Risk Scores, Empirical Z-estimators and its application to Censored Regression

K. Pelckmans, J. De Brabanter †, J.A.K. Suykens, B. De Moor

K.U.Leuven, ESAT, SCD/sista, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
† Hogeschool KaHo Sint-Lieven, (Associatie K.U.Leuven), B-9000 Gent, Belgium
Kristiaan.Pelckmans@esat.kuleuven.be.

Abstract. This work explores the use of risk scores and Zero-estimators (Z-estimators) into learning theory. Those notions are introduced in analogy with the likelihood scores in parametric probabilistic methods, but instead studies its relation with empirical risk minimization and its finite sample properties. The case of the L_1-norm location estimator is discussed in some detail, and it is shown how this mechanism can be used for the design of a learning machine for regression in the context of censored observations where the risk is not completely accessible. This work aims at the further integration of optimization theory in the analysis of learning algorithms.

1 MOTIVATION

A profound disadvantage of parametric statistical techniques is that valid inferences assume a given parametric model class containing the *true parameter* [6]. Robust statistics [5] relaxed this assumption somewhat by allowing the true parameter to lie in a close neighborhood of the parametric model class. Distribution-free learning theory [13, 4] takes a different viewpoint by adopting a more flexible non-parametric class of distributions, not necessarily indexed by a parameter. The objective here is the design and analysis of algorithms which result in estimates which are guaranteed to perform well on future samples which are generated based on the same unknown stochastical model as before. Although the *predictive model* class is to be of finite capacity [13] before effectively implementing risk minimization, minimal risk inferences are valid for almost any underlying *stochastic model*. We further investigate this line of thought for the task of estimating unknowns in a predictive model (θ are called the parameters of the predictive model, or in short *parameters*). We discuss finite sample risk guarantees of such estimates using a pre-specified loss function ℓ and for an arbitrary distribution in a non-parametric class $\mathcal{F}$. Herefor, the notion of risk scores is introduced and investigated, resulting in a class of learning algorithms which are named empirical Z-estimators, This idea is developed in analogy to the classical Z-estimators for maximum likelihood and M-estimators, see e.g. [12], Chapter 5. This work is restricted to the case where the parameters belong to a finite dimensional vector space.

This paper presents four main results. At first, we introduce the formal notion of risk scores and risk based Z-estimators. Secondly, we show that empirical Z-estimators can be used to give results which will converge to the minimal risk estimate under

certain conditions. Hereto, we require that the loss function under consideration is differentiable and Lipschitz bounded enabling a technique of quantization. This technique is advantageous over classical approaches based on covering numbers (see e.g. [13, 4, 2]) as no assumption is to be made concerning finiteness nor the existence of second moments (the support of the distribution need not be bounded). This property paves the way towards the design of robust methods. Thirdly, the case of the one norm (denoted as ℓ_1) for location estimation is considered in some detail. Finally, the usefulness of empirical Z-estimators is indicated by constructing a convex learning machine for regression in the context of censored observations.

This paper is organized as follows. Section 2 discusses the notion of risk scores in the case of differentiable, Lipschitz bounded loss functions and also considers in some detail the case of the L_1-norm for estimating the minimal risk location. Section 3 explores its use in the construction and analysis of a convex approach towards inference in the presence of randomly censored observations. Capital letters are used for random variables, scalars for deterministic values or vectors. An exception is made for the parameter vector $\hat{\theta}_n$ minimizing the risk, which is a function of random variables and thus a random variable itself. The symbol '$\xrightarrow{P}$' denotes convergence in probability when the limit is taken for $n \to \infty$.

2 RISK SCORES

2.1 Empirical Z-estimators

Let Z be a (multivariate) random variable generated from a fixed but unknown stochastic model as represented by the distribution function $F_Z \in \mathcal{F}$. Here, $\mathcal{F}$ is a non-parametric class of distributions with existing first moment. Assume further that any joint distribution in the class $\mathcal{F}$ is linearly independent, e.g. two random variables Z_i and Z_j are collinear with zero probability for any $i < j$. Let the data-samples $\{Z_i\}_{i=1}^n$ be sampled iid from the same distribution, abbreviated as $Z_{1:n}$ for notational convenience.

Definition 1 (Risk Scores) *Let $\ell : \mathbb{R}^D \to \mathbb{R}^+$ in $\mathcal{C}^1(\mathbb{R}^D, \mathbb{R})$ be a loss function measuring the discrepancy between parameters $\theta \in \Theta = \mathbb{R}^D$ and the fixed distribution F_Z (in some sense). (We do not consider the generalization to the constrained case where $\Theta \subset \mathbb{R}^D$.) The risk score for a given parameter vector θ is then defined as $r : \mathbb{R}^D \to \mathbb{R}$ in $\mathcal{C}^0(\mathbb{R}^D, \mathbb{R})$ as follows*

$$r_d(\theta^*; Z) = \frac{\partial \ell(\theta; F_Z)}{\partial \theta_d}\Big|_{\theta=\theta^*}, \quad \forall d = 1, \ldots, D, \tag{1}$$

where with a slight abuse of notation, the dependence of the lhs on the chosen loss function ℓ is omitted for notational convenience. The expected risk score can then be defined as

$$r_d(\theta^*; F_Z) = E\left[\frac{\partial \ell(\theta; F_Z)}{\partial \theta_d}\Big|_{\theta=\theta^*}\right], \quad \forall d = 1, \ldots, D, \tag{2}$$

The expected risk score $r : \Theta \to \mathbb{R}^D$ is defined as the vector valued function $r(\theta; F_Z) = (r_1(\theta; Z), \ldots, r_d(\theta; Z))^T \in \mathbb{R}^D$. Let the empirical counterpart $r^d_{1:n} : \mathbb{R}^D \to \mathbb{R}$ in $\hat{C}^0(\mathbb{R}^D, R)$ be defined as follows

$$r_{d,1:n}(\theta^*; Z_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(\theta; Z_i)}{\partial \theta_d}\Big|_{\theta=\theta^*}, \quad \forall d = 1, \ldots, D, \tag{3}$$

for the realizations $\{Z_i\}_{i=1}^n$ sampled iid from F_Z. The empirical vector valued function $r_{1:n} : \Theta \to \mathbb{R}^D$ is defined as $r_{1:n}(\theta; F_Z) = (r_{1,1:n}(\theta; F_Z), \ldots, r_{D,1:n}(\theta; F_Z))^T \in \mathbb{R}^D$.

Note the subtle notational difference between the expectation $r(\theta; F_Z)$, the random vector $r(\theta; Z)$, its empirical counterparts $r_{1:n}(\theta; Z_{1:n})$, and the point evaluation $r(\theta; Z_i)$ of sample Z_i. As a prototype, consider the case $Z = (X, Y) \subset \mathbb{R}^D \times \mathbb{R}$ and $\ell(\theta, F_{XY}) = \frac{1}{2}(\theta^T X - Y)^2$ for $\theta \in \mathbb{R}^D$. The corresponding expected risk score becomes $E[(\theta^T X - Y)]$. The following proposition summarizes some properties of this quantity.

Proposition 1 (Properties of Risk Scores) *Under sufficient regularity conditions on the distribution F_Z, the following equalities follow straightforwardly from the linearity of the derivative and the expectation:*

1. *The expectation and the derivative may be switched under the usual regularity conditions on the distribution F_Z and the loss function (the interchange of integration and differentiation is based on Lebesgue's Dominated Convergence theorem, see e.g. [10]) :*

$$r_d(\theta; F_Z) = E\left[\frac{\partial \ell(\theta; Z)}{\partial \theta_d}\right] = \frac{\partial}{\partial \theta_d}\mathcal{R}(\theta; F_Z), \quad \forall d = 1, \ldots, D, \tag{4}$$

   where $\mathcal{R}(\theta; F_Z) = E[\ell(\theta; F_Z)]$ denotes the actual risk.
2. *The empirical risk score provides an unbiased estimate of the actual risk scores*

$$E[r_{d,1:n}(\theta; Z_{1:n})] = r_d(\theta; F_Z), \quad \forall d = 1, \ldots, D. \tag{5}$$

3. *If $\ell \in \mathcal{C}^1(\mathbb{R}^D, R)$ is Lipschitz smooth with constant $L > 0$, the corresponding risk score is bounded in the interval $]-L, L[$.*
4. *The risk score of a sample Z_i, defined as*

$$r_d(\theta_0; Z_i) = \frac{\partial \ell(\theta; Z_i)}{\partial \theta_d}\Big|_{\theta=\theta_0}, \quad \forall d = 1, \ldots, D, \tag{6}$$

   for all $i = 1, \ldots, n$. The terms $r_d(\theta; Z_i)$ for fixed θ are iid random variables bounded in the interval $]-L, L[$ when ℓ is Lipschitz smooth. Consider a fixed parameter $\theta \in \Theta$. In this case, the following probabilistic bound holds

$$P\left(|r_d(\theta; F_Z) - r_{d,1:n}(\theta; Z_{1:n})| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{2L^2}\right), \quad \forall d = 1, \ldots, D. \tag{7}$$

Proof: Statements 1 to 3 follow from the linearity of the expectation and the derivative operator. Statement 4 follows readily from Hoeffding's inequality (see e.g. [4]). ■

The primal objective of these notions is to be able to formalize the notion of an actual Z-estimator and its empirical counterpart.

Definition 2 (Z-estimator for Empirical Risk Minimization) *An actual Z-estimator for loss ℓ and domain Θ outputs a parameter vector $\hat{\theta} \in \Theta$ for a given distribution F_Z such that the following set of equalities hold*

$$r_d(\hat{\theta}; F_Z) = 0, \ \ \forall d = 1, \dots, D, \tag{8}$$

for r the risk score r associated with the fixed loss ℓ. An empirical Z-estimator outputs a parameter vector $\hat{\theta}_n \in \Theta$ for given sample $Z_{1:n}$ such that

$$r_{d,1:n}(\hat{\theta}_n; Z_{1:n}) = 0, \ \ \forall d = 1, \dots, D. \tag{9}$$

Those estimators are inspired by optimization routines which try to set the gradients of the objective to zero. The following section gives such a Z-estimator for the task of regression in the case of censored observations: here the empirical risk cannot be computed completely due to censoring. An important advantage of this kind of estimators is that in the case where the expectation of the risk is not finite (e.g. outliers of Y ocuur or when Y is Cauchy distributed), the risk scores are still bounded and can as such be used for robust inference. Note that in classical theorems on learning of real-valued functions as e.g. in [2], one requires typically bounded response variables excluding those cases.

Let the risk score r_d be associated with the loss ℓ, and let the quantized risk score $q_d : \mathbb{R}^D \to [-L, L]$ for a given quantization level $Q \in \mathbb{N}_0$ be defined as follows

$$q_d(\theta; Z, Q) = \frac{L}{Q} \left\lceil \frac{Q}{L} r_d(\theta; Z) \right\rceil, \tag{10}$$

which has $2Q+1$ different possible output levels $G_{Q,k} = \left(\frac{kL}{Q}\right) - L$ for $k = 0, \dots, 2Q$. Figure 1 exemplifies such quantification for a risk score function in the case of regression $\ell(\theta; (X, Y)) = \ell(\theta^T X - Y)$. Let the empirical counterpart $q_{d,1:n}$ be defined correspondingly. Define the quantification error $0 \leq \epsilon_Q$ as $\sup_e |r_d(e) - q_d(e)|$ which satisfies by construction $\epsilon_Q \leq \frac{L}{Q}$.

Lemma 1 (Uniform Convergence of Risk Scores) *Assume the loss $\ell \in \mathcal{C}^1(\mathbb{R})$ is a Lipschitz bounded function with constant $0 < L < \infty$. The random vector Z is assumed to occur in the loss through a linear combination, where $\theta \in \Theta$ parameterizes this combination in a linear independent way. Let $0 < \alpha < 1$ be a constant confidence level, then with probability higher than $1 - \alpha$, the following inequality holds*

$$\sup_{\theta \in \Theta} |r(\theta; F_Z) - r_{1:n}(\theta; Z_{1:n})|_\infty \\ \leq \frac{4L}{Q} + \sqrt{\frac{32L^2}{n} \left(\ln \left(16D(2Q+1) \sum_{i=0}^{D-1} \binom{n-1}{i} \right) - \ln \alpha \right)}. \tag{11}$$

where we denote the rhs in short as $\nu(n, \alpha; D, Q, L)$. Specifically, the rhs goes to zero when $Q \to \infty$ and $n \to \infty$ at sufficient rates. This proofs uniform convergence of $r_{1:n} \xrightarrow{P} r$.

Proof: Then the large deviation can be bounded as follows

$$\begin{aligned}\sup_{\theta\in\Theta} |r_{d,1:n}(\theta; Z_{1:n}) - r_d(\theta; F_Z)| \leq & \sup_{\theta\in\Theta} |r_{d,1:n}(\theta; Z_{1:n}) - q_{d,1:n}(\theta; Z_{1:n})| \\ & + \sup_{\theta\in\Theta} |q_{d,1:n}(\theta; Z_{1:n}) - q_d(\theta; F_Z)| \\ & + \sup_{\theta\in\Theta} |q_d(\theta; F_Z) - r_d(\theta; F_Z)| \,. \end{aligned} \tag{12}$$

Both the first and last term of the rhs can be bounded by ϵ_Q by linearity of the expectation operator:

$$|E[r_d(\theta; F_Z)] - E[q_d(\theta; F_Z)]| \leq E[|r_d(\theta; F_Z) - q_d(\theta; F_Z)|] \leq E[\epsilon_Q] = \epsilon_Q. \tag{13}$$

The midterm of the rhs of equation (12) is bounded by reducing the terms to a classification problem. Herefor, let the indicator function $I(z < c)$ be 1 if $z < c$ and zero otherwise, and let the variables δ_d^Q be defined as

$$\delta_d^Q(\theta; Z_i, k) = I(q_d(\theta; Z_i, Q) \leq G_{Q,k}) \in \{0, 1\}, \tag{14}$$

for $k = 0, \ldots, 2Q$, indicating whether a risk score $r_i^d(\theta; Z_i)$ is smaller or equal than the k-th quantized level. Notice that the set $\{\delta_d^Q(\theta; Z_i, k)\}_{i=1}^n$ for given Q, d and k is iid (by construction). Application of the union bound (as thoroughly used in e.g. [13, 4]) yields for any $\epsilon > 0$:

$$P\left(\sup_{\theta\in\Theta} \left| E[\delta_d^Q(\theta; Z, k)] - \frac{1}{n}\sum_{i=1}^n \delta_d^Q(\theta; Z_i, k) \right| \geq \epsilon \right) \leq 16 \sum_{i=0}^{D-1} \binom{n-1}{i} \exp\left(-\frac{n\epsilon^2}{8}\right), \tag{15}$$

where the bound on the shattering number of the vector space Θ with dimension D is as described in [3, 4] under the assumption that the D variables are linearly independent. Rewriting the term q_d as $q_d(\theta; Z, Q) = \sum_{k=0}^{2Q} \left(\frac{L}{Q}\right) \delta_d^Q(\theta; Z, k) - L$ (weighted sum of the indicators) and application of the above bound gives the bound

$$\begin{aligned} & P\left(\sup_{\theta\in\Theta} \left| q_d(\theta; F_Z, Q) - \frac{1}{n}\sum_{i=1}^n q_d(\theta; Z_i, Q) \right| \geq \epsilon \right) \\ & \quad \leq P\left(\sup_{\theta\in\Theta} \frac{L}{Q} \sum_{k=0}^{Q} \left| \delta_d^Q(\theta; F_Z, k) - \frac{1}{n}\sum_{i=1}^n \delta_d^Q(\theta; Z_i, k) \right| \geq \epsilon \right) \\ & \quad \leq P\left(\sup_{\theta\in\Theta} \sup_k \left| \delta_d^Q(\theta; F_Z, k) - \frac{1}{n}\sum_{i=1}^n \delta_d^Q(\theta; Z_i, k) \right| \geq \frac{Q}{L}.\frac{\epsilon}{Q} \right) \\ & \quad \leq 16(2Q+1) \sum_{i=0}^{D-1} \binom{n-1}{i} \exp\left(-\frac{n\epsilon^2}{32L^2}\right), \end{aligned} \tag{16}$$

by linearity of the terms q_d and δ_d^Q; application of the inequality $P(\sum_{i=1}^n |Y_i| > \epsilon) \leq P(\sup_i |Y_i| > \epsilon/n)$; and application of the union bound over the $2Q+1$ different quantized intervals respectively. Application of the union bound over the dimensions $d = 0, \ldots, D$ and substitution of the rhs by $0 < \alpha$ gives (11). ■

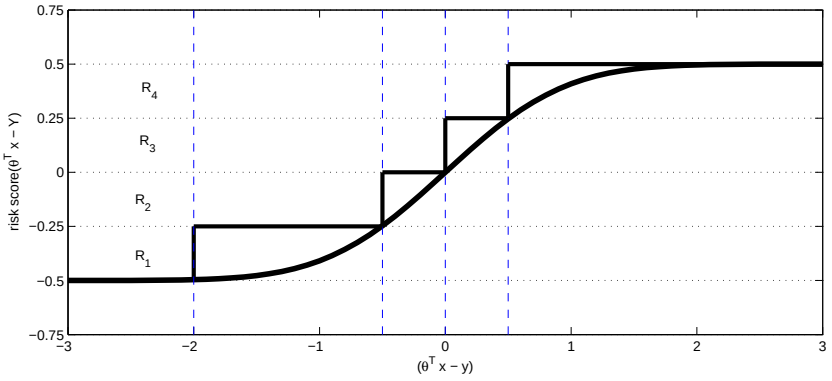

Fig. 1. *An example of a risk score (solid smooth line) as a function of the parameter θ in a regression setting where $\ell(\theta; F_Z) = \ell(\theta^T X - Y)$. The horizontal dotted lines indicate the $Q+1 = 4$ different quantized intervals R_1, R_2, R_3 and R_4. The staircase (solid line) indicates the quantized risk score as in (10). The vertical dashed lines indicate the corresponding quantization of the residuals $(\theta^T x - y)$.*

At this point, we can state our main result. Let a loss function ℓ with associated risk score r for distribution F_Z have a separable minimum in case for all $\epsilon > 0$, the following inequality holds

$$\inf_{|\theta - \theta_*|_\infty > \epsilon} |r(\theta; F_Z)|_\infty > 0 = |r(\theta_*; F_Z)|_\infty \,. \tag{17}$$

Theorem 1 (Convergence of Empirical Z-estimators) *Let the set $\{Z_i\}_{i=1}^n$ be sampled iid from a fixed but unknown underlying joint distribution $F_Z \in \mathcal{F}$ possessing regularity conditions as stated previously. Assume the continuous separable function $\ell : \mathbb{R}^D \to \mathbb{R} \in \mathcal{C}_1(\mathbb{R}^D, R)$ have a separable root as in (17). Then $r(\theta_*; F_Z) = 0 \Leftrightarrow \theta_* = \arg\min_{\theta \in \Theta} \mathcal{R}(\theta; F_Z)$. Moreover, any sequence of estimators $\{\hat{\theta}_n\}$ with bounded deviation from the optimum $r(\hat{\theta}_n; F_Z)$ converges to the theoretical risk minimizer $\hat{\theta}_n \xrightarrow{P} \theta_*$ with $|r(\theta_*; F_Z)|_\infty = 0$.*

Proof: The proof follows from Lemma 1 and taking the limit $Q \to \infty$ of inequality (11) for $n \to \infty$. A formal justification can be found in Theorem 5.7 of [12]. ■

Particularly, an algorithm which has its risk scores converging to zero - or $r_{1:n}(\theta; Z_{1:n}) \xrightarrow{P} 0$ - effectively implements risk minimization. This theorem motivates the construction

of algorithms which try to find zero risk scores based on a finite sample as exemplified in Section 3.

Corollary 1 (Testing a Point Prediction with a Finite Sample) *Consider a fixed estimate $\hat{\theta} \in \Theta$, and a Lipschitz bounded loss function $\ell \in \mathcal{C}^1(\mathbb{R}^D, \mathbb{R})$ with constant $L < \infty$. The probability of occurrence of the empirical risk $r_{1:n}(\hat{\theta})$ under the null hypothesis that F_Z is such that $|r(\hat{\theta}; F_Z)|_\infty = 0$ obeys the tail inequality*

$$P\left(\left|r_{1:n}(\hat{\theta}; Z_{1:n})\right|_\infty \geq \epsilon \;\middle|\; |r(\hat{\theta}; F_Z)|_\infty = 0\right) \leq D \exp\left(-\frac{\epsilon^2 n}{2L^2}\right). \tag{18}$$

Proof: This statement follows readily by application of Hoeffding's inequality as in Proposition 1.4, together with a union bound over the D dimensions. ■

This can be used for testing a single estimate $\hat{\theta}$ on an independent test-set $Z_{1:n}$. To see this, substitute the rhs of (18) by the confidence level $0 < \alpha < 1$ as with confidence level smaller than α, the following inequality holds

$$\left|r_{1:n}(\hat{\theta}; Z_{1:n})\right|_\infty \geq \sqrt{\frac{2L^2(\ln D - \ln \alpha)}{n}} \tag{19}$$

under the condition that $r(\hat{\theta}; F_Z) = 0$. Now, the null-distribution becomes F_Z is such that $r(\hat{\theta}; F_Z) = 0$, the test-statistic is $r_{1:n}(\hat{\theta})$, and the finite sample (!) test distribution has a tail bound as in the rhs of (18), with derived critical value which is an upper-bound to the p-value as in (19) (see e.g. [6]). Testing for a finite number (say $1 < m < \infty$) of hypothesis $\{\theta_1, \ldots, \theta_m\}$ can be done similarly by adding a Bonferroni correction term (or equivalently, application of the union bound) such that the rhs of (19) becomes $\sqrt{\frac{2L^2(\ln(Dm) - \ln \alpha)}{n}}$. The following result studies the extension towards infinite sets, which occurs e.g. if one wants to have a confidence set of the results after optimization over an infinite set, say $\Theta = \mathbb{R}^D$.

Definition 3 (Confidence Set for a Finite Sample) *Let $\theta_* \in \mathbb{R}^D$ be the Bayes estimate such that $r(\theta_*; F_Z) = 0$. Let $0 < \alpha < 1$ be a pre-specified confidence of covering. Given a finite sample $\{Z_i\}_{i=1}^n$ sampled iid from F_Z, a nontrivial set $\mathcal{S}_{\alpha,n} \subset \Theta$ can be constructed which includes the Bayes estimate with high probability, formally*

$$P\left(\theta_* \in \mathcal{S}_{\alpha,n}\right) \geq \alpha. \tag{20}$$

To construct this set, all vectors $\theta \in \Theta$ should be included whose empirical risk would not be too far of the minimum actual risk. In terms of scores, this can be formalized as follows by application of Theorem 1:

$$\mathcal{S}_{\alpha,n} = \left\{\theta \in \Theta \;\middle|\; |r_{1:n}(\theta; Z_{1:n})|_\infty \leq \nu(n, \alpha; D, Q, L)\right\}. \tag{21}$$

2.2 ℓ_1 LOSS FUNCTION FOR LOCATION ESTIMATION

We consider the prototypical case of estimating the minimum risk location parameter $\theta \in \mathbb{R}$ of univariate random variable Y with iid samples $\{Y_i\}_{i=1}^n$, and with the ℓ_1 loss in some detail. The L_1-norm loss is defined as $\ell_1(\theta; Z) = \ell_1(\theta; Y) = |\theta - Y|$. Since the L_1-norm $\ell_1 \in \mathcal{C}_0(\mathbb{R}^D, \mathbb{R})$, is not differentiable, we resort to a closely related loss function which belongs to $\mathcal{C}_\infty(\mathbb{R}^D, \mathbb{R})$. Let the loss ℓ_{a_s} be defined for any $e \in \mathbb{R}$ as

$$\ell_{a_s}(e) = \sqrt{e^2 + a_s}, \tag{22}$$

with $a_s > 0$ an appropriate constant (see Figure (2)).

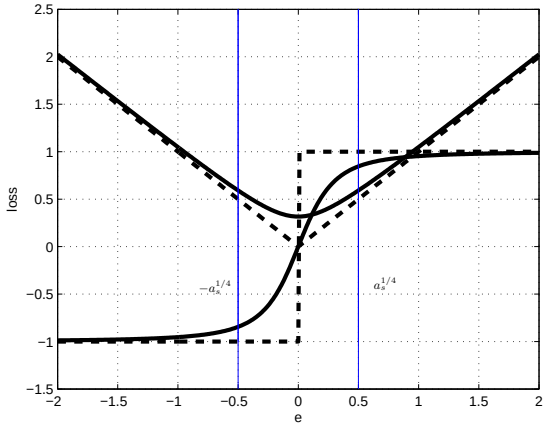

Fig. 2. *The ℓ_1 loss function and its discontinuous derivative (dashed lines), and the differentiable approximation ℓ_{a_s} and its respective derivative (smooth solid lines). The horizontal thin lines indicate the critical region ($]-a_s^{1/4}, a_s^{1/4}[$) where the approximation r_{a_s} is has a distance from q^{a_s} smaller than one.*

Lemma 2 (Uniform Convergence) *The approximation ℓ_{a_s} converges uniformly to ℓ_1 when $a_s \to 0$, formally*

$$\lim_{a_s \to 0} \sup_{e \sim F} |\ell_{a_s}(e) - \ell_1(e)| = 0. \tag{23}$$

Moreover, the optimal estimate using the approximate norm ℓ_{a_s}, defined as $\hat{\theta}_{a_s} = \arg\min_{\theta \in \Theta} \ell_{a_s}(\theta - Y)$ corresponding with ℓ_{a_s} for $a_s > 0$ will converge to $\hat{\theta}_1$ which is the parameter which has optimal risk with loss $\ell(\cdot) = |\cdot|$, or formally, $\hat{\theta}_1 = \arg\min_{\theta \in \Theta} \ell_1(\theta - Y)$. Then both will converge uniformly when a_s shrinks sufficiently fast, formally

$$\lim_{a_s \to 0} \hat{\theta}_{a_s} = \hat{\theta}_*, \tag{24}$$

where $\hat{\theta}_* = \arg\min_{\theta\in\Theta} \mathcal{R}_{1:n}(\theta; Z_{1:n})$.

This result is proven in [8, 14]. Let now r^{d,a_s} be defined as the dth risk score of the differentiable loss ℓ_{a_s} such that in the case of the univariate location estimate $Z = Y \in \mathbb{R}$ one can write

$$r^{a_s}(\theta; Y) = \frac{\ell_{a_s}(\theta - Y)}{\partial\theta} = \frac{\theta - Y}{\sqrt{(\theta - Y)^2 + a_s}}. \tag{25}$$

The following quantified version q^{d,a_s} is defined for fixed $0 < a_s$ as follows

$$q^{a_s}(\theta; F_Y) = \begin{cases} -1 & \text{iff} \quad (\theta - Y) < -a_s^{1/4} \\ 1 & \text{iff} \quad (\theta - Y) > a_s^{1/4} \\ 0 & \text{elsewhere.} \end{cases} \tag{26}$$

A first result bounds the deviation of this quantification q^{a_s} to the continuous scores r^{a_s} for any fixed θ.

Proposition 2 (Bounded Deviation of Quantization) *Let any set* $\{Y \;\big|\; |\theta - Y| \leq a_s^{1/4}\}$ *contain a probability mass of at most* $0 \leq p(a_s) < 1$ *for any parameter* $\theta \in \mathbb{R}$ *and underlying distribution* F_Y. *Then the expected deviation between* r^{a_s} *and* q^{a_s} *can be bounded in terms of this probability mass for any fixed* $\theta \in \mathbb{R}$ *as follows*

$$E\Big[\, |r^{a_s}(\theta; Y) - q^{a_s}(\theta; Y)| \,\Big] \leq (1 - p(a_s)) \left(1 - \frac{1}{\sqrt{1 + \sqrt{a_s}}}\right) + p(a_s). \tag{27}$$

Proof: This argument can be proven purely geometrically (see Figure 2). Let e be defined as $e = (\theta - Y)$. First, assume that $e \geq a_s$ (or equivalently $e \leq a_s$), then $1 - |r^{a_s}(\theta; Y)| = 1 - \left|\frac{e}{\sqrt{e^2 + a_s}}\right|$. This quantity can be bounded as $1 - \frac{1}{\sqrt{1+\sqrt{a_s}}}$ by substitution of e by $a_s^{1/4}$ and by using the fact that r^{a_s} is monotonically increasing. When $|e| < a_s^{1/4}$, the deviation is bound as $|r^{a_s}(\theta; Y)| < 1$ by construction. This together with the definition of $p(a_s)$ proves the inequality (27). ■

We now specialize this result for the parameter $\hat{\theta}_n$ which solves the convex problem

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} \ell_1(\theta - Y_i), \tag{28}$$

for given iid sample $Y_{1:n}$ of size n. This is done by relating the Lagrange multipliers to the quantization as proposed in (26). Without loss of generality, the following Proposition is stated in a deterministic setting where the samples are fixed such that $Y_i = y_i$ for all $i = 1, \ldots, n$. At first, the dual derivation of problem (28) is given. The ℓ_1 based location estimation problem can be formulated as a convex constrained programming problem by introducting slack variables $\{e_i\}_{i=1}^n$ as follows

$$\min_{\theta\in\mathbb{R}, e\in\mathbb{R}^n} \mathcal{J}(e) = \sum_{i=1}^{n} e_i \quad \text{s.t.} \quad -e_i \leq \theta - y_i \leq e_i, \;\; \forall i = 1, \ldots, n. \tag{29}$$

The Lagrangian becomes $\mathcal{L}(e,\theta;\alpha^+,\alpha^-) = \sum_{i=1}^n e_i - \sum_{i=1}^n \alpha_i^+(e_i+\theta-y_i) - \sum_{i=1}^n \alpha_i^-(e_i-\theta+y_i)$. The first order conditions for optimality of maximizing the Lagrangian w.r.t. the primal variables e_i and θ become

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial e_i} = 0 \rightarrow 1 = \alpha_i^+ + \alpha_i^- & \forall i = 1,\ldots,n \\ \dfrac{\partial \mathcal{L}}{\partial \theta} = 0 \rightarrow \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) = 0. \end{cases} \tag{30}$$

Those conditions are satisfied by strong duality property, see e.g. [9]. Substitution of $(\alpha_i^+ - \alpha_i^-)$ by a new variable α_i for all $i = 1,\ldots,n$ gives

$$\begin{cases} -1 \leq \alpha_i \leq 1 & \forall i = 1,\ldots,n & (a) \\ \sum_{i=1}^n \alpha_i = 0 & & (b) \\ \alpha_i = -1 \Leftrightarrow \theta - y_i > 0 & \forall i = 1,\ldots,n & (d) \\ \alpha_i = 1 \Leftrightarrow \theta - y_i < 0 & \forall i = 1,\ldots,n & (e) \\ -1 < \alpha < 1 \Leftrightarrow \theta - y_i = 0 & \forall i = 1,\ldots,n & (f) \end{cases} \tag{31}$$

where the last three statements follow from the complementary slackness conditions.

Proposition 3 (Lagrange Multipliers and Risk Scores) *Let $\hat{\theta}_n \in \Theta$ be the empirical optimum obtained by minimizing problem (28), and let the vector α be defined as $\alpha = (\alpha_i,\ldots,\alpha_s)^T \in \mathbb{R}^n$. For any $a_s > 0$, the following inequality holds*

$$\left|\alpha_i - r^{a_s}(\hat{\theta}_n; Y_i)\right| \leq \begin{cases} \left(1 - \frac{1}{\sqrt{1+\sqrt{a_s}}}\right) & \textit{iff} \quad |\hat{\theta}_n - Y_i| > a_s^{1/4} \\ 2 & \textit{iff} \quad |\hat{\theta}_n - Y_i| \leq a_s^{1/4}. \end{cases} \tag{32}$$

Proof: This statement is prooven by relating the quantization as defined in (26) with conditions (31.cde). The factor 2 appears from the case $|\theta - Y| = 0$, and its counterpart condition (31.e). ■

Lemma 3 (Convergence of Empirical Scores) *Under the regularity conditions of $F_Y \in \mathcal{F}_Y$ and the Lebesgue smoothness assumption, the following inequality holds for an iid sample $Y_{1:n}$ of F_Y:*

$$\lim_{a_s \to 0} E\left[\left|r_{1:n}^{a_s}(\hat{\theta}_n, Y_{1:n})\right|\right] \leq \frac{2D}{n}. \tag{33}$$

Herefor, the convergence $\hat{\theta}_n \xrightarrow{P} \arg\min_{\theta\in\Theta} \mathcal{R}_1(\theta; F_Y)$ takes place.

Proof: Remark that from the conditions for optimality as given in the previous proof the equality $\sum_{i=1}^N \alpha_i = 0$ holds exactly. We take the expectation of expression (32) as in Proposition 2. Substitution of the term $p(a_s)$ by D/n follows from the fact that $\lim_{a_s\to 0}$, at most D different points can be interpolated by any θ and $\lim_{a_s\to 0} p(a_s) = \frac{D}{n}$. This can be made formal through use of the assumption of Lebesgue smooth distributions as e.g. in [12]. The first term in the rhs of expression (27) vanishes as $\lim_{a_s\to 0}\left(1 - \frac{1}{\sqrt{1+\sqrt{a_s}}}\right) = 0$. Taking the limit of n to infinity gives

$\lim_{n\to\infty}\lim_{a_s\to 0} E\left[\left|r^{a_s}_{1:n}(\hat{\theta}_n, Y_{1:n})\right|\right] = 0$. Interchanging the limits and application of Theorem 1 and Lemma 1 proofs convergence. ■

3 CENSORED REGRESSION

Let V be a univariate random variable representing the response corresponding with covariate X, such that the set $\{Z_i = (X_i, V_i)\}_{i=1}^n$ is iid sampled from a joint distribution $F_{XV} \subset \mathbb{R}^D \times \mathbb{R}$ and $F_{XV} \in \mathcal{F}$. Now, V_i is subject to right-censoring at point C_i, iid from $F_C \subset \mathbb{R}$. The observed response is as such $Y_i = \min(C_i, V_i)$. Let the random variable Δ denote wether the observation (X, V) is subject to censoring. Formally, for any $i = 1, \ldots, n$

$$\begin{cases} (X_i, V_i) \sim F_{XV} \\ \Delta_i = I(V_i > C_i) \\ Y_i = \min(V_i, C_i). \end{cases} \tag{34}$$

In our setup, both X, Y and Δ are fully observed. The response V_i is only observed in case $\Delta_i = 0$. Risk Minimization in case of censored responses targets the minimization of the actual risk $\mathcal{R}(\theta; F_{XV})$ through the observed random variables X and Y. Note at first that this setup is unlike the likelihood approach of the Tobit model [11, 1] describing the data generation model including the censoring process (corresponding to F_{XY}). Our approach differs as well from the censored least absolute deviation approach (CLAD) of Powell [7] where the observed values are used to fit a model $\min(C_i, \theta^T X_i)$ with a L_1-norm. In our case, a standard linear programming problem solves the estimation problem, unlike the iterative optimizers needed in the referred techniques.

At first, we extend the risk score to the case of censored observations.

Definition 4 (Risk Score of Censored Observations) *Let (X_i, Y_i) be a right-censored observation such that $\Delta_i = 1$. Let $\theta \in \Theta$ be a fixed parameter. The set of possible risk scores for this censored sample, (or the* possible risk scores*) are then*

$$r_\Delta(\theta; X_i, Y_i) = \{r(\theta; X_i, V_i) \quad \textit{s.t.} \quad V_i \geq Y_i\}. \tag{35}$$

If this set is a singleton, then the point is informative *for the loss and the fixed parameter θ.*

Let for such informative censored points $r^\delta(\theta; X_i, Y_i)$ be equal to $r_\Delta(\theta; X_i, Y_i)$ as $E[r^\delta(\theta; X_i, Y_i)] = E[r^\delta(\theta; X_i, V_i)]$. In case the non-singleton set $r_\Delta(\theta; X_i, Y_i)$ contains zero, we define $r^\delta(\theta; X_i, Y_i) = 0$, as $E[r^\Delta(\theta; X_i, V_i)] = 0$ without other knowledge on the problem. The censored empirical risk score of a fixed parameter $\theta \in \Theta$ can then be defined as

$$r^\delta_{1:n}(\theta; Z_{1:n}) = \sum_{i:\Delta_i=0} r(\theta; X_i, Y_i) + \sum_{i:\Delta_i=1,} r^\delta(\theta; X_i, Y_i), \tag{36}$$

in the case of loss where the empirical risk score takes almost surely a value in the set $\{-L, L\}$ as for ℓ_1. As such, the above expansion consists of n iid terms.

Proposition 4 (Unbiased Risk Score Estimate) *The expectation of $r^{\delta}_{1:n}$ reduces to the actual risk score as*

$$E\left[r^{\delta}_{1:n}(\theta; Z_{1:n})\right] \triangleq E\left[\sum_{i:\Delta_i=0} r(\theta; X_i, Y_i)\right] + E\left[\sum_{i:\Delta_i=1} r^{\delta}(\theta; X_i, Y_i)\right] = r(\theta; F_{XV}), \tag{37}$$

Proof: This follows from the definitions and by exploiting the iid property of the censoring mechanism. Now the previous methodology of risk scores can be applied to give finite sample properties of an estimator setting this censored empirical risk score to zero as in the following learning machine. ■

This censored empirical risk score then gives an improved estimate of the actual risk score with respect to the one that omits all censored observations. See Fig. 3 for a schematical representation for the case of the ℓ_1 loss.

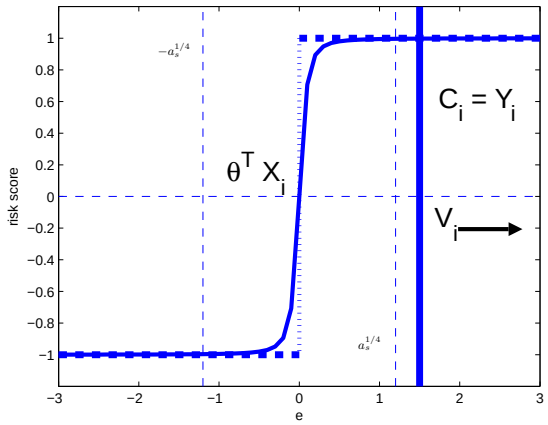

Fig. 3. *Schematical representation of a sample (X_i, V_i) which is right-censored with active censoring at C_i (and thus $\Delta_i = 1$). The observed value Y_i is the minimum $\min(C_i, V_i)$ and thus equals C_i (vertical solid line) in this case. It is clearly seen that the risk score for ℓ_{a_s} (horizontally smooth solid line) for any $V_i > Y_i$ is approximatively one for any $V_i > Y_i$.*

Definition 5 (Right-Censored Regression) *The parameter $\theta \in \mathbb{R}^D$ corresponding to the loss function ℓ describing optimally the latent variable V given its covariate X based on the censored sample of size n can be found by solving*

$$(\hat{\theta}_n, \hat{e}) = \underset{\theta\in\Theta, e\in\mathbb{R}^n}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \ell_1(e_i) \quad \textit{s.t.} \quad \begin{cases} \theta^T x_i + e_i = y_i & \forall i: \ \Delta_i = 0 \\ \theta^T x_i + e_i \leq y_i & \forall i: \ \Delta_i = 1. \end{cases} \tag{38}$$

Theorem 2 (Consistency of ℓ_1 Risk Score Estimator) *This Z-estimate $\hat{\theta}_n$ converges with high probability to the optimal risk estimate*

$$\lim_{n\to+\infty} \hat{\theta}_n = \arg\min_{\theta\in\Theta} \mathcal{R}(\theta; F_{XV}). \tag{39}$$

The key element of the proof is to proof that the empirical risk scores r^{a_s} and the Lagrange multipliers as used in the programming problem are with high probability equal for any $a_s > 0$ and $\theta \in \Theta$. The result then follows from Lemma 1. At first, let the risk score corresponding to $\ell_{a_s}(\theta; F_{XY})$ for regression be defined for all $d = 1, \ldots, D$ as

$$\frac{\partial \ell_{a_s}(\theta^T X - Y)}{\partial \theta_d} = \frac{\partial \ell_{a_s}((\theta^T X) - Y)}{\partial (\theta_d X^d)} \frac{\partial (\theta_d X^d)}{\partial \theta_d} = r^{a_s,d}(\theta; X, Y), \tag{40}$$

by application of the chain rule. Note at this point the relation with Proposition 2: this quantity can again be related to the Lagrange multipliers of the constrained problem.

Proposition 5 (Dual and Scores) *Let $\hat{\theta}_n \in \Theta$ be the optimum obtained by minimizing problem (38), and let $\alpha = (\alpha_i, \ldots, \alpha_s)^T \in \mathbb{R}^n$ be the corresponding Lagrange multipliers. For any $a_s > 0$ and for non-censored samples (X_i, Y_i) (where $\Delta_i = 0$), the following inequality holds as previously $\left|(\alpha_i X_i) - r^{a_s}(\hat{\theta}_n; X_i, Y_i)\right|_\infty \leq B(1 - |\frac{1}{\sqrt{1+\sqrt{a_s}}}|)$ whenever $|\hat{\theta}_n^T X_i - Y_i| > a_s^{1/4}$ and $\leq 2B$ otherwise. For all censored samples (X_i, Y_i) with $\Delta_i = 1$, the following inequality holds*

$$\left|(\alpha_i X_i) - r^{\delta,a_s}(\hat{\theta}_n; X_i, Y_i)\right|_\infty \leq \begin{cases} B\left(1 - \left|\frac{1}{\sqrt{1+\sqrt{a_s}}}\right|\right) & \textit{iff } \hat{\theta}_n^T X_i < Y_i - a_s^{1/4} \\ 2B & \textit{iff } |\hat{\theta}_n^T X_i - Y_i| \leq a_s^{1/4} \\ 0 & \textit{iff } \hat{\theta}_n^T X_i > Y_i + a_s^{1/4} \end{cases} \tag{41}$$

where $B < \infty$ is fixed such that $B \geq |X|_\infty$.

Proof: The censored ℓ_1 regression problem (38) can be formulated as a linear programming problem as follows

$$\min_{\theta\in\mathbb{R}^D, e\in\mathbb{R}^n} \mathcal{J}(e) = \sum_{i=1}^n e_i \quad \text{s.t.} \quad \begin{cases} -e_i \leq \theta^T X_i - Y_i \leq e_i & \forall i: \ \Delta_i \neq 1 \\ e_i \geq 0, \ \theta^T X_i - Y_i \leq e_i & \forall i: \ \Delta_i = 1. \end{cases} \tag{42}$$

The Lagrangian becomes $\mathcal{L}(e, \theta; \alpha^+, \alpha^-) = \sum_{i=1}^n e_i - \sum_{i:\Delta_i=0} \alpha_i^+ (e_i + \theta^T X_i - Y_i) - \sum_{i:\Delta_i=0} \alpha_i^- (e_i - \theta^T X_i + Y_i) - \sum_{i:\Delta_i=1} \alpha_i^+ e_i - \sum_{i:\Delta_i=1} \alpha_i^- (e_i - \theta^T X_i + Y_i)$. Due to the theorem of strong duality in linear programming problems [9], in the optimum the first order conditions for optimality w.r.t. the primal variables e_i and θ are satisfied

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial e_i} = 0 \rightarrow 1 = \alpha_i^+ + \alpha_i^- & \forall i = 1, \ldots, n \\ \frac{\partial \mathcal{L}}{\partial \theta_d} = 0 \rightarrow \sum_{i:\Delta_i=0} (\alpha_i^+ - \alpha_i^-) X_i^d - \sum_{i:\Delta_i=1} \alpha_i^- X_i^d = 0 & \forall d = 1, \ldots, D \end{cases} \tag{43}$$

Replacing $(\alpha_i^- - \alpha_i^+)$ by a new variable α_i for all i such that $\Delta_i = 0$ and $\alpha_i = -\alpha_i^-$ when $\Delta_i = 1$ gives:

$$\begin{cases} \sum_{i=1}^n \alpha_i X_i^d = 0 & \forall d = 1, \ldots, D \quad (a) \\ -1 \leq \alpha_i \leq 1 & i: \ \Delta_i = 0 \quad (b) \\ 0 \leq \alpha_i \leq 1 & i: \ \Delta_i = 1 \quad (c) \\ \alpha_i = 1 \Leftrightarrow (\theta^T X_i - Y_i) > 0 & \forall i = 1, \ldots, n \quad (d) \\ \alpha_i = -1 \Leftrightarrow (\theta^T X_i - Y_i) < 0 & i: \ \Delta_i = 0 \quad (e) \\ -1 < \alpha_i < 1 \Leftrightarrow (\theta^T X_i - Y_i) = 0 & i: \ \Delta_i = 0 \quad (f) \\ \alpha_i = 0 \Leftrightarrow \theta^T X_i < Y_i & i: \ \Delta_i = 1 \quad (g) \\ 0 < \alpha_i < 1 \Leftrightarrow (\theta^T X_i - Y_i) = 0 & i: \ \Delta_i = 1 \quad (h) \end{cases} \tag{44}$$

We relate the quantity to the risk scores $r^{\delta,a_s}(\theta; Z_i)$ as follows. Relating this with the definition of the risk scores r^{δ,a_s} as in (25) and conditions (44.bde) shows that one can use $\alpha_i X_i^d$. as quantized versions of the risk score for ℓ_{a_s} where $a_s \to 0$. ■

Corollary 2 (Bounded Expected Deviation) *Let $\hat{\theta}_n \in \Theta$ be the minimizer of problem (38), then the expected deviation of the censored empirical risk score to zero is bounded as follows*

$$E\left[|r_{1:n}^{a_s,\delta}(\hat{\theta}_n; F_{XV})|\right] \leq (1 - p(a_s))B\left(1 - \frac{1}{\sqrt{1+\sqrt{a_s}}}\right) + 2Bp(a_s). \tag{45}$$

where $p(a_s)$ denotes the probability mass

$$0 \leq p(a_s) \triangleq \sup_{\theta \in \Theta} P_{XV}\left(\left\{(X,V) \;\middle|\; |\theta^T X - V| \leq a_s^{1/4}\right\}\right) \leq 1. \tag{46}$$

Given this result, Theorem 2 is proved as before in Lemma 3.

4 CONCLUSION

This paper advances results relating results of learning theory with notions which are indispensable in optimization theory. Hereto, we outlined an approach for incorporating the notion of risk scores and Lagrange multipliers in the theory of empirical risk minimization. Convergence of an empirical Z-estimators and the actual risk minimizing solution is shown. This framework relies on differentiability and Lipschitz boundedness of the loss function, but it is shown that ideas can be extended to the non-differentiable L_1-norm risk by using an appropriate approximation. Finally, it is shown how this methodology can be used to design a learning machine for regression in the context of censored observations.

Acknowledgments. This research work was carried out at the ESAT laboratory of the KULeuven. Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666, GOA-Ambiorics IDO, several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0407.02, G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0499.04, G.0211.05, G.0080.01, research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s, STWW-Genprom, GBOU-McKnow, Eureka-Impact, Eureka-FLiTE, several PhD grants); Belgian Federal Government: DWTC IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006) (2002-2006), Program Sustainable Development PODO-II (CP/40); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS is an associate professor and BDM is a full professor at K.U.Leuven Belgium, respectively.

References

1. T. Amemiya. Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41(6):997–1016, 1973.
2. M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
3. T. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–224, 1965.
4. L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.
5. P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.
6. L.P. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*. Springer Series in Statistics. Springer, 2nd. edition edition, 1986.
7. J.L. Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25:303–325, 1984.
8. C.C. Pugh. *Real Mathematical Analysis*. Springer-Verlag, 2002.
9. R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
10. W. Rudin. *Principles of Real Analysis*. McGraw-Hill, New-York, 1976.
11. J. Tobin. Estimatin of relationships for limited dependent variables. *Econometrica*, 26:22–36, 1958.
12. A. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.
13. V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.
14. C.R. Vogel. *Computational Methods for Inverse Problems*. Siam, 2002.