=1}^B c_b = 1$ . Therefore,

$$\begin{aligned} 0 &\leq \int L(Y, f_{\mathcal{S},B,(\lambda_{n_b})}^{RLB}(X)) d\mathbb{P} - \mathcal{R}_{L,P}^* \\ &= \int L\left(Y, \sum_{b=1}^B c_b f_{\mathcal{S}_b, \lambda_{n_b}}(X)\right) d\mathbb{P} - \mathcal{R}_{L,P}^* \\ &\leq \int \sum_{b=1}^B c_b L(Y, f_{\mathcal{S}_b, \lambda_{n_b}}(X)) d\mathbb{P} - \mathcal{R}_{L,P}^* \end{aligned} \quad (30)$$

$$= \sum_{b=1}^B c_b \left[ \int L(Y, f_{\mathcal{S}_b, \lambda_{n_b}}(X)) d\mathbb{P} - \mathcal{R}_{L,P}^* \right] \xrightarrow{\mathbb{P}} 0, \quad (31)$$

if  $\lim_{n \rightarrow \infty} \min_{1 \leq b \leq B} n_b \rightarrow \infty$ . Here we used the convexity of  $L$  in (30), the  $L$ -risk consistency of the original estimator in (31), and (14).  $\square$

**Proof of Theorem 11.** The minimum number of points needed to modify  $T_{S_b}$  in bite  $b$  such that there is breakdown is given by  $n_b \cdot \varepsilon_{n_b}^*(T_{S_b}) + 1$ ,  $b = 1, \dots, B$ . The RLB estimator breaks down if at least  $B\varepsilon_B^*(\hat{\mu}) + 1$  of the estimators  $T_{S_1}, \dots, T_{S_B}$  break down. This gives the assertion.  $\square$

**Proof of Theorem 14.** Let  $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$  and  $P$  be a probability distribution on  $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$ . By assumption the RLB estimator has the property (20), i.e.  $T_{S,B}^{RLB} = \sum_{b=1}^B c_b T(P_{S_b})$ , where  $P_{S_b}$  denotes the empirical distribution of bite  $S_b$ ,  $b = 1, \dots, B$ . Further, the influence function  $\text{IF}(z; T, P)$  exists by assumption of the theorem. It follows

$$\begin{aligned} \text{IF}(z; T_B^{RLB}, P) &= \lim_{\varepsilon \downarrow 0} \frac{T_B^{RLB}((1 - \varepsilon)P + \varepsilon\delta_z) - T_B^{RLB}(P)}{\varepsilon} \\ &= \lim_{\varepsilon \downarrow 0} \frac{\sum_{b=1}^B c_b T((1 - \varepsilon)P + \varepsilon\delta_z) - \sum_{b=1}^B c_b T(P)}{\varepsilon} \\ &= \sum_{b=1}^B c_b \lim_{\varepsilon \downarrow 0} \frac{T((1 - \varepsilon)P + \varepsilon\delta_z) - T(P)}{\varepsilon} \\ &= \text{IF}(z; T, P), \end{aligned}$$

which gives the assertion.  $\square$

## References

- Aelst, S. V., Rousseeuw, P. J., Hubert, M., and Struyf, A. (2002). The deepest regression method. *Journal of Multivariate Analysis*, **81**, 138–166.
- Bai, Z. D. and He, X. (1999). Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *Annals of Statistics*, **27**, 1616–1637.
- Bartlett, P. L. and Tewari, A. (2004). Sparseness versus estimating conditional probabilities: Some asymptotic results. In *Proceedings of the 17th Annual Conference on Learning Theory, Vol. 3120*, Lecture Notes in Computer Science, Vol. 3120, pages 564–578, New York. Springer.
- Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, to appear. <http://stat-www.berkeley.edu/tech-reports/638.pdf>.
- Breiman, L. (1999). Pasting bites together for prediction in large data sets. *Machine Learning*, **36**, 85–103.
- Chawla, N. V., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. (2004). Learning ensembles for bites: a scalable and accurate approach. *Journal of Machine Learning Research*, **5**, 421–451.

- Christmann, A. (2005). On a strategy to develop robust and simple tariffs from motor vehicle insurance data. *Acta Mathematicae Applicatae Sinica (English Series, Springer)*, **21**, 193–208.
- Christmann, A. and Steinwart, I. (2004). On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, **5**, 1007–1034.
- Christmann, A. and Steinwart, I. (2005). Consistency and robustness of kernel based regression. University of Dortmund, SFB-475, TR-01/05. Submitted. .
- Croux, C., Van Aelst, S., and Dehon, C. (2003). Bounded influence regression using high breakdown scatter matrices. *Annals of the Institute of Statistical Mathematics*, **55**, 265–285.
- David, H. A. and Nagaraja, H. N. (2003). *Order Statistics, 3rd ed.* . Wiley & Sons, New York.
- Davies, P. L. (1990). The asymptotics of S-estimators in the linear regression model. *Annals of Statistics*, **18**, 1651–1675.
- Davies, P. L. (1993). Aspects of robust linear regression. *Annals of Statistics*, **21**, 1843–1899.
- Devroye, L., Györfi, L., and Lugosi, G. (1997). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.
- Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In P. J. Bickel, K. A. Doksum, and J. L. Hodges, editors, *A Festschrift for Erich L. Lehmann*, pages 157–184. Jr. , Belmont, California, Wadsworth.
- Hampel, F. R. (1968). *Contributions to the theory of robust estimation*. Ph. D. thesis, Dept. Statistics, Univ. California, Berkeley.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383–393.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: The Approach Based on Influence Functions*. Wiley, New York.
- Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. MIT Press, Cambridge, Massachusetts.
- Hipp, J., Güntzer, U., and Grimmer, U. (2001). Data quality mining - making a virtue of necessity. Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD, Santa Barbara, CA, [http://www.cs.cornell.edu/johannes/papers/dmkd2001-papers/p5\\_hipp.pdf](http://www.cs.cornell.edu/johannes/papers/dmkd2001-papers/p5_hipp.pdf).
- Keerthi, S. S., Duan, K., Shevade, S. K., and Poo, A. N. (2004). A fast dual algorithm for kernel logistic regression. In *Machine Learning: Proceedings of the Nineteenth International Conference (Eds: C. Sammut, A. G. Hoffmann)*, pages 299–306. Kaufmann, San Francisco.

- Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, **18**, 191–219.
- Le Cam, L. (1980). In: Discussion of "minimum chi-square, not maximum likelihood!" *Annals of Statistics*, **8**, 473–478.
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871–880.
- Rousseeuw, P. J. and Bassett, G. W. (1990). The remedian: a robust averaging method for large data sets. *Journal of the American Statistical Association*, **85**(409), 97–104.
- Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.
- Rousseeuw, P. J. and Driessen, K. V. (2000). An algorithm for positive-breakdown regression based on concentration steps. In O. O. W. Gaul and M. Schader, editors, *Data Analysis: Scientific Modeling and Practical Application*, pages 335–346, New York. Springer.
- Rousseeuw, P. J. and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, **94**, 388–402.
- Rüping, S. (2003). myKLR - kernel logistic regression. University of Dortmund, Department of Computer Science, <http://www-ai.cs.uni-dortmund.de/SOFTWARE>.
- Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts.
- Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, **2**, 67–93.
- Steinwart, I. (2002). Support vector machines are universally consistent. *J. Complexity*, **18**, 768–791.
- Steinwart, I. (2003). Sparseness of support vector machines. *Journal of Machine Learning Research*, **4**, 1071–1105.
- Steinwart, I. (2005a). Consistency of support vector machines and other regularized kernel machines. *IEEE Transactions on Information Theory*, **51**, 128–142.
- Steinwart, I. (2005b). How to compare loss functions and their risks. Technical report, Los Alamos National Laboratory. <http://www.c3.lanl.gov/ingo/pubs.shtml>.
- Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, **32**, 135–166.
- Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York.
- Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S, 4<sup>th</sup>ed.* . Springer.

- Wahba, G. (1999). Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 69–88, Cambridge, MA. MIT Press.
- Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, **15**, 642–656.
- Zhang, T. (2004). Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**, 56–134.
- Zuo, Y. and Cui, H. (2005). Depth weighted scatter estimators. *Annals of Statistics*, **33**(1), 381–413.

ANDREAS CHRISTMANN  
UNIVERSITY OF DORTMUND  
DEPARTMENT OF STATISTICS  
44221 DORTMUND  
GERMANY  
E-MAIL:  
christmann@statistik.uni-dortmund.de

INGO STEINWART  
CCS-3  
MAIL STOP B256  
LOS ALAMOS NATIONAL LABORATORY  
LOS ALAMOS, NM 87545  
USA  
E-MAIL: ingo@lanl.gov

MIA HUBERT  
DEPARTMENT OF MATHEMATICS  
KATHOLIEKE UNIVERSITEIT LEUVEN  
W. DE CROYLAAN 54  
B-3001 LEUVEN  
BELGIUM  
E-MAIL: Mia.Hubert@wis.kuleuven.ac.be