Robust Learning from Bites

By Andreas Christmann[1]

University of Dortmund, Department of Statistics
christmann@statistik.uni-dortmund.de

Many robust statistical procedures have two drawbacks. Firstly, they are computer-intensive such that they can hardly be used for massive data sets. Secondly, robust confidence intervals for the estimated parameters or robust predictions according to the fitted models are often unknown. Here, we propose a general method to overcome these problems of robust estimation 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. The method additionally offers distribution-free confidence intervals for the median of the predictions. The method is illustrated for two situations: robust estimation in linear regression and kernel logistic regression from statistical machine learning.

1. Introduction

Data sets with millions of observations occur nowadays in many areas. An insurance company or a bank collects many variables to develop tariffs and scoring methods for credit risk management, respectively. Other examples are data mining projects and micro-arrays. For such data sets parametric assumptions are often violated, outliers are present, or some variables can only be measured in an imprecise manner. The application of robust statistical methods is important in such situations. However, many robust methods have the following drawbacks which are serious limitations for the application of robust methods. (a) They are computer-intensive such that they can hardly be used for massive data sets, say for several millions of observations with many 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 goal of the proposal is to broaden in application of robust methods for massive data. The idea is to split the huge data set S by random into disjoint subsets S_b, $b = 1, \ldots, B$. Then the robust method is applied to each subset, and the results are summarized in a robust manner. The proposal yields robust predictions for the median together with distribution-free confidence intervals. The method is scalable to the memory of the computer by choosing B appropriately and the computation can easily distributed on several processors which helps to reduce the computation time substantially.

1. Address: University of Dortmund, Department of Statistics, 44221 Dortmund, GERMANY.
Supported in part by the Deutsche Forschungsgemeinschaft (SFB 475), and DoMuS.
2. *AMS 2000 subject classification. Primary 62G08, 62G35; secondary 68Q32, 62G20.*
3. *Keywords and Phrases.* Convex risk minimization, distributed computing, finite sample breakdown point, influence function, logistic regression, robustness, scalability, statistical machine learning.

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 a discussion. 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. The goal of the proposal is to make robust estimation methods usable for data sets which are too large for currently available algorithms due to memory or time limitations.

The goal of regression and classification is to estimate an approximated functional relationship between an observation random variable X and a response random variable Y using n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ 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 $\mathrm{P}(\,\cdot\,|x)$, $x \in \mathcal{X}$, on $\mathcal{Y}$. For the case of binary classification we have $\mathcal{Y} = \{-1, +1\}$.

Under the classical signal plus noise assumption $Y_i|(X = x_i)$ is distributed as $f(x_i) + \varepsilon_i$, where f is an unknown function and ε_i are independent and identically distributed error terms, $1 \leq i \leq n$. In the parametric setup we have $f = f_\theta$, $\theta \in \Theta \subset \mathbb{R}^d$. In the non-parametric setup f belongs to some functional subspace $\mathcal{H}$ of all measurable functions $f : \mathcal{X} \to \mathbb{R}$. The main goals are to obtain a robust estimator $\hat{f}$ and good predictions $\hat{f}(x)$.

In this paper we always assume that the sample size n is large. Instead of modelling the full training data set, we split the training data set by random into $B \geq 1$ parts (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 the (componentwise) median or the mean of the estimators $T_{n_b}(\mathcal{S}_b) = \hat{f}_b$ and summarize the predictions by the median of $\hat{f}_b(x)$ from the B fitted models. Table 1 gives the three main steps of the procedure which we call robust learning from bites (RLB). For classification problems the median can also be computed for the predicted event probabilities $\mathrm{P}(Y_i = 1|X = x_i)$. Of course other robust estimators can be used instead of the median, e.g. M-estimators, S-estimators or Hodges-Lehmann-type estimators. In Section 3 it will be shown that the mean instead of the median can yield estimators with bad robustness properties even if the estimators computed in each bite are highly robust.

If B is large enough, say around 20, precision estimates can additionally be obtained by computing standard deviations of the predictions $\hat{f}_{RLB,n,B}(x)$ using the central limit theorem. However, in general we favor an alternative distribution-free method based on the median. If B is small or if it is unknown whether $T_{RLB,n,B}(x)$ has a finite variance, one can construct distribution-free confidence intervals for the median of $T_{RLB,n,B}(x)$ and distribution-free tolerance regions based on order statistics, see David and Nagaraja (2003, Chap. 7). For e.g. if $B = 17$, the 5^{th} and the 12^{th} order statistics yield a distribution-free confidence interval at the 95% level for the median without any distributional assumption. Table 2 lists some values of B, the corresponding pair of order statistics determining

Step 1: construct bites.
Split data set S by random into B disjoint subsets S_b of sample sizes $n_b \approx n/B$, $b = 1, \ldots, B$.
Step 2: fit bites.
for $(b = 1, \ldots, B)$
{ Compute the robust estimator T_{n_b} based on bite S_b. }
Step 3: aggregate estimators and predictions by the median.
Compute $T_{RLB,n,B} = \text{median}_{1 \le b \le B} T_{n_b}$ (componentwise) and $T_{RLB,n,B}(x_i) = \text{median}_{1 \le b \le B} T_{n_b}(x_i)$ for all $x_i \in S$.
Compute distribution-free $(1 - \alpha)$ confidence intervals for the median based on the pair (r, s) of order statistics, i.e. $[T_{RLB,n,(r:B)}(x), T_{RLB,n,(s:B)}(x)]$.
Step 3': aggregate estimators and predictions by the mean.
Compute $T_{RLB,n,B} = \frac{1}{B} \sum_1^B T_{n_b}$ and $T_{RLB,n,B}(x_i) = \frac{1}{B} \sum_1^B T_{n_b}(x_i)$ for all $x_i \in S$.

Table 1: Principle of RLB.

the confidence interval $[T_{RLB,n,(r:B)}(x), T_{RLB,n,(s:B)}(x)]$, the finite sample breakdown point of the confidence interval, and the lower bound of the actual confidence level which is $0.5^B \sum_{j=r}^{s} \binom{B}{j}$. In Section 3 it will be shown that RLB using the median in the aggregation step offers also nice robustness properties. The actual confidence intervals can be conservative of small choices of B, see Table 2. The last column in Table 2 gives the finite sample breakdown point for the distribution-free confidence interval for the median. 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^{th} and the 13^{th} order statistics give a confidence interval at the level 95% for the median. Because the results of the four lowest and the four highest predictions are not considered, the breakdown point of this confidence interval is $4/17 = 0.235$.

If the robust estimator is based on hyper-parameters and if their values must be determined from the data set itself, a common approach is to split huge data sets into three parts. A description of RLB in this case is given in Table 3. As usual the validation data set is used to determine good values for the hyper-parameters by optimizing an appropriate goodness-of-fit criterion or by minimizing the generalization error. Finally, the test data set is used to estimate the goodness-of-fit criterion or the generalization error for new data points.

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 kernel based 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.

confidence level $1-\alpha$	B	r	s	lower bound of actual confidence level	finite sample breakdown point $\min\{r-1, B-s\}/B$
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 2: Selected pairs (r, s) of order statistics for non-parametric confidence intervals of the median.

3.1 General properties

By construction of RLB the estimators T_{n_b} from the B bites are independent and computed from disjoint parts of the data set. The computation time and the memory space for RLB

Step 1: split data set.
Split data set by random into training data set, validation data set, and test data set.
Step 2: construct bites.
Split training data set by random into B disjoint subsets S_b of sample sizes $n_b \approx n/B$, $b = 1, \ldots, B$.
Step 3: fit bites.
for $(b = 1, \ldots, B)$
{ Compute the robust estimator T_{n_b} based on bite S_b . }
Step 4: aggregate estimators and predictions.
Compute $T_{RLB,n,B} = \text{median}_{1 \le b \le B} T_{n_b}$ (componentwise) and $T_{RLB,n,B}(x_i) = \text{median}_{1 \le b \le B} T_{n_b}(x_i)$ for all $x_i \in S$.
[analogous for the mean]
Step 5: Validation and testing.
Use the validation data set to optimize the hyper-parameters.
Use the test data set to measure the overall behavior of the method.

Table 3: Principle of RLB in the case of hyper-parameters.

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.

Proposition 1 (Computation time, k CPUs) *Assume that the computation time of the estimator T_n for a data set with $n = Bn_b$ observations and d explanatory variables is of order $O(g(n,d))$, where g is some positive function. Then the computation time of RLB with B bites for the same data set is approximately of order $O(k_B \cdot g(n/B, d))$.*

Proposition 2 (Memory space, k CPUs) *Assume that the estimator T_n for a data set with $n = Bn_b$ observations and d explanatory variables needs memory space and hard disk space of order $O(g_1(n,d))$ and $O(g_2(n,d))$, respectively, where g_1 and g_2 are positive functions. Then the computation of RLB with B bites for the same data set needs approximately memory space and hard disk space of order $O(k \cdot g_1(n/B, d))$ and $O(k \cdot g_2(n/B, d))$, respectively.*

Proposition 3 (Consistency) *Consider RLB where the* mean *is used in the aggregation step. Denote the estimator based on the whole data set by T_n 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 $T_{RLB,n,B}$.*
(i) If $\mathrm{E}(T_b) = \mathrm{E}(T_n)$ for all $b \in \{1, \ldots, B\}$, then $\mathrm{E}(T_n) = \mathrm{E}(T_{RLB,n,B})$.
(ii) If T_n converges in probability (or almost sure) to $T(\mathrm{P})$ for $n \to \infty$ and if $(n/n_b) \to B$, B fixed, then $T_{RLB,n,B}$ converges in probability (or almost sure) to $T(\mathrm{P})$.
(iii) Assume that $n_b^{1/2}(T_{n_b} - T(\mathrm{P}))$ converges in distribution to a multivariate normal distribution $N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite, and that $(n/n_b) \to B$, $1 \le b \le B$, B fixed. Then $n^{1/2}(T_{RLB,n,B} - T(\mathrm{P}))$ converges in distribution to a multivariate normal distribution $N(0, \Sigma)$, $n \to \infty$.

Proposition 4 (Consistency) *Consider RLB where the* median *is used in the aggregation step. Denote the estimator based on the whole data set by T_n and denote the corresponding RLB estimator based on B bites by $T_{RLB,n,B}$. If T_n converges in probability (or almost sure) to $T(\mathrm{P})$ and if $\lim_{n\to\infty}(n/n_b) \equiv B$, B fixed, then $T_{RLB,n,B}$ converges in probability (or almost sure) to $T(\mathrm{P})$.*

3.2 Properties of RLB using the mean for kernel based methods

Now we consider kernel based estimators for f given by

$$\arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda||f||_{\mathcal{H}}^2\,, \tag{1}$$

where $L : \mathcal{Y}\times[0,\infty)$ is a convex loss function and $\mathcal{H}$ is the reproducing kernel Hilbert space defined via the kernel k, *cf.* Vapnik (1998) and Schölkopf and Smola (2002). The kernel based estimator $\hat{f}_b(x)$, $x\in\mathcal{X}$, defined as the solution of (1) for bite $\mathcal{S}_b$ can be written as

$$\hat{f}_{n_b}(x) = \sum_{i=1}^{n_b} \alpha_{i,b}\, k(x, x_i), \quad i\in\mathcal{S}_b,\ b=1,\ldots,B,\ x\in\mathcal{X}, \tag{2}$$

where $\alpha_{i,b}\in\mathbb{R}$. If $\alpha_{i,b}\neq 0$, then (x_i, y_i) is called a support vector (SV). Special cases of such kernel based methods are the support vector machine, support vector regression, and kernel logistic regression. Obviously, the minimization problem (1) can be interpreted as a stochastic approximation of the minimization of the theoretical regularized risk

$$f_{\mathrm{P},\lambda} := \arg\min_{f\in\mathcal{H}} \mathrm{E}_{\mathrm{P}}\, L(Y, f(X)) + \lambda\|f\|_{\mathcal{H}}^2\,. \tag{3}$$

Theorem 5 (RLB for kernel based methods) *Assume that the estimator $\hat{f}_n$ for the whole data set with $n = B\cdot n_b$ observations, B fixed, and d explanatory variables is a kernel based estimator based on (1). Assume that the* mean *is used in the aggregation step of RLB. Then the RLB estimator is a kernel based estimator and can be written as*

$$\hat{f}_{RLB,n,B}(x) = \sum_{i=1}^{n} \alpha_{i,RLB}\, k(x, x_i) \tag{4}$$

$$= \sum_{i\,\in\, SV(\mathcal{S}_1)\,\cup\,\ldots\,\cup\, SV(\mathcal{S}_B)} \alpha_{i,RLB}\, k(x, x_i), \quad x\in\mathcal{X}, \tag{5}$$

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

Now we investigate the number of support vectors of RLB. For part (ii) of our next result we need the quantity

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

see Steinwart (2003, p.1082). Here P_X denotes the marginal distribution of X, $X_0 := \{x\in\mathcal{X}; \mathrm{P}(1|X=x) = 1/2\}$, $X_{cont} := \{x\in\mathcal{X}; \mathrm{P}_X(\{x\}=0\}$, and $\partial_2 L$ denotes the subdifferential

operator of the loss function L with respect to the second variable. Further, F_L^* denotes the set-valued function given by

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

Theorem 6 (Number of support vectors) *Under the assumptions of Theorem 5 the RLB estimator using the* mean *in the aggregation step has the following properties.*
(i) The number of support vectors, i.e. $\alpha_{i,RLB} \neq 0$, of the RLB estimator is given by

$$\#\left\{SV(\mathcal{S}_1) \cup \ldots \cup SV(\mathcal{S}_B)\right\}. \tag{6}$$

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

$$\Pr^{*n}\left(\mathcal{S}_1 \cup \ldots \cup \mathcal{S}_B \in (\mathcal{X}\times\mathcal{Y})^n; \#SV(\hat{f}_{RLB,n,B}) \geq \sum_{b=1}^{B}(S_{L,\mathrm{P}} - \varepsilon)n_b\right) \to 1\,. \tag{7}$$

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

Part (ii) of the above result was proved for $B = 1$ by Steinwart (2003). It has the following interpretation: with probability tending to 1 when the total sample size $n = Bn_b \to \infty$, but B is fixed, the fraction of support vectors of the kernel based RLB estimator $\hat{f}_{RLB,n,B}(x)$ is essentially greater than the average of the Bayes risks for the bites.

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

$$\mathcal{R}_{L,\mathrm{P}}(f) := \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x))\, d\mathrm{P}(x,y) = \int_{\mathcal{X}}\int_{\mathcal{Y}} L(y,f(x))\,\mathrm{P}(dy|x)\,\mathrm{P}_X(dx).$$

The above integral is always defined since L is non-negative and continuous, although it may be infinite. Consider a kernel based estimator $\hat{f}_{n,\lambda_n}$ for the whole data set $S = S_n$. The estimator $\hat{f}_{n,\lambda_n}$ is called *L-risk consistent*, if

$$\mathcal{R}_{L,\mathrm{P}}(\hat{f}_{n,\lambda_n}) \to \mathcal{R}_{L,\mathrm{P}} := \inf\left\{\mathcal{R}_{L,\mathrm{P}}(f)\,;\, f : X \to \mathbb{R} \text{ measurable}\right\} \tag{8}$$

holds in probability for $n \to \infty$ for suitable chosen regularization sequences (λ_n). Of course, such convergence can only hold if the used RKHS is rich. One way of describing the richness of $\mathcal{H}$ is the following definition taken from Steinwart (2001).

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

Several authors have given conditions to guarantee that kernel based estimators are $L-$risk consistent, *cf.* Steinwart (2002, 2005) and Zhang (2004) for classification.

If $\hat{f}_{n,\lambda_n}$ is L-risk consistent, $B \geq 1$ fixed, and $n/n_b \to B$ for $n \to \infty$, we obtain by Slutzky's theorem

$$\frac{1}{B}\sum_{b=1}^{B} \mathcal{R}_{L,\mathrm{P}}(\hat{f}_{n_b,\lambda_{n_b}}) \to \mathcal{R}_{L,\mathrm{P}} \tag{9}$$

in probability for $n \to \infty$.

The next result gives L-risk consistency of RLB estimators, *i.e.*

$$\mathcal{R}_{L,\mathrm{P}}\left(\frac{1}{B}\sum_{b=1}^{B} \hat{f}_{n_b,\lambda_{n_b}}\right) \to \mathcal{R}_{L,\mathrm{P}} \tag{10}$$

in probability for $n \to \infty$. A loss function L is called *Lipschitz continuous*, if there exists a constant $c \in (0,\infty)$ such that

$$|L(y,t) - L(y,t')| \leq c \cdot |t - t'| \tag{11}$$

for all $y \in \mathcal{Y}$, $t, t' \in \mathbb{R}$. In this case we denote the smallest possible constant c in (11) by $|L|_1$. Kernel based regression estimators based on convex and Lipschitz continuous loss functions are under weak conditions L-risk consistent and have nice robustness properties, see Christmann and Steinwart (2005).

Theorem 8 ($L-$risk consistency) *Consider a kernel based estimator $\hat{f}_{n,\lambda_n}$ based on (1) with a convex loss function which is Lipschitz continuous with Lipschitz constant $|L|_1$. Assume that $\hat{f}_{n,\lambda_n}$ is L-risk consistent and that there exists a measurable function f^* such that $\mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}_{L,\mathrm{P}}$ in (8) and*

$$\int \left|\hat{f}_{n,\lambda_n}(x) - f^*(x)\right| \, d\mathrm{P}(x,y) \to 0\,, \quad n \to \infty. \tag{12}$$

Further, assume that the mean *is used for the RLB estimator in the aggregation step and that $B \geq 1$ is fixed with $n/n_b \to B$, if $n \to \infty$. Then the RLB estimator $\hat{f}_{RLB,n,\lambda_n,B} := \frac{1}{B}\sum_{b=1}^{B} \hat{f}_{n_b,\lambda_{n_b}}$ is L-risk consistent.*

3.3 Robustness properties of RLB

Now we derive results which show that certain robustness properties are inherited from the original estimator T_n to the RLB estimator. Here, two different robustness approaches are considered. Donoho and Huber (1983) proposed the finite sample breakdown point to measure the worst case behavior of a statistical estimator. The influence function was proposed by F.R. Hampel, see Hampel *et al.* (1986), and measures the impact on the estimation due to an infinitesimal small contamination of the distribution P in direction of a Dirac-distribution.

Definition 9 (Finite-sample breakdown point) *Let $\mathcal{S}_n = \{(x_i, y_i), i = 1, \ldots, n\}$ be a data set with values in $\mathcal{X} \times \mathcal{Y}$. The finite-sample breakdown point of an estimator $T_n(\mathcal{S}_n)$ is defined by*

$$\varepsilon_n^*(T_n, \mathcal{S}_n) = \min\left\{\frac{m}{n}\,; Bias(m; T_n, \mathcal{S}_n) \;\; is \;\; finite\right\}\,, \tag{13}$$

where

$$Bias(m; T_n, \mathcal{S}_n) = \sup_{\mathcal{S}'_n} \parallel T_n(\mathcal{S}'_n) - T_n(\mathcal{S}_n) \parallel \tag{14}$$

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 10 (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 the finite sample breakdown point of the RLB estimator is given by*

$$\varepsilon^*_{RLB,n,B} = \varepsilon^*_{n_b}(T_b; \mathcal{S}_b) \cdot \left(\varepsilon^*_B(\hat{\mu}) + \frac{1}{B} \right) + \frac{B}{n} \cdot \varepsilon^*_B(\hat{\mu}) \,. \tag{15}$$

Remark 11 *(a) If B and $n_b = B/n$ are both large, we obtain from (15) the lower bound*

$$\varepsilon^*_{RLB,n,B} \geq \varepsilon^*_{n_b}(T_b; \mathcal{S}_b) \cdot \varepsilon^*_B(\hat{\mu}) \,. \tag{16}$$

(b) If the mean or any other estimator with $\varepsilon^*_B(\hat{\mu}) = 0$ is used in the aggregation step, RLB has a finite sample breakdown point of $\varepsilon^*_{n_b}(T_b; \mathcal{S}_b)/B \to 0$, if $B \to \infty$.

Example 12 (Univariate location model) Consider the univariate location problem, where $x_i \equiv 1$ and $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, $n = 55$. The finite sample breakdown point of the median is $[\![n/2]\!]/n = 0.49$. The mean has a finite sample breakdown point of 0. Now let us investigate the robustness of the RLB approach with $B = 5$ and $n_b = 11$, $b = 1, \ldots, B$. (a) If the median is used as the location estimator in each bite and if the median is used in the aggregation step, the finite sample breakdown point of the RLB estimator is $\varepsilon^*_{RLB,n,B} = 0.309$. This value is reasonably high, but lower than the finite sample breakdown point of the median for the whole data set, which is 0.49. Note that in a *fortunate* situation the impact of up to $(2 \cdot 11 + 5 \cdot 3)/55 = 0.672$ extreme large data points (say 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 $n_b\, \varepsilon^*_{n_b}(\hat{f}) = 5$ data points in the remaining $B(1 - \varepsilon^*_B(\hat{\mu})) = 3$ bites. This is no contradiction to (15) 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, the finite sample breakdown point of the RLB estimator is $\varepsilon^*_{RLB,n,B} = (1/B)\varepsilon^*_{n_b}(\hat{f}) = 0.09$. (c) If the mean is used as the location estimator in each bite and also in the aggregation step we obtain of course $\varepsilon^*_{RLB,n,B} = 0$. $\square$

Now we investigate the influence function of the RLB estimator $T_{RLB,n,B}$ for the case that the *mean* is used in the aggregation step, i.e. $T_{RLB,n,B} = \frac{1}{B} \sum_{b=1}^{B} T_{n_b}(S_b)$. To this end we assume the existence of a map T which assigns to every distribution P on a given set Z an element $T(\mathrm{P})$ of a given Banach space E such that our RLB estimator for a data set $S = S_1 + \ldots + S_B$ has the representation

$$T_{RLB,n,B} = T_{RLB,B}(\mathrm{P}_n) = \frac{1}{B} \sum_{b=1}^{B} T_{n_b}(\mathrm{P}_{n_b}). \tag{17}$$

Here P_n and P_{n_b} denote the empirical distributions of the whole sample S and of the bite S_b, $b = 1, \ldots, B$, respectively. For parametric models we have $T(\mathrm{P}) = \theta \in E = \mathbb{R}^d$. For kernel based methods defined by (1) $E = \mathcal{H}$ and $T(\mathrm{P}) = f_{\mathrm{P},\lambda}$.

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)*

$$IF(z; T, \mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{T\big((1-\varepsilon)\mathrm{P} + \varepsilon \Delta_z\big) - T(\mathrm{P})}{\varepsilon} \quad , \tag{18}$$

where Δ_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 $T(\mathrm{P})$. Therefore, in the robustness approach based on influence functions it is desirable that a statistical method which can be written as $T(\mathrm{P})$ has a *bounded* influence function.

Theorem 14 (Influence function of RLB) *Assume that the original estimator $T_n(S)$ has the representation $T(\mathrm{P}_n)$, where P_n is the empirical distribution of the sample S, and that the influence function of the map $T(\mathrm{P})$ exists for the probability distribution* P*. Then the RLB estimator based on the mean in the aggregation step with a fixed number B of bites exists and equals the influence function of $T(\mathrm{P})$.*

Hence, if $T(\mathrm{P})$ 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 have a bounded influence function, see e.g. Hampel *et al.* (1986) for M-estimators and GM-estimators in parametric models, and Davies (1990) for S-estimators in the linear regression model. Recently, Christmann and Steinwart (2004, 2005) showed that the influence function of various kernel based methods using Vapnik's convex risk minimization principle exists for the case of binary classification and regression. This is true e.g. for kernel logistic regression. Further the influence function of such methods can be bounded by choosing a loss function L with bounded first derivative and a bounded and universal kernel k, e.g. a Gaussian radial basis function kernel with $\gamma \in (0, \infty)$ is given by

$$k(x, x') = \exp(-\gamma ||x - x'||^2) \,, \quad x, x' \in \mathcal{X} \,.$$

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 general 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, hard disk) 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. Further, B should depend on the dimensionality d of the explanatory vectors $x_i \in \mathcal{X}$. *E.g.* a rule of thump 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 $n \to \infty$, to obtain consistency of RLB. The results from some numerical experiments not given here can be summarized as follows.

- If B is too large, the computational overhead increases and the danger of bad fits increases, because n_b is too small to provide reasonable estimators.
- A major decrease in computation time and memory saving is often already present, if B is chosen in a way such that the numerical algorithms to fit each bite fits nicely into the computer (CPU, RAM, hard disk). Nowadays robust estimators can often be computed for sample sizes up to $n_b = 10^4$ or $n_b = 10^5$. In this case $B = [\![n/n_b]\!]$ can be a reasonable choice.
- If distribution-free confidence intervals at the $(1 - \alpha)$ level for the median of the predictions, i.e. $T_{RLB,n,B}(x) = \text{median}_{1 \le b \le B} T_{n_b}(x)$, $x \in \mathcal{X}$, 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.

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). All 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 $\mathrm{P} = 0.8\mathrm{P}_1 + 0.2\Delta_{(x,y)}$, 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 θ is zero because the intercept term was set to zero. Further, this model contains outliers in $y-$direction due to the use of a Student distribution.

Table 4 shows the computation times in seconds, the bias of an MM-estimator and of the RLB estimator for $B = 17$ and the width of the componentwise confidence intervals 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). 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^{th} and the 12^{th} order statistics. Because the bias terms 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 4: 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 are very small due to the large sample size, the values in Table 4 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 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 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 the algorithm used by `rlm` using the RAM space of the computer, RLB only saves a little bit of computation time. However, RLB can be processed for much larger data sets for which the algorithm used by `rlm` 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 $\mathrm{P}(Y = 1|X = x)$, $x \in \mathcal{X}$, which is not true for the support vector machine (SVM), 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 of the algorithm proposed by Keerthi *et al.* (2002) to solve the dual problem. We choose KLR for two reasons. Firstly, the computation of KLR needs much more time than the SVM, because the latter solves a quadratic instead of a convex program in dual space. Secondly, the

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 5: Computation times for kernel logistic regression using myKLR.

number of support vectors of KLR is often approximately equal to n, which slows down the computation of predictions.

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 $x_i = (x_{i,1}, \ldots, x_{i,8})$ are simulated independently from a uniform distribution on $(0, 1)$. The responses y_i are simulated independently from a logistic regression model according to $\mathrm{P}(Y_i = +1|X_i = x_i) = 1/(1 + \exp[-f(x_i)])$ and $\mathrm{P}(Y_i = -1|X_i = x_i) = 1 - \mathrm{P}(Y_i = +1|X_i = x_i)$. We set

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

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 5. 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 16 minutes, if there is 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 $B = 10$ needs approximately 11 minutes which is a reduction by a factor of 55.

Concluding, RLB can be quite useful for kernel logistic regression for large data sets. Christmann (2004) describes a strategy to construct insurance tariffs for a data set from 15 German motor vehicle insurance companies. The whole data contains data from around 4.6 million customers. Although a strategy was used to reduce the computational effort by exploiting certain characteristic features of such data sets, RLB can help to reduce to computation time in a substantial manner.

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 *e.g.* in data mining problems or in the construction of insurance tariffs. RLB is especially designed for situations under which the original robust method cannot be used due to

excessive computation time or memory space problems. In these situations RLB offers robust estimates and additionally robust confidence intervals. Although RLB estimators will in general not fulfill certain optimality criteria, the method has the following advantages.

- *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 because the estimators computed from the B bites are independent and identically distributed. Such confidence intervals for the predictions are especially interesting for kernel based methods (*e.g.* support vector machine 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 RLB approach has connections 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 is 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 kernel based 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, 2002) 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-samples, 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 and have a complicated non-normal limiting distribution, see e.g. Rousseeuw (1984), Davies (1990), and Kim and Pollard (1990). If distribution-free confidence intervals for the median of the predictions are needed for such estimators RLB can be useful for data sets of only moderate sample size, too.

Acknowledgments

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") and of the Forschungsband DoMuS (University of Dortmund) is gratefully acknowledged.

Appendix

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

Proof of Proposition 1. Obvious. □

Proof of Proposition 2. Obvious. □

Proof of Proposition 3. (i) follows from the linearity of the expectation operator. (ii) and (iii) follow from Slutzky's theorem. □

Proof of Proposition 4. By construction of RLB the bites are disjoint and the estimators from the bites are independent. Assume that the original estimator $T_n(S)$ is consistent in probability. Then we have for all $\varepsilon > 0$ that

$$\begin{aligned} & \mathrm{P}\left(||\mathrm{median}_{b=1,\ldots,B}\, T_{n_b}(S_b) - T(\mathrm{P})|| < \varepsilon\right) \\ \geq\ & \mathrm{P}\left(||T_{n_b}(S_b) - T(\mathrm{P})|| < \varepsilon \text{ for all } b = 1, \ldots, B\right) \\ =\ & \prod_{b=1}^{B} \mathrm{P}\left(||T_{n_b}(S_b) - T(\mathrm{P})|| < \varepsilon\right) \to 1, \quad n \to \infty\,, \end{aligned}$$

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

$$\mathrm{P}\left(\lim_{n\to\infty} \mathrm{median}_{b=1,\ldots,B}\, T_{n_b}(S_b) = T(\mathrm{P})\right) \geq \prod_{b=1}^{B} \mathrm{P}\left(\lim_{n_b\to\infty} T_{n_b}(S_b) = T(\mathrm{P})\right) = 1,$$

because B is fixed and $\lim_{n\to\infty}\,(n/n_b) = B$. □

Proof of Proposition 5. By assumption each bite $\mathcal{S}_b$ is fitted with a kernel based estimator having the representation

$$\hat{f}_b(x) = \sum_{i=1}^{n_b} \alpha_{i,b}\, k(x, x_i), \quad i \in \mathcal{S}_b\,,\ b = 1, \ldots, B,\ x_i \in \mathcal{X}. \tag{19}$$

Because the bites $\mathcal{S}_b$, $b = 1, \ldots, B$, are disjoint, the RLB estimator using the mean in the aggregation step is given by

$$\begin{aligned} \hat{f}_{RLB,B}(x) &= \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n_b} \alpha_{i,b}\, k(x, x_i) && (20) \\ &= \sum_{i=1}^{n} \frac{\alpha_{i,b}}{B}\, k(x, x_i)\,, \quad x \in \mathcal{X}. && (21) \end{aligned}$$

The formula (5) follows immediately. □

Proof of Proposition 6. (i) This follows immediately from (5).
(ii) Steinwart (2003, Th.9) proved that the kernel based estimator evaluated for the whole data set $\mathcal{S}$ has the property

$$\Pr^{*n}\left(\mathcal{S}\in(\mathcal{X}\times\mathcal{Y})^n; \#\mathrm{SV}(\hat{f}_n)\geq(S_{L,\mathrm{P}}-\varepsilon)\right)\to 1\,, \tag{22}$$

if the conditions of the proposition are satisfied. Denote the outer probability measure of the product measure P^{n_b} by $\Pr^{*n_b}$. The bites $\mathcal{S}_b$, $b=1,\ldots,B$, are independent and identically distributed by construction of RLB. Using (22) we obtain

$$\Pr^{*n}\left(\mathcal{S}=(\mathcal{S}_1,\ldots,\mathcal{S}_B)\in(\mathcal{X}\times\mathcal{Y})^n; \#\mathrm{SV}(\hat{f}_{RLB,n,B})\geq\sum_{b=1}^{B}(S_{L,\mathrm{P}}-\varepsilon)n_b\right) \tag{23}$$

$$\geq \Pr^{*n}\left(\text{ for all }\mathcal{S}_b\in(\mathcal{X}\times\mathcal{Y})^{n_b}, b=1,\ldots,B\,;\,\#\mathrm{SV}(\hat{f}_{n_b})\geq(S_{L,\mathrm{P}}-\varepsilon)n_b\right) \tag{24}$$

$$=\prod_{b=1}^{B}\Pr^{*n_b}\left(\mathcal{S}_b\in(\mathcal{X}\times\mathcal{Y})^{n_b}; \#\mathrm{SV}(\hat{f}_{n_b})\geq(S_{L,\mathrm{P}}-\varepsilon)n_b\right)\to 1\,,\quad n\to\infty, \tag{25}$$

because B is fixed and $n_b\to\infty$. □

Proof of Theorem 8. Under the assumptions of the theorem we have

$$\begin{aligned}
0 &\leq \int L(y,\hat{f}_{RLB,n,\lambda_n,B}(x))\,d\mathrm{P}(x,y)-\mathcal{R}_{L,\mathrm{P}}\\
&= \int\left[L\left(y,\frac{1}{B}\sum_{b=1}^{B}\hat{f}_{n_b,\lambda_{n_b}}(x)\right)-L\left(y,f^*(x)\right)\right]d\mathrm{P}(x,y)\\
&\leq \int\left|L\left(y,\frac{1}{B}\sum_{b=1}^{B}\hat{f}_{n_b,\lambda_{n_b}}(x)\right)-L\left(y,f^*(x)\right)\right|d\mathrm{P}(x,y)\\
&\leq |L|_1\int\left|\frac{1}{B}\sum_{b=1}^{B}\hat{f}_{n_b,\lambda_{n_b}}(x)-f^*(x)\right|d\mathrm{P}(x,y) \qquad (26)\\
&\leq \frac{|L|_1}{B}\sum_{b=1}^{B}\int\left|\hat{f}_{n_b,\lambda_{n_b}}(x)-f^*(x)\right|d\mathrm{P}(x,y)\ \to\ 0, \qquad (27)
\end{aligned}$$

because B is fixed and $n_b\to\infty$, if $n\to\infty$. Here we used the Lipschitz continuity of the loss function in (26) and the consistency assumption (12) in (27). □

Proof of Theorem 10. The minimum number of points needed to modify $T_b(S_b)$ in bite $\mathcal{S}_b$ such that there is breakdown is given by $n_b\cdot\varepsilon^*_{n_b}(T_b;\mathcal{S}_b)+1$, $\quad b=1,\ldots,B$. The RLB estimator breaks down if at least $B\varepsilon^*_B(\hat{\mu})+1$ of the estimators $\hat{f}(S_1)$, …, $\hat{f}(S_B)$ break down. Therefore

$$\varepsilon^*_{RLB,n,B}=\frac{(n_b\,\varepsilon^*_{n_b}(T_b;\mathcal{S}_b)+1)\cdot(B\,\varepsilon^*_B(\hat{\mu})+1)-1}{n}\,, \tag{28}$$

which gives the assertion. □

Proof of Theorem 14. Let $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ and P be a probability distribution on $\mathcal{X} \times \mathcal{Y}$. By assumption the RLB estimator has the property (17), i.e. $T_{RLB,n,B} = \frac{1}{B}\sum_{b=1}^{B} T_{n_b}(\mathrm{P}_{n_b})$, where P_{n_b} denotes the empirical distribution of bite S_b, $b = 1, \ldots, B$. Further, the influence function $IF(z; T, \mathrm{P})$ exists by assumption of the theorem. It follows

$$\begin{aligned} IF(z; T_{RLB,B}, \mathrm{P}) &= \lim_{\varepsilon \downarrow 0} \frac{T_{RLB,B}\big((1-\varepsilon)\mathrm{P} + \varepsilon\Delta_z\big) - T_{RLB,B}(\mathrm{P})}{\varepsilon} \\ &= \lim_{\varepsilon \downarrow 0} \frac{\frac{1}{B}\sum_{b=1}^{B} T\big((1-\varepsilon)\mathrm{P} + \varepsilon\Delta_z\big) - \frac{1}{B}\sum_{b=1}^{B} T(\mathrm{P})}{\varepsilon} \\ &= \frac{1}{B}\sum_{b=1}^{B} \lim_{\varepsilon \downarrow 0} \frac{T\big((1-\varepsilon)\mathrm{P} + \varepsilon\Delta_z\big) - T(\mathrm{P})}{\varepsilon}\,, \end{aligned}$$

which gives the assertion. □

References

Bartlett, P. and Tewari, A. (2004). Sparseness vs estimating conditional probabilities: Some asymptotic results. Technical report, University of California, Berkeley.

Breiman, L. (1999). Pasting bites together for prediction in large data sets. *Machine Learning*, **36**, 85–103.

Chawla, N., Hall, L., Bowyer, K., and Kegelmeyer, W. (2004). Learning ensembles for bites: a scalable and accurate approach. *Journal of Machine Learning Reseach*, **5**, 421–451.

Christmann, A. (2004). An approach to model complex high-dimensional insurance data. *Allg. Statist. Archiv*, **88**, 375–397.

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). Robustness properties of kernel based regression. University of Dortmund, Department of Statistics, Technical report 01/05, SFB 475, submitted.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics, 3rd ed.* Wiley & Sons, New York.

Davies, P. (1990). The asymptotics of S-estimators in the linear regression model. *Ann. Statist.*, **18**, 1651–1675.

Donoho, D. and Huber, P. (1983). The notion of breakdown point. In P. Bickel, K. Doksum, and J. Hodges, editors, *A Festschrift for Erich L. Lehmann*, pages 157–184. Jr., Belmont, California, Wadsworth.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust statistics: The Approach Based on Influence Functions*. Wiley, New York.

Keerthi, S., Duan, K., Shevade, S., and Poo, A. (2002). A fast dual algorithm for kernel logistic regression. National University of Singapore. Preprint.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.*, **18**, 191–219.

Rousseeuw, P. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinat estimator. *Technometrics*, **41**, 212–223.

Rousseeuw, P. and Driessen, K. V. (2002). Computing lts regression for large data sets. *Estadistica*, **54**, 163–190.

Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, **79**, 871–880.

Rüping, S. (2003). myKLR - kernel logistic regression. Technical report, University of Dortmund, Department of Computer Science.

Schölkopf, B. and Smola, A. (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. *J. Mach. Learn. Res.*, **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. *J. Mach. Learn. Res.*, **4**, 1071–1105.

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel machines. *IEEE Trans. Inform. Theory*, **to appear**. `http://www.c3.lanl.gov/~ingo/publications/info-02.ps`.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.

Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer, New York.

Wahba, G. (1999). Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 69–88, Cambridge, MA. MIT Press.

Yohai, V. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.*, **15**, 642–656.

Zhang, T. (2004). Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, **32**, 56–134.

Andreas Christmann
University of Dortmund
Department of Statistics
44221 Dortmund
Germany
E-mail:
christmann@statistik.uni-dortmund.de