Annals of the Institute of Statistical Mathematics (2023) 75:335–368
https://doi.org/10.1007/s10463-022-00847-1

Quantitative robustness of instance ranking problems

Tino Werner[1]

Received: 4 November 2021 / Revised: 3 May 2022 / Accepted: 7 July 2022 /
Published online: 30 August 2022

Abstract
Instance ranking problems intend to recover the ordering of the instances in a data set with applications in scientific, social and financial contexts. In this work, we concentrate on the global robustness of parametric instance ranking problems in terms of the breakdown point which measures the fraction of samples that need to be perturbed in order to let the estimator take unreasonable values. Existing breakdown point notions do not cover ranking problems so far. We propose to define a breakdown of the estimator as a sign-reversal of all components which causes the predicted ranking to be potentially completely inverted; therefore, we call it the order-inversal breakdown point (OIBDP). We will study the OIBDP, based on a linear model, for several different carefully distinguished ranking problems and provide least favorable outlier configurations, characterizations of the order-inversal breakdown point and sharp asymptotic upper bounds. We also compute empirical OIBDPs.

Keywords Breakdown point · Quantitative robustness · Instance ranking problems · Sparsity

1 Introduction

A well-known issue when analyzing data is that the data usually are not clean but consist of perturbations that can severely distort an estimator. Instances that are distant from the majority of the data are often termed as "outliers". For a well-founded analysis, it is required to define an underlying ideal model so that the data points are interpreted as independent realizations from this model. However, wrong model assumptions let the real data appear as contaminated data. In these cases, methods

The online version of this article contains supplementary material.

✉ Tino Werner
tino.werner1@uni-oldenburg.de

[1] Institute for Mathematics, Carl von Ossietzky University Oldenburg, Carl-von-Ossietzky-Strasse 9-11, P/O Box 5634, 26046 Oldenburg, Germany

from robust statistics (Huber and Ronchetti, 2009; Hampel et al., 1986; Rieder, 1994; Maronna et al., 2019) to handle these phenomena are necessary to incorporate even contaminated data points appropriately since just removing possible outliers is the wrong way as for example discussed in Hampel et al. (1986). Even worse, in contrast to the classical convex contamination model [see, e.g., (Rieder, 1994, Sec. 4.2)] where an instance either stems from a contaminated distribution or from the ideal distribution, the cell-wise outlier model from Alqallaf et al. (2009) allows for contaminating the single predictor components for each instance independently which causes the probability to have even one clean instance in the data to tend to zero which again is a manifestation of the curse of dimensionality.

Robust statistics provides two concepts to measure the quantitative robustness of an estimator. Since robust statistics identifies estimators as statistical functionals (Huber and Ronchetti, 2009; Hampel et al., 1986; Rieder, 1994; Maronna et al., 2019), functional derivatives (e.g., Averbukh and Smolyanov, 1967) can be applied in order to linearize this functional in a first-order expansion which goes back to Von Mises (1947). The functional derivative, usually a Gâteaux derivative, has been identified in Hampel (1974) with the influence curve which is an important diagnostic tool which measures the infinitesimal impact of one data point on the estimator.

In contrast to the influence curve which quantifies the local robustness of an estimator, i.e., only allowing for an infinitesimal fraction of the data being contaminated, the breakdown point (BDP), introduced in Hampel (1971, Sec. 6) in a functional version and in Donoho and Huber (1983) in a finite-sample version, studies the global robustness of an estimator. The finite-sample BDP from Donoho and Huber (1983) quantifies the minimum fraction of instances in a data set so that contaminating any such fraction of data points arbitrarily can let the estimator "break down", while the functional BDP essentially quantifies the allowed maximum Prokhorov distance between the ideal and the contaminated distribution without the estimator breaking down. There has yet been a lot of work on BDPs, see for example (Rousseeuw, 1984, 1985; Davies, 1993; Hubert, 1997; Genton, 1998; Becker and Gather, 1999; Gather and Hilker, 1997) or (Donoho and Stodden, 2006) which cover location, scale, regression and spatial estimators and (Hubert et al., 2008) who study the BDP for multivariate estimators. Recently, a BDP for classification (Zhao et al., 2018) and for multiclass-classification (Qian et al., 2019) has been proposed.

While regression and classification aim for an exact fit of the response value, there exist types of problems where one is only interested in an ordering of the instances and not of the particular responses. These problems are ranking problems which are very important in for example in document ranking (Page et al., 1999; Herbrich et al., 1999a; Cao et al., 2006), medicine (Agarwal and Sengupta, 2009), credit risk-screening (Clémençon et al., 2013b) or biology and chemistry (Agarwal, 2010; Kayala et al., 2011; Morrison et al., 2005). Due to the global nature of ranking problems where essentially each instance pair is compared, the existing global robustness measures, i.e., the existing BDP concepts, are not suitable here.

Consider the problem to order instances in a data set. If responses are available, this can be identified with minimizing a pair-wise loss function, i.e., which operates on pairs of responses and their predictions by checking if their ordering coincides, as shown in the seminal work of (Herbrich et al., 1999a, b). We then speak of

instance ranking problems in the terminology of Fürnkranz and Hüllermeier (2011). Clémençon et al. (2008) proposed the statistical framework for such instance ranking problems which emerge from ordinal regression (Herbrich et al., 1999a) and proved that the common approach of empirical risk minimization (ERM) is indeed suitable for such ranking problems. There are three ways of casting a ranking problem, i.e., either the ordering of all instances has to be correct (hard ranking), one just wants to identify the top K instances for a given K (weak ranking, (Clémençon and Vayatis, 2007)) or the best K instances have to be identified and the ordering of these instances has to be correct [localized ranking, Clémençon and Vayatis (2007)]. Furthermore, one distinguishes between binary responses which lead to binary or bipartite ranking problems [e.g., Joachims (2002), Freund et al. (2003), Clémençon and Vayatis (2010)], d-partite ranking problems for categorical responses with d categories [e.g., Clémençon et al. (2013c), Fürnkranz et al. (2009)] and continuous ranking problems (Sculley, 2010; Clémençon and Achab, 2017).

Instance ranking problems are usually solved by learning a real-valued, here parametric, scoring function which assigns a score to each instance with the goal to minimize some ranking error between the predicted ordering of the instances according to the scores and the true ordering. The peculiarity of ranking problems is that they have an inherent equivariance nature, i.e., multiplying each response with the same positive factor or adding the same fixed value to it does not alter the ordering. Therefore, the norm of the coefficient which is considered by the regression BDP is not suitable here. It is out of question that for a ranking prediction, it would be even worse to predict an inverted ordering than to perform random guessing (to which zero or infinite coefficients essentially correspond) which exactly motivates our so-called OIBDP which is the minimum fraction of perturbed data points so that the nonzero coefficient components can be inverted. At the first glance, we get unreasonable BDPs in high-dimensional settings, i.e., if the predictor dimension is no longer smaller than the number of observations, but this can be remedied by assuming sparse underlying models resp. sparse model selection which is natural in such settings [e.g., Bühlmann and Van De Geer (2011)].

Our contribution is threefold: **(i)** We propose the definition of the order-inversal BDP for ranking problems which embeds the BDP concept of robust statistics into that area of machine learning; **(ii)** we provide explicit worst-case outlier configurations; and **(iii)** we compute upper bounds for the corresponding OIBDPs for different ranking problems.

The rest of this work is organized as follows. Section 2 compiles necessary preliminaries in terms of a more concise definition of the loss functions corresponding to the different ranking problems as well as the BDP concept. In Sect. 3, we show why neither the classical BDP for regression nor the angular BDP for classification is suitable for ranking problems and propose the OIBDP for ranking. In Sects. 4, 5 and 6, we propose outlier schemes and prove asymptotic bounds for the OIBDP for hard continuous resp. hard binary and hard d-partite resp. localized continuous ranking problems. In Sect. 7, we discuss the applicability of BDP concepts to the remaining instance ranking problems. In Sect. 8, we relate the computed BDPs to sparse underlying models and outline how robust ranking can be achieved in practical applications. We also provide empirical OIBDPs based on simulated and real

data. Section 9 is an outlook devoted to SVM- resp. SVR-type approaches. Further results and selected proofs can be found in the supplementary file.

2 Preliminaries

Let $D = (X, Y)$ be a data set with regressors $X_i \in \mathcal{X} \subset \mathbb{R}^p$ and responses $Y_i \in \mathcal{Y} \subset \mathbb{R}$. We start by revisiting suitable loss functions for different types of instance ranking problems and the breakdown point concept.

2.1 Ranking problems and motivating example

Example 1 Consider data from tax fraud detection where for n tax payers, represented by predictor vectors $X_i \in \mathbb{R}^p$, one has information about their past tax compliance. This information is either represented by a binary response variable (fraudulent/compliant) or by a continuous pseudo-response variable, e.g., the damage (which is negative if the corresponding tax payer indeed gets a refund). Due to the limited capacities of finance offices, it is desirable to use machine learning in order to facilitate the selection of income tax statements that need to be investigated further. This strategy is known as risk-based auditing [e.g., Pickett (2006)]. Since just classifying instances as fraudulent or compliant leads to the potential problem that there are more income tax statements classified as fraudulent than the finance offices can investigate, i.e., one has a prioritization problem. Training a ranking model which learns an ordering of the instances, i.e., in which order the income tax statements should be investigated, avoids this, as already done for example in Torgo and Ribeiro (2007).

Assuming that one has continuous (pseudo-)responses, a regression model should predict $\hat{Y}_i$ that are close to the true Y_i. In contrast, a ranking model works perfectly if it correctly predicts the orderings, i.e., if $Y_j > Y_i$, any predictions Y_i and Y_j so that $\hat{Y}_j > \hat{Y}_i$ holds is correct, while the predicted values do not need to be close to the true responses.

However, it is well-known from regression that one single outlying instance can make the regression model worthless. Therefore, it is important to investigate the robustness of ranking problems, i.e., whether outliers can perturb a ranking model and whether it is possible to make the ranking model completely worthless, which would especially be the case if the predicted ordering inverts the true ordering of all instances, leading to rigorous investigations of the most compliant tax payers.

In this work, we assume a linear model $Y_i = s_\beta(X_i) + \epsilon_i$ for the ϵ_i being i.i.d. realizations from some centered distribution with existing second moments and a scoring function $s_\beta : \mathcal{X} \to \mathbb{R}$ for some parameter $\beta \in \Theta \subset \mathbb{R}^p$. X_{ij} refers to the j-th entry of row X_i and $X_{\cdot,j}$ to the j-th column of X. The j-th component of the coefficient vector β is denoted by β_j.

In the case of hard ranking problems, the goal is to retrieve the correct ordering of all instances. Therefore, the parameter β in the linear model is optimized according to

$$\min_{\beta\in\Theta}\left(L_n^{hard}(\beta)=\frac{1}{n(n-1)}\sum\sum_{i\neq j}L(X_i,X_j,Y_i,Y_j,\beta)\right) \tag{1}$$

where $L:\mathcal{X}\times\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\times\Theta\to[0,\infty]$ is some loss function that compares instance pairs. In (Herbrich et al., 1999b) or (Clémençon et al., 2008), L is the indicator function

$$L(X_i,X_j,Y_i,Y_j,\beta)=I((Y_i-Y_j)(s_\beta(X_i)-s_\beta(X_j))<0)$$

which just checks whether a misranking occurred, i.e., if the true resp. the predicted pair-wise orderings did not coincide, but the actual magnitude of the product is not taken into account. Since this loss function is not even continuous, one often considers surrogate losses (see Werner, 2021a for an overview). In the following BDP computations, we will always consider general loss functions that can be rewritten as $L((Y_i-Y_j)(s_\beta(X_i)-s_\beta(X_j))$ in the same manner as classification loss functions are frequently rewritten as $L(ys_\beta(x))$.

For weak ranking problems (Clémençon and Vayatis, 2007), the goal is to correctly retrieve the best K instances for a user-defined K without ordering them internally. The empirical counterpart of the misclassification risk can be expressed by

$$L_n^{\text{weak},K}(\beta)=\frac{2}{n}\sum_{i\in Best_K}I(\text{rk}\,(s_\beta(X_i))>K) \tag{2}$$

with the set $Best_K$ of the true top K indices where the ranks correspond to a descending ordering. Again, the indicator function may be replaced by any classification loss function $L:\mathcal{X}\times\mathcal{Y}\times\beta\to[0,\infty]$.

Localized ranking problems (Clémençon and Vayatis, 2007) aim to correctly retrieve the best K instances but also to order these K instances correctly. The optimization problem is

$$\begin{aligned}L_n^{loc,K}(\beta)\ :=\ &\frac{n-K}{n}L_n^{\text{weak},K}(\beta)+\frac{2}{n(n-1)}\\&\sum_{i<j,i,j\in Best_K}\sum I((s_\beta(X_i)-s_\beta(X_j))(Y_i-Y_j)<0).\end{aligned} \tag{3}$$

One can again replace the indicator function by surrogates. Note that one may replace the set $\widehat{Best_K}$ in the double sum by $Best_K$. We will discuss both cases in Sect. 6.

For further discussions of these loss functions and for instance ranking, see Werner (2021a).

2.2 Quantitative robustness

Robustness of an estimator can be understood in the sense that it allows for perturbations or even large contaminations of the underlying sample without the quality of the estimator being significantly affected. One can distinguish between quantitative and qualitative robustness. The latter goes back to Hampel (1971) and essentially indicates the continuity of the underlying statistical functionals.

As for quantitative robustness, one further has to distinguish between global and local robustness. Local robustness is devoted to the effect of small perturbations of the data where the term "small" means that, for finite samples, only one observation may be contaminated, so in other words, the influence curve or influence function which is the diagnostic tool for local robustness measures the infinitesimal influence of a single observation on the estimator. In contrast, global robustness allows for large perturbations, i.e., a considerable fraction of the data points being contaminated arbitrarily. The maximum fraction which an estimator can cope with, i.e., without taking unreasonable values, is measured by the breakdown point.

2.2.1 The breakdown point concept

Let Z_n be a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let $\hat{\beta}(Z_n)$ be the estimated coefficient for the scoring function s_β based on Z_n. The finite-sample BDP of Donoho and Huber (1983) is defined as follows.

Definition 1 The **finite-sample breakdown point** of an estimator $\hat{\beta}$ is defined as

$$\epsilon^*(\hat{\beta}, Z_n) = \min \left\{ \frac{m}{n} \;\middle|\; sup_{Z_n^m}(||\hat{\beta}(Z_n^m)||) = \infty \right\} \tag{4}$$

where Z_n^m denotes any sample that has exactly $(n - m)$ instances in common with Z_n, i.e., m instances can be modified arbitrarily.

Note that this definition assumes that $\beta \in \mathbb{R}^p$. In cases where $\beta \in \Theta \subset\subset \mathbb{R}^p$, the situation would get more difficult since here a breakdown may be defined in the sense that $\hat{\beta}$ is located at the boundary of Θ. In this case, one would require some transformation that moves the boundaries of Θ to infinite values, see for example (He, 2005) who proposed to use a log-transformation for computing the BDP of scale estimators in order to move the value 0 to $-\infty$.

A variety of extensions of the BDP concept have been proposed in the literature. Stromberg and Ruppert (1992) and Sakata and White (1995) proposed BDPs for regression, Sakata and White (1998) suggested a BDP definition for location-scale estimators, while Genton (1998) propose the spatial BDP for variogram estimators and Genton and Lucas (2003) and Genton (2003) introduce a BDP for dependent samples (time series). Donoho and Stodden (2006), Donoho (2006) propose a BDP for model selection, Kanamori et al. (2004) study the BDP for SVMs and Hennig (2008) transferred the BDP concept to the dissolution point concept for clustering. Ruckdeschel and

Horbenko (2012) suggest an expected BDP that respects the ideal distribution of the original data. See Davies and Gather (2005) for a notable discussion paper on BDPs.

2.2.2 Angular breakdown point for classification

Recently, Zhao et al. (2018) proposed the following definition of a breakdown point that is suitable for classification, calling it "angular breakdown point" since it is based on the angle between the decision hyperplane of the original coefficient and the one estimated on a contaminated sample. The following definition stems from Zhao et al. (2018, Def. 1) and assumes linear classifiers.

Definition 2 *(Angular breakdown point for classification)* The **(population) angular breakdown point for classification** is given by

$$\epsilon(\beta, Z_n) = \min \left\{ \frac{m}{n} \;\middle|\; \hat{\beta}(Z_n^m) \in S^- \right\}, \quad S^- = \{\tilde{\beta} \mid \tilde{\beta}^T \beta \leq 0\}. \tag{5}$$

Zhao et al. (2018, Def. 1') also proposed a sample counterpart of this breakdown point where β is replaced by $\hat{\beta}(Z_n)$ and therefore S^- by $\hat{S}^-$ with the respective replacement. The angular breakdown point indicates that modifying more than $\epsilon(\beta, Z_n)$ of the sample Z_n by arbitrary points can induce an angle between the original decision hyperplane and the hyperplane of the coefficient corresponding to the contaminated sample of at least $\pi/2$, leading to very low discriminative power if the classifier corresponding to $\hat{\beta}(Z_n)$ was sufficiently well. This setting has been extended to multi-class classification in Qian et al. (2019).

3 Outliers and breakdown for ranking with linear scoring functions

As a motivation, we consider continuous ranking problems where the responses are continuously valued (taking values in wlog. the whole space $\mathbb{R}$) in this section.

3.1 Why neither the regression nor the classification breakdown point work

We start by proving a counterpart of Zhao et al. (2018, Prop. 3.1) showing that the finite-sample breakdown point in Eq. 4 in also not reasonable in the ranking context. To this end, let the objective function of regularized continuous ranking with linear scoring functions be given by

$$L_{\lambda,n}(b, \beta, Z_n) = \frac{1}{n(n-1)} \sum_{i<j} \sum L((Y_i - Y_j)(s_{b,\beta}(X_i) - s_{b,\beta}(X_j))) + J_\lambda(\beta)$$

where $s_{b,\beta}(x) := x\beta + b$ so that $s(X_i) =: \hat{Y}_i$ is a parametric scoring function for some optional intercept b with $|b| < \infty$, a loss function L as introduced in Sect. 2 and a regularizer $J_\lambda(\beta)$ satisfying

$$\begin{aligned} &\textbf{i)}\ J_\lambda \geq 0, J_0 \equiv 0, \quad \textbf{ii)}\ J(\beta) = 0 \iff \beta = 0_p, \\ &\textbf{iii)}\ J(-\beta) = J(\beta), \quad \textbf{iv)}\ J(\beta) \overset{||\beta|| \to \infty}{\longrightarrow} \infty, \end{aligned} \tag{6}$$

where 0_p is the vector of length p containing only zeroes. The fourth property is also known as coercivity [e.g. Werner (2006)]. The regularizer encourages sparse models and therefore does not take the intercept b into account.

Definition 3 A sample Z_n is **linearly inrankable** if there exists no linear scoring function (linear in β) $s_{b,\beta}(x) = x\beta + b$ such that we can perfectly replicate the ranking of the responses in Z_n; otherwise, we call the sample **linearly rankable**.

Graphically, in the most simple case, linear rankability can be easily understood in the sense that ordering the instances w.r.t. their j-th component leads to a perfect ordering of the responses, for all j. If the re-ordered responses are strictly monotonically increasing, $\beta_j > 0$ perfect replicates their ordering and vice versa. If this holds for all j, the coefficient β consisting of these β_j also retrieves the ordering perfectly. However, linear inrankability **is generally not given** if this property is violated for some axes.

Example 2 Consider the sample ((1, 1), 1), ((2, 4), 2), (2.5, 10), 50). Evidently, linear rankability is given and each coefficient with $\beta_1, \beta_2 > 0$ leads to a perfect ranking. Now, consider the sample ((1, 1), 1), ((0, 3), 2), ((3, 2), 3). In contrast to the sample before, the responses clearly do neither strictly monotonically increase or decrease with increasing $X_{\cdot,1}$ nor with increasing $X_{\cdot,2}$. However, for $\beta := (1, 1)$ and arbitrary but finite b, we have $\hat{Y}_1 = 2 + b, \hat{Y}_2 = 3 + b, \hat{Y}_3 = 5 + b$, so the ranking is perfect.

This example points out that linear inrankability does not only depend on some strict monotonicity of the responses w.r.t. some variable but also on the variables themselves (unless the strict monotonicity is satisfied along all axes). This is a first motivating aspect which makes the angular breakdown point from Zhao et al. (2018) inappropriate for ranking. Let us state the following counterpart to Zhao et al. (2018, Prop. 3.1) whose proof can be found in the supplementary file.

Lemma 1 *Let L be a nonnegative loss function, $L(0) < \infty$, and let assumptions (6) hold.*

(a) For $\lambda > 0$, it holds that $||\hat{\beta}(Z_n)|| < \infty$ and $||\hat{\beta}(Z_n^m)|| < \infty$ for any Z_n and Z_n^m.
(b) For $\lambda = 0$, norm finiteness of the estimated coefficient cannot be guaranteed.

Lemma 1 indicates that the usual finite-sample breakdown point in Eq. 4 is insufficient for measuring the robustness of regularized ranking problems since any contamination keeps the norm of the estimated coefficient finite if the assumptions in the lemma hold. As for the angular BDP for classification introduced in Zhao et al. (2018), we similarly can conclude that it is inappropriate for

ranking if the variables are scaled differently or if they take values in different spaces.

Example 3 Let the sample ((5, 0.2), 0.9), ((6, 0.3), 1.2), ((1, 0.1), 0.3) be given an let $\beta = (0.1, 2)$ be the true coefficient (wlog. let $b = 0$). Then for $\tilde{\beta} = (0.2, -1)$ we have $\beta^T \tilde{\beta} < 0$, but the ordering of the predictions w.r.t. $\tilde{\beta}$ is still correct, making the angular BDP insufficient for ranking.

Remark 1 *(Interpretation of the classical BDP)* Let us additionally highlight the fact that one has to be very cautious in interpreting the classical BDP. Focusing on the values $\pm\infty$ can be highly misleading since it does not fully reflect the real meaning of the BDP.

Consider a least squares regression estimator. Having control over one of the observations allows us to produce an arbitrary estimated regression coefficient, as the proof of of Alfons et al. (2013, Thm. 1) reveals. However, it is impossible to achieve an estimated coefficient with $||\beta|| = \infty$ for arbitrary data. Just consider the case $p = 1$. A breakdown can be achieved by for example letting the response of the mostright observation which is w.l.o.g. (X_n, Y_n) grow. Clearly, the regression coefficient would also grow. Now, consider the limit case that the contaminated response takes the value ∞. This does not enforce "$\beta = \infty$" but allows for any coefficient. For example, $\beta = 0$ produces an infinite loss on the n-th instance, but the same loss is suffered for any other finite β. The case "$\beta = \infty$" would similarly lead to an infinite loss on any other instance, while it would be hard to meaningfully define the loss on the n-th instance. Therefore, this limit case would essentially make all coefficients equal and definitely not enforce an infinite coefficient.

Example 3 and Remark 1 motivate a new BDP notion for instance ranking problems.

3.2 The order-inversal breakdown point for ranking

We have seen that reverting the sign of some components of the coefficient does not guarantee any effect on the ranking quality unless all coefficient components are sign-reverted. Even this does not guarantee an inverted ordering but guarantees that the predicted ordering cannot be perfect anymore. Taking all these arguments into account, we now state the following definition for the OIBDP for ranking.

Definition 4 *(Order-inversal breakdown point for ranking)***(a)** The population order-inversal breakdown point for ranking is defined by

$$\check{\epsilon}(\beta, Z_n) := \min \left\{ \frac{m}{n} \;\middle|\; \hat{\beta}(Z_n^m) \in S_\cap^- \right\}, \quad S_\cap^- := \bigcap_{j:\beta_j \neq 0} \{\tilde{\beta}_j \mid \tilde{\beta}_j \beta_j < 0\}.$$

(b) The sample order-inversal breakdown point for ranking is defined by

$$\check{e}(\hat{\beta}, Z_n) := \min\left\{ \frac{m}{n} \;\middle|\; \hat{\beta}(Z_n^m) \in \hat{S}_\cap^- \right\}, \quad \hat{S}_\cap^- := \bigcap_{j:\hat{\beta}_j(Z_n)\neq 0} \{\tilde{\beta}_j \mid \tilde{\beta}_j \hat{\beta}_j(Z_n) < 0\}.$$

One could ask why we do not include the case that the $\hat{\beta}_j$ are zero in the definition of the OIBDP. The reason is that this definition should be tailored to ranking problems and their inherent property of interest which are the orderings of the instances. Of course, allowing several estimated coefficient components to become zero where the original ones are nonzero, while the others being sign-reverted also leads to a nonperfect ranking; however, this situation refers to model selection itself. Evidently, not selecting all relevant variables leads to models without predictive power, but this is a topic for its own which has recently been studied in Werner (2021b).

Remark 2 *(Nonlinear scoring functions)* The restriction to linear scoring functions (i.e., linear in x) is not necessary since sign-reverting all components of β would clearly also revert the ordering for any scoring function of the form $s_{b,\beta}(x) = f(x)\beta + b$ where $f : \mathbb{R}^p \to \mathbb{R}^{p'}$ maps the regressors from the original regressor space $\mathcal{X}$ to some feature space $\mathcal{X}' \subset \mathbb{R}^{p'}$ where we allow $p \neq p'$ which refers for example to very natural situations like facing categorical regressors whose encoding enlarges the column number of the regressor matrix. The respective outlier configurations that we provide in the remainder then have to be concentrated on regions where the score is strictly monotonic w.r.t. the original coefficient.

The reduction to linear scoring functions is done for the sake of simplicity and illustrativeness and no restriction (as long as our scoring functions are still linear in the parameter β which evidently is the case) since one could essentially approximate any nonlinear scoring function by piece-wise linear scoring functions. The case of kernel-based scoring functions will be discussed in Sect. 9.

4 Asymptotic upper bounds for the OIBDP of the hard continuous ranking problem

4.1 Univariate case

Assumption 1 Let L be a continuous and strictly monotonically decreasing function with $\lim_{u\to\infty}(L(u)) = 0$ and $\lim_{u\to-\infty}(L(u)) = \infty$.

Unbounded loss functions arise once convex surrogates are used, as in RankBoost (Freund et al., 2003), RankingSVM (Herbrich et al., 1999a; Joachims, 2002) or the p-Norm-Push (Rudin, 2009).

Remark 3 *(Ties)* Zhao et al. (2018) also respect the case of zero coefficients. As shown in Werner (2022), the case of ties has to be handled differently, but we always assume that ties have zero probability in continuous ranking problems in this work.

Remark 4 *(Nonzero assumption)* We always assume that the true coefficient is never 0_p. Together with the assumption of linear rankability with the original coefficient that we use in the proofs, this is the counterpart of assuming that the points are in general position.

Remark 5 *(Immunization against particular regularization terms)* We will not consider the regularization term directly in the following lemmas and theorems. Having $\lambda > 0$ which defines a feasible set of the form $B_{r,c_\lambda} := \{\beta \mid ||\beta||_r \leq c_\lambda\}$ for some $0 < c_\lambda < \infty$ and some $r > 0$, we can always assume that we project the true coefficient onto this set by standardizing all components uniformly (which does not alter the ranking) since the issue of a sign-reversal does not depend on the magnitude of the respective coefficient components. We will discuss to the case $r = 0$ in Sect. 8.

Lemma 2 *Let $p = 1$. For the hard ranking problem with the loss function $L(u) = I(u < 0)$, the sample and population OIBDP for ranking is*

$$\frac{\breve{m}}{n}, \quad \breve{m} = \min \left\{ m \,\middle|\, m(n-m) + \frac{m(m-1)}{2} > \frac{(n-m)(n-m-1)}{2} \right\} \tag{7}$$

and asymptotically, the BDP is given by $1 - \sqrt{0.5}$.

Proof The proof is given for the population version; the sample version is proven completely analogously. Assume wlog. that $\beta > 0$. Consider the worst-case outliers shown in Fig. 1.

A correct ranking of two instances does not suffer a loss, while any incorrect ranking suffers the same loss. This makes it impossible to achieve a breakdown by letting the response of one single outlier tend to $-\infty$ for all n that are reasonably high (≥ 4). Observe that, due to symmetry, we have $n(n-1)/2$ effective pairwise comparisons and that a single outlier like the rightmost point in Fig. 1 leads to $(n-1)$ misrankings for a coefficient $\beta > 0$. Consider to add one outlier. Then, comparing each of the $(n-2)$ noncontaminated instances with one outlier leads to $(n-2)$ misrankings for $\beta > 0$, but since the ordering of the outliers is also incorrect, we get a total of $2(n-2)+1$ misrankings. Now, let $m \geq 1$ outliers be given. Then, we get a total of $m(n-m) + m(m-1)/2$ misrankings for $\beta > 0$, while the number of misrankings that we make for $\beta < 0$ is evidently given by $(n-m)(n-m-1)/2$ since every pair of clean observations of which we have $(n-m)$ ones is misranked, so the number of outliers $\breve{m}$ that we require for a breakdown is as stated in Eq. 7.

Figure 2 shows the sample OIBDP for ranking for $n \in \{4, 5, ..., 500\}$. Note that the maximal sample BDP is attained for $n = 4$ where $\breve{m} = 2$ and the minimal sample BDP is attained for both $n = 7$ and $n = 14$, namely $\breve{m}/n = 2/7$.

As for the asymptotic setting, we set $m = cn$ and solve the inequality given in 7 for c.

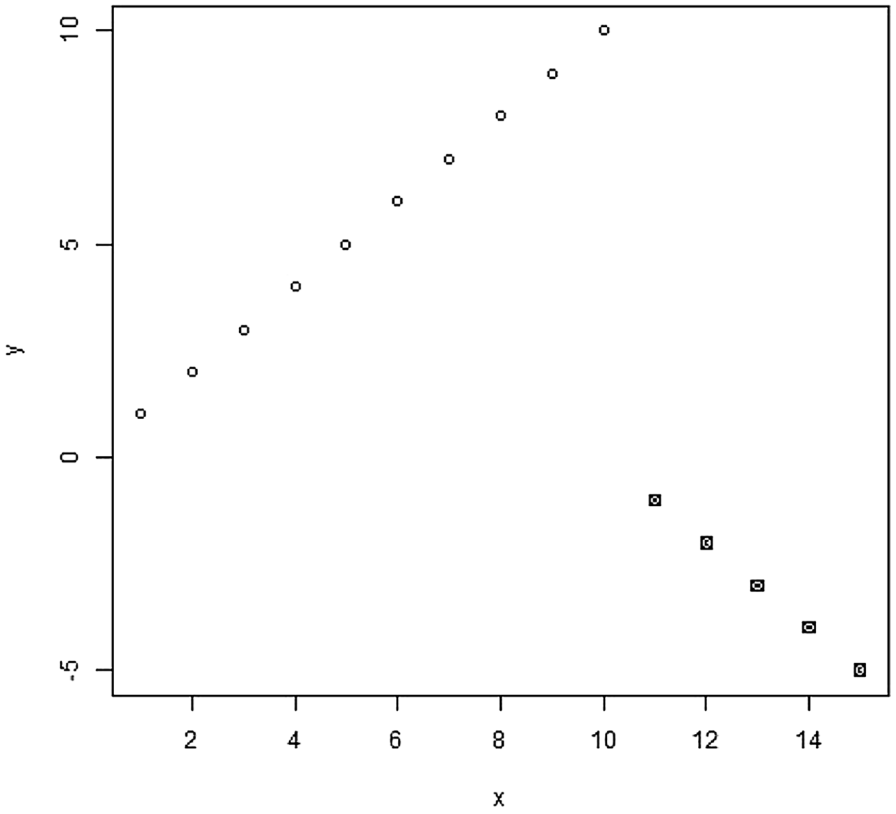

Fig. 1 Worst-case outliers for $p = 1$

$$cn^2 - c^2n^2 + \frac{c^2n^2}{2} - \frac{cn}{2} \overset{!}{>} \frac{(n - cn)(n - cn - 1)}{2} = \frac{n^2 - 2cn^2 + c^2n^2 + cn - n}{2}$$
$$\overset{n>0}{\iff} n\left[-c^2 + 2c - \frac{1}{2}\right] + [-c + 0.5] \overset{!}{>} 0$$

where the notation $\overset{!}{>} 0$ indicates that we search for c so that this inequality holds. Asymptotically, we just require that the value in the bracket of the left-hand side is positive. We can easily conclude that this holds for $c > 1 - \sqrt{0.5}$, so this value is the sharp asymptotic upper bound for the OIBDP. □

Remark 6 Note that this result equals the asymptotic BDP of the Hodges-Lehmann estimator [see Hodges Jr (1967, Sec. 11)]. This is not surprising since the Hodges-Lehmann estimator is given as the median of the set of all possible pairs of univariate samples. In order to achieve a breakdown of such an estimator, at least the half of the underlying observations which, in case of the Hodges-Lehmann estimator, are pairwise comparisons, have to be contaminated. This equals our setting since for the Hodges-Lehmann estimator, the sum of the outlier–outlier pairs and the outlier–non-outlier pairs has to be more than the half number of data points.

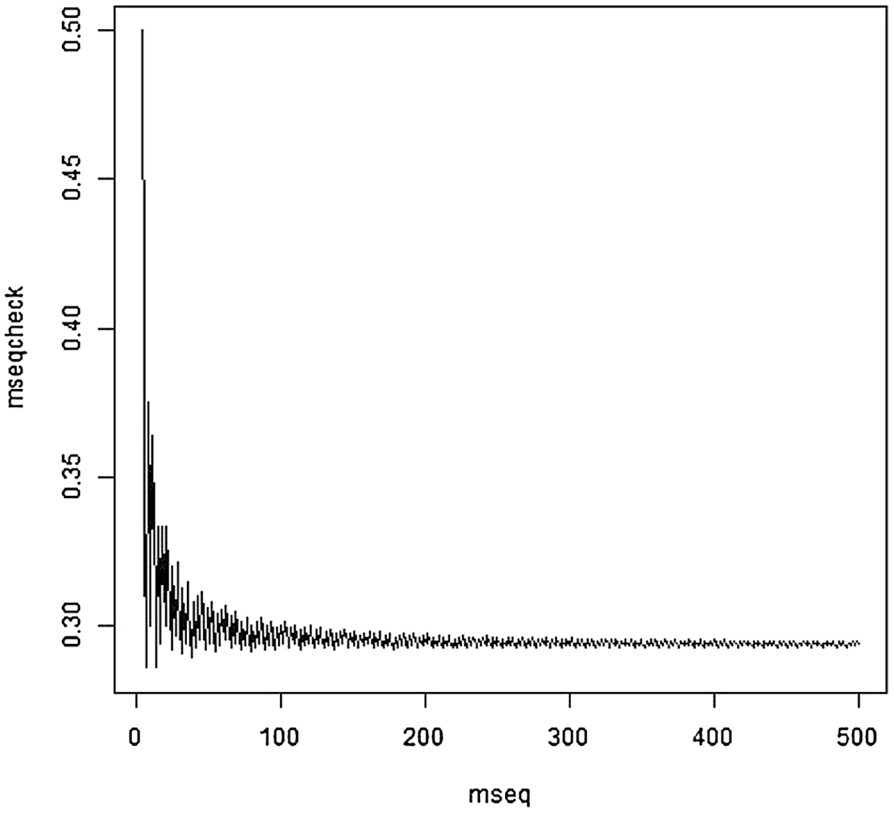

Fig. 2 OIBDP for hard ranking for $p = 1$

We do not consider unbounded loss functions here since we cover this case with Theorem 1 in the next subsection. Before we proceed with bounded loss functions, we argue why it suffices to consider indicator functions here.

Remark 7 *(Reduction to indicator loss functions)* Assume that the loss function is bounded, i.e., $\lim_{u\to-\infty}(L(u)) = C_l < \infty$. Then, we can obviously generate losses w.r.t. the original coefficient on the contaminated instances which are (close to) C_l by using the outlier scheme introduced above. However, since BDP computations have to consider all possible data configurations, we also have to take into account that the losses w.r.t. the broken coefficients on the original instances may be similarly close to C_l. Therefore, such extreme cases essentially lead to losses that are arbitrarily close to zero and arbitrarily close to C_l, allowing for a reduction to the indicator loss function case.

4.2 Multivariate case

We considered the case $p = 1$ separately since the arguments and results for $p > 1$ are different for ranking. Now, the question arises if a breakdown in the sense of the OIBDP for ranking can always be achieved, disregarding the particular configuration of the original data and the dimension.

4.2.1 Unbounded loss function

Theorem 1 *Let L satisfy Assumption 1. Then, the upper bound for the sample and population OIBDP for ranking is* $(p+1)/n$ *provided that* $1 < p < n-1$, p/n *for* $p = n-1$, $1/n$ *for* $p = 1$ *and not existent otherwise.*

Proof Let us first illustrate our outlier configuration for $p = 2$ and as usual, we only prove the population variant. Let wlog. be $\beta_1, \beta_2 > 0$ and let the original data points be linearly rankable according to β. Consider $X' = (\max_i(X_{i1}), \max_i(X_{i2}))$ and let $X^{(1)} = (X'_1 + 1, X'_2)$ and $X^{(2)} = (X'_1, X'_2 + 1)$. Set $Y' < \min_i(Y_i)$ and let $Y^{(1)} = Y^{(2)} < Y'$. This special configuration ensures that along each axis, any coefficient β with positive components will produce a misrankings w.r.t. the outliers and, in addition, that there is no consistency with the original data since it is guaranteed that the response for all outliers would be greater than the response for all original variables according to any β with positive components. Therefore, letting $Y^{(1)}, Y^{(2)} \to -\infty$, we produce an unlimited loss unless $\beta_1, \beta_2 < 0$. If any component, say, the first component of the original coefficient is negative, use $X^{(1)} = (X'_1 - 1, X'_2)$ and proceed along the same lines.

This strategy obviously is applicable to the general case $p > 2$, requiring at most $(p+1)$ outliers. In the special case $p = n-1$, we just need p instead of $p+1$ outliers by using the last remaining original data point as starting point for the construction of the p outliers. The special case $p = 1$ does not require a starting point.

As for the case $p \geq n$, consider for simplicity again $p = 2$ and let $\beta_1, \beta_2 > 0$ and let $X_{11} < X_{21}$, $X_{12} < X_{22}$ and $Y_1 < Y_2$. Regardless of the outlier configuration, one can only enforce the sign-inversal of one component. Even if one modifies (X_2, Y_2) by letting $Y_2 \to -\infty$, any coefficient with $\beta_1 > 0$ and $\beta_2 < 0$ resp. $\beta_1 < 0$ and $\beta_2 > 0$ produces a perfect ranking provided that the negative component dominates here. Enforcing a sign-reversal of both components stays impossible and carries over to higher dimensions $p > 2$. □

Remark 8 Note that although we cannot enforce multiple components to be sign-reverted by a single outlier, for particular algorithms this may not hold. When computing the BDPs, we considered all variables separately by our outlier schemes. This guarantees that our results provide conservative but valid upper bounds for the BDP which are sharp for the situations assumed in the respective theorems and lemmas. However, from an algorithmic point of view, the true BDP may be considerably

lower. This is true for example for gradient-based approaches which update multiple coefficient components at once by a joint gradient step so that a single outlier in a remote location may pull the estimated coefficient toward a broken coefficient. However, this is a property of the numerical procedure that intends to minimize the corresponding objective and no general property. One may alleviate this issue by for example single gradient steps like in Gradient Boosting [e.g., Bühlmann and Hothorn (2007)].

Theorem 1 has a very interesting consequence: In a high-dimensional setting where $p > n - 1$, it is impossible to find outlier configurations that guarantee a breakdown of the estimator in the sense of the OIBDP for ranking! Even if the whole data set would be replaced by outliers, it can only be enforced that n coefficients are sign-reversed. Also note that it is not unusual that the dimension enters the BDP which also appeared for example in the BDP of the Least Trimmed Squares (LTS) estimator introduced in Rousseeuw (1984), see also Rousseeuw and Van Driessen (2006) for its fast computation, given in Rousseeuw and Leroy (2005), where however an increasing dimension leads to a decreasing BDP. The sparse variant SLTS (Alfons et al., 2013) also has a dimension-independent BDP.

Going back to our ranking setting, the asymptotic case has to take the behavior of the predictor dimension into account. See Sect. 8 for further discussions.

Corollary 1 *Asymptotically, we have to distinguish between four cases.*

- **(i)** If p is fixed, then the asymptotic breakdown point is zero.
- **(ii)** If $p = p(n) = b_n n$ such that $b_n \to b \in [0, 1[$, the asymptotic breakdown point is b.
- **(iii)** If $p = p(n) = b_n n$ such that $b_n \to b \geq 1$, the asymptotic breakdown point does not exist, i.e., it is impossible to achieve a breakdown for ranking.

4.2.2 Bounded loss function

Theorem 2 *Let wlog. L be the indicator loss function used in the hard ranking loss and let $p \geq 2$. Then, the upper bound for the OIBDP for ranking is given by*

$$\frac{m^*}{n}, \quad m^* = 1 + pk^*, \quad k^* = \min \left\{ k \;\middle|\; \frac{k(k+1)}{2} > \frac{(n - pk - 1)(n - pk - 2)}{2} \right\}. \tag{8}$$

This quantity always exists for $p \leq n - 1$.

Proof Let us again illustrate our idea for $p = 2$. The problem is that when having a starting point X' as in the proof of Theorem 1, generating one outlier by altering one component may not suffice to ensure a breakdown if the original data points still

dominate. Moreover, we have to guarantee that all components of the coefficient are sign-reversed. We propose the following outlier algorithm:

$k = 1$;
while *No breakdown* $\wedge\ 1 + p(k+1) < n$ **do**
  for $j = 1, ..., p$ **do**
    Generate an outlier around X' on the j-th axis as in the proof of Thm. 1;
    if *Breakdown* **then**
      Stop
    end
  end
  $k = k + 1$;
end
$m = 1 + pk$

For illustration, let $p = 2$ and $k = 2$. Then we generate a further outlier on each axis (note that "axis" has to be understood in the sense $(X'_1, \mathbb{R})$ resp. $(\mathbb{R}, X'_2)$ here) by proceeding on the respective axis, i.e., if $X^{(1)} = (X'_1 + 1, X'_2)$, the next outlier is $X^{(3)} = (X'_1 + 2, X'_2)$ and $Y^{(3)} = Y^{(4)} < Y^{(1)} = Y^{(2)}$ for $X^{(2)} = (X'_1, X'_2 + 1)$ and $X^{(4)} = (X'_1, X'_2 + 2)$. Applying this strategy, we get $k(k+1)/2$ comparisons along each axis; more precisely, in the example, we have all pair-wise comparisons of the points (X', Y'),$(X^{(2)}, Y^{(2)})$, $(X^{(4)}, Y^{(4)}), \ldots, (X^{(2k)}, Y^{(2k)})$ along the axis corresponding to the first component of the coefficient vector, which are $k(k+1)/2$ in total; additionally, we have all pair-wise comparisons of the points (X', Y'),$(X^{(1)}, Y^{(1)})$, $(X^{(3)}, Y^{(3)}), \ldots, (X^{(2k-1)}, Y^{(2k-1)})$ along the axis corresponding to the second component of the coefficient vector, which are again $k(k+1)/2$. In total, keeping the original sign of the corresponding coefficient component leads to $k(k+1)/2$ misrankings. In contrast, we still have $(n - 2k - 1)$ original data points which, in the worst case, cause $(n - 2k - 1)(n - 2k - 2)/2$ misrankings provided that the coefficient has at least one component with the original sign.

Note that comparisons of original and contaminated data points are not informative. Let us elaborate this argument a bit further. By construction, the responses of the outliers are lower than the responses of the original data which makes their ranking prediction perfect if all components of the coefficient are sign-reversed. In this case, the loss suffered due to these $(n - 2k - 1)(2k + 1)$ comparisons is zero, so such a coefficient indeed leads to $k(k-1)/2$ misrankings. On the other hand, the original coefficient induces $(n - 2k - 1)(2k + 1)$ misrankings, but any other coefficient in between these two extreme cases potentially predicts the respective orderings perfectly, so we have to be conservative and assume this "least favorable case" (from the view of the attacker) that such a coefficient also achieves a loss of zero like the completely sign-reverted coefficient when comparing outliers and original data points. Therefore, a breakdown is guaranteed once k is large enough such that

$$\frac{k(k+1)}{2} > \frac{(n - 2k - 1)(n - 2k - 2)}{2}$$

which leads to stated formula 8 for $p = 2$.

In the general case $p > 2$, we consider at most p-chunks of k new outliers, i.e., $m = 1 + pk$, and by the same arguments, a breakdown occurs if

$$\frac{k(k+1)}{2} > \frac{(n - pk - 1)(n - pk - 2)}{2}.$$

Clearly, there exist cases where such a k^* does not exist. Here, we have to distinguish between two cases: **(i)** $p \geq n$; **(ii)** $p \leq n - 1$.

(i) This case is already discussed in Theorem 1 where we concluded that it is impossible to guarantee a breakdown in such high dimensions. This evidently also holds for the case of bounded loss functions.

(ii) A breakdown may be achieved before a p-chunk is complete. In the worst case, we can stop once $m = n - 1$ since then, using the last remaining point as starting point, we can generate at least one outlier along each axis. In general, provided that k^* exists, the true upper bound BDP therefore lies in the set $\left\{\frac{1+p(k^*-1)+1}{n}, ..., \frac{1+pk^*}{n}\right\}$. Note that there exist configurations in which $m^* = 1 + pk^*$ is indeed sharp which is true for example for $p = n - 1$ as already discussed. □

Example 4 To illustrate the case (ii) in the proof above, consider the case $p = 2$ and $n = 8$. Generating an outlying starting point and two outliers along each axis leads to $m = 5$ and $k = 2$, but we have only three comparisons of outliers along each axis and three comparisons of original instances. The loss for each coefficient β with $\beta_1, \beta_2 > 0$ is obviously greater than the loss for each coefficient with $\beta_1, \beta_2 < 0$, but there is no guarantee that such a sign-reversed coefficient would achieve a lower loss than a coefficient with only one sign-reverted component. However, adding one additional outlier according to our outlier scheme, disregarding on which of the two axes, leads to a breakdown since the number of comparisons between original data boils down to one, leading finally to $m^* = 6$ instead of $m^* = 7$.

Corollary 2 *The asymptotic upper bound for the OIBDP for ranking*

(i) is given by $p/(p + 1)$ for fixed p,
(ii) is given by 1 for $p = p(n) = b_n n$ with $b_n \to b \in]0, 1[$,
(iii) does not exist for $p = p(n) = b_n n$ with $b_n \to b \geq 1$.

Remark 9 We do not exclude that there may exist even more sophisticated outlier schemes than ours which leads to a faster breakdown. However, our outlier scheme guarantees a breakdown, provided than p resp. $p(n)$ is small enough, which would be very hard to show for outlier schemes than are not axis-based. An intuitive alternative that however does not work is given in the supplementary file.

Summarizing, we showed under which conditions and with which outlier scheme a breakdown for the hard instance ranking problem can be achieved. This again relates to out motivating example Example 1 with continuous pseudo-responses, e.g., the amount of damage. If the data would be suitably perturbed, the resulting ranking model would suggest the tax offices to investigate essentially those income tax statements which lead to a negative damage, i.e., a refund to the tax payer, which may be undesirable for the government.

4.3 Expected OIBDP

Evidently, there are always pathological configurations of the original data that even immediately cause a breakdown or that hinder a breakdown but being extremely artificial, see the supplementary file for an example. A comparable situation has already been investigated in Ruckdeschel and Horbenko (2012) who consider the expectation of the BDP w.r.t. the ideal distribution (which the original instances are assumed to follow), leading to a so-called expected BDP. Their motivation was to account for the fact that unfavorable configurations of the original data points only appear with low probabilities which helped them to get nonzero expected BDPs in the context of heavy-tailed distributions or when only partial equivariance is valid.

However, in our setting, we have to be very cautious how to define an **expected OIBDP**. Evidently, assuming iid. instances (X_i, Y_i) and computing the expectation w.r.t. the joint distribution would make no sense since iid. instances are all ranked equally in expectation. We indeed require a fixed design of the regressor matrix which, for every fixed n, assumes that observations $X_{n,i}, i = 1, \dots, i_n$, are given. Then, the responses are computed by $Y_{n,i} = X_{n,i}\beta + \epsilon_{n,i}$ for $\epsilon_{n,i} \sim F_\epsilon$ iid. for some centered distribution F_ϵ. Therefore, the points $(X_{n,i}, X_{n,i}\beta)$ are trivially linearly rankable, but the points $(X_{n,i}, Y_{n,i})$ do not necessarily be linearly rankable since this property depends on the realizations of the error terms. In the proofs, we always consider linear rankability w.r.t. the original coefficient β which can be interpreted as taking the expectation of the data w.r.t. F_ϵ. This motivates the following definition which mimicks Ruckdeschel and Horbenko (2012, Def. 3.2).

Definition 5 *(Expected OIBDP for ranking)* Let $Z_n(\epsilon)$ be the sample consisting of the data points $(X_{n,1}, Y_{n,1}(\epsilon_{n,1})), \dots, (X_{n,i_n}, Y_{n,i_n}(\epsilon_{n,i_n}))$.

(a) The **expected population order-inversal breakdown point for ranking** is defined by $\mathbb{E}_\epsilon[\check{\epsilon}(\beta, Z_n(\epsilon))]$. **b)** The **expected sample order-inversal breakdown point for ranking** is defined by $\mathbb{E}_\epsilon[\check{\epsilon}(\hat{\beta}, Z_n(\epsilon))]$.

Remark 10 One has to be very cautious when considering the sample OIBDP (or general sample BDPs). This fact has been respected by Zhao et al. (2018, Thm. 4) who indeed assume that the estimator does not yet break down on the original sample. As for ranking, our theoretical results on BDP bounds are founded on the expectation w.r.t. the error term, making the data linearly rankable w.r.t. β. Any tie or other inconsistency reduces the required amount of outliers to let the estimator break down. This can indeed be problematic if the sample BDP is considered and if,

maybe due to a large error variance, the estimated coefficient is insufficiently supported by the data. Let $p = 1$ and let the original coefficient have a small magnitude. Then, a large error variance may cause the data points to oscillate with growing regressor value, so just imposing one outlier may already change the sign of the estimated coefficient. We think that such issues are prone for a low signal-to-noise ratio (SNR) and that there may exist something like a "noise gap" between the population and sample BDP variants.

5 Hard binary and hard *d*-partite ranking problems

The goal of binary hard ranking problems is to find the correct ordering of all instances w.r.t. the probability to belong to class 1 for binary responses. In fact, one computes a real-valued scoring function so that the ordering of the scores is equivalent to an ordering of the respective probabilities. In d-partite ranking problems, one proceeds as in ordered logit regression by binning the scores. However, while an ordered classification model would be perfect if all instances get a score that is contained in the correct interval, hard d-partite ranking problems require that the ordering of all scores, and therefore also in the respective chunks, is correct.

As for the OIBDP computation, let us distinguish between two cases: **(i)** We have access to the real-valued pseudo-responses, so we are again in the usual continuous setting, making the results from Sect. 4 applicable; **(ii)** the more realistic case is that we indeed only observe the categorical responses and that we only can produce outliers with responses in the respective discrete set. The main difference to the continuous case is that the outlier configuration becomes far less flexible. Let the loss function operate on the score scale, i.e., we use $s_{b,\beta}(X_i)$ instead of $\hat{Y}_i$ where the latter would be ± 1 for binary ranking; otherwise, we were in a classification setting. All proofs can be found in the supplement.

Corollary 3 *If the loss function satisfies Assumption 1, the upper bound of the sample and population OIBDP for ranking is* $(p+1)/n$ *for* $p \leq n-2$, p/n *for* $p = n-1$, $1/n$ *for* $p = 1$ *and not existent otherwise.*

Let us now translate Lemma 2 and Theorem 2 to the case of hard binary ranking with the indicator loss function. As already elaborated in the proof of Corollary 3, we do not have access to the true underlying real-valued scores but only to the binary responses which severely restricts the possible outlier configurations. Then, the idea is essentially the same as in AUC maximizing approaches (for example done in Rakotomamonjy (2004) for SVM-type and in Clémençon and Vayatis (2008), Clémençon et al. (2013a) for tree-type approaches for ranking), i.e., the score for each instance of class 1 has to be higher than the score for each instance of class -1.

Lemma 3 *Let* $p = 1$. *For the hard binary and d-partite ranking problem with the loss function* $L(u) = I(u < 0)$, *the sample and population OIBDP for ranking is given by*

$$\frac{\check{m}}{n}, \quad \check{m} = 2\check{k}, \quad \check{k} = \min\left\{ k \,\middle|\, k\lfloor\frac{n}{2}\rfloor + k\Big(\lceil\frac{n}{2}\rceil - k\Big) > \Big(\lceil\frac{n}{2}\rceil - k\Big)\Big(\lfloor\frac{n}{2}\rfloor - k\Big) \right\} \tag{9}$$

and asymptotically, the BDP is given by $1 - \sqrt{0.5}$.

Now, we consider the general case $p > 2$.

Theorem 3 *Let L be the indicator loss function and let* $p \geq 2$. *Then, the upper bound for the OIBDP for ranking for hard bipartite and hard d-partite ranking problems is given by*

$$\frac{m^*}{n}, \quad m^* = 1 + 2pk^*, \quad k^* = \min\left\{ k \,\middle|\, k(k+1) > \frac{(n - 2pk - 1)(n - 2pk - 2)}{2} \right\}. \tag{10}$$

This quantity always exists for $p \leq n - 1$.

Corollary 4 *The asymptotic upper bound for the OIBDP for bipartite and d-partite ranking*

(i) Is given by $(2p^2 - \sqrt{2}p)/(2p^2 + 1)$ for fixed p,
(ii) Is given by 1 for $p = p(n)$ with $p(n)/n \to b \in]0, 1[$,
(iii) Does not exist for $p = p(n)$ with $p(n)/n \to b \geq 1$.

The ranking problems considered in this section have applications for example in fraud detection according to Example 1 if the response is binary (fraudulent/compliant), but also for example for rating (Clémençon et al., 2013b) or gene identification (Agarwal and Sengupta, 2009) where especially rating usually corresponds to more than two ordered classes, leading to a d-partite ranking problem. Note that although we described how the ranking model can be broken down, s_β assigns a real-valued score to each instance, requiring a discretization step. Therefore, the effect on the actual binary or d-partite predictions depends on the selected discretization method; however, it an order-inversal BDP has occurred; one can assume that at least a significant fraction of the instances is wrongly classified due to the inverted ranking of the instances w.r.t. their probabilities for the classes.

6 Localized ranking problems

Localized ranking problems follow two goals, i.e., identifying the top K instances and retrieve the ordering of the true or fitted top K instances correctly. As for robustness analysis, we have to consider the OIBDP for ranking instead of the angular BDP for classification from Zhao et al. (2018) since the former one is stricter, so letting the ranking break down directly guarantees a breakdown of the classification

due to the fixed number K of class 1 instances (for $K < n/2$; the other case will also be discussed below). The proofs can be found in the supplementary file.

Corollary 5 *If the loss function used for the ranking part satisfies Assumption 1, the upper bound of the sample and population OIBDP for localized ranking is* $(p+1)/n$ *for* $p \leq K-2$, p/n *for* $p = K-1$ *and not existent otherwise. If the classification loss function satisfies Assumption 1, the BDP is* $(p+1)/n$ *for* $p \leq K-1$, p/n *for* $p = K$ *and not existent otherwise. In either case, it is* $1/n$ *for* $p = 1$.

As for the case of bounded loss functions, wlog. the indicator loss functions, we have to distinguish between a couple of cases, i.e., if $K \leq n/2$ or $K > n/2$ and if the ranking part of the localized loss is based on $Best_K$ or on $\widehat{Best}_K$.

In this section, we require that the ranking of the true best K instances is predicted correctly (see the discussion below Eq. 3). The case of localizing on $\widehat{Best}_K$ can be found in the supplementary file.

Lemma 4 *Let* $p = 1$. *For the localized continuous ranking problem optimizing*

$$\frac{n-K}{n}\frac{2}{n}\sum_{i\in Best_K} I(\mathrm{rk}\,(s_\beta(X_i)) > K) + \frac{2}{n(n-1)}\sum_{i<j,i,j\in Best_K}\sum I((s_\beta(X_i)-s_\beta(X_j))(Y_i-Y_j)<0), \tag{11}$$

the sample and population OIBDP for ranking

(i) is given by

$$\frac{\check{m}}{n},\quad \check{m} = \min\left(K, \min\left\{m \;\middle|\; \frac{n-K}{n}\cdot\frac{2(K-m)}{n} + \frac{(K-m)(K-m-1)}{2n(n-1)} < \frac{n-K}{n}\cdot\frac{2m}{n} + \frac{1}{n(n-1)}\left[\frac{m(m-1)}{2} + m(K-m)\right]\right\}\right) \tag{12}$$

for $K \leq (n+m)/2$,

(ii) is given by

$$\frac{\check{m}}{n},\quad \check{m} = \min\left\{m \;\middle|\; \frac{n-K}{n}\cdot\frac{2n-K}{n} + \frac{(K-m)(K-m-1)}{2n(n-1)} < \frac{n-K}{n}\cdot\frac{2m}{n} + \frac{1}{n(n-1)}\left[\frac{m(m-1)}{2} + m(K-m)\right]\right\} \tag{13}$$

for $K \geq (n+m)/2$ *and* $K \leq n-m$ *provided that* $n-m \geq (n+m)/2$,

(iii) is given by Eq. 7 in Lemma 2 where n in the definition of $\check{m}$ is replaced by K for $K \geq n-m$.

Theorem 4 *Let $p \geq 2$. Then, the upper bound for the OIBDP for localized ranking with the optimization problem as in Eq. 11 is given by*

$$\frac{m^*}{n}, \quad m^* = \min(1 + pk^*, K), \quad k^* = \min \left\{ k \;\middle|\; \frac{n-K}{n} \frac{2\min(K-1-pk, n-K)}{n} + \frac{(K-1-pk)(K-2-pk)}{2n(n-1)} < \frac{n-K}{n} \frac{2(1+pk)}{n} + \frac{k(k+1)}{2n(n-1)} \right\} \tag{14}$$

for the case $K < n - m$. For $K > n - m$, we get the same k^* as in Eq. 8 in Theorem 2. This quantity always exists for $p \leq K - 1$.

See the supplementary file for the asymptotic counterparts of Lemma 4 and Theorem 4.

Our results can for example be applied to recommender systems [e.g., Chu et al. (2020), Yoganarasimhan (2020)] as their goal is indeed to identify the top products for the costumer and to present them in the correct order. Our results show that such models can indeed be perturbed, although model training mainly focuses on the few top instances. If the ranking model breaks down, it is expected that it promoted products that are very unsuitable for the respective costumer.

7 Other ranking problems

Weak ranking problems (Clémençon and Vayatis, 2007) are nothing but binary classification problems with the peculiarity that one has to predict exactly K class 1 instances. Since a binary classification loss function is used for weak ranking problems, the notion of the angular breakdown point of Zhao et al. (2018) (resp. Zhao et al. (2018, Def. 2+2') for kernel-based classification) is directly applicable, but the results of Zhao et al. (2018) are only valid if the loss function is a suitable surrogate of the 0/1-loss function since continuity is assumed there. As for the outlier scheme, note that the number K leads to an additional constraint in the proposed outlier set in the proof of Zhao et al.(2018, Thm. 2). Based on the mere classification loss, we can produce outliers that lead to a breakdown of the coefficient in terms of the OIBDP. The proofs are in the supplementary file.

Corollary 6 *If the (classification) loss function satisfies Assumption 1, the upper bound of the sample and population OIBDP for weak continuous ranking is p/n for $p \leq K - 1$ and not existent otherwise.*

Theorem 5 *For the weak continuous ranking problem with the 0/1-loss function, the OIBDP*

(a) is given by m/n for $m = \lfloor K/2 \rfloor + 1$ for $p = 1$,

(b) is bounded from above by K/n for $1 < p < K$,
(c) Does not exist for $p \geq K$.

As for Corollary 6, note the difference to Zhao et al.(2018, Thm. 1) where an angular BDP of $1/n$ is proven, resulting from the different definitions of the angular BDP and the OIBDP. For the special case $p = 1$, the definitions coincide, but for Theorem 5, the results would only coincide with similar results for the angular BDP if K equals the number of positives in the data set.

We abstain from detailing out possible results for localized binary and localized d-partite ranking problems as well as for weak binary ranking problems. The reason is that these problems are essentially ill-posed from the perspective of the OIBDP. The reason is that when localizing, the top K instances may all have the same label which makes them indistinguishable and therefore not rankable in any sense. We suggest to focus only on the classification part, inevitably requiring to measure the robustness in terms of the angular BDP for binary (Zhao et al., 2018) or of the angular BDP for d-partite localized ranking (Qian et al., 2019).

8 Discussion and simulations

8.1 The nonexistence issue

Example 5 Let $p > 1$ and assume that $X_{ij} = 0 \; \forall j \neq j_0$ and $\beta_j = I(j = j_0)$ for some $1 \leq j_0 \leq p$. Then, it suffices to use the worst-case outlier configuration from Fig. 1 only on the j_0-th axis. Although we cannot guarantee that our estimated coefficient maintains the zero components, but however, $\beta_j \hat{\beta}_j < 0$, hence $\hat{\beta} \in S_{\cap}^{-}$, is clearly satisfied.

The computed breakdown points depend on the dimension p and generally grow with p. Even worse, if p is at least as large as n resp. K, a breakdown can no longer be achieved. However, the tides turn once sparsity of the true underlying model is assumed as the example above showed.

Definition 6 The linear model $Y = X\beta$ is called sparse with true dimension q if $||\beta||_0 = q$. In this setting, denote the set of the q relevant variables by S^0.

If the outlier scheme exactly knows which q predictors are relevant (we may call the outliers **"oracle outliers"** here), the outlier scheme is only applied to the corresponding q axes.

Corollary 7 *Let n be fixed and let $q \leq n - 1$ resp. $q \leq K - 1$ be the true dimension of the linear model, i.e., $||\beta||_0 = q$. Then the order-inversal breakdown point for every ranking problem that we considered in this work exists.*

Corollary 8 *Let $q = q(n) = b_n n$ such that $b_n \in]0, 1[$ $\forall n$. Then the asymptotic order-inversal breakdown point for all nonlocalized ranking problems considered in this work exists. For localized ranking problems with $K = K(n)$ with $K(n)/n \to d \in]0, 1]$, we have to assume that $b_n \to b < d$.*

We are aware of the fact that very high-dimensional true models for which $q \geq n$ holds cannot break down in the sense of the OIBDP. In many situations, one can reduce this dimension to $q' < n$ by only considering the most relevant predictors (e.g., Meinshausen and Bühlmann (2010)), although there are situations in which more than n selected predictors are desired [e.g., Wang et al. (2011)]. We do not think of this issue as being a weakness of our OIBDP notion since the OIBDP is quite intuitive and since the global nature of ranking problems that take at least pairs of instances into account and no single instances defines a significantly different setting than for example regression for which higher dimensions generally reduce the BDP. The OIBDP can still be used to compare the robustness of competing algorithms by considering the $q < n$ case which identifies which algorithm is more robust.

Remark 11 One could ask why one cannot just multiply the responses with (-1) in order to achieve a breakdown which also holds for SVR-type ranking estimators below in Lemma 5. Honestly speaking, from an algorithmic perspective, we believe that one can indeed let the ranking estimator break down for any reasonable algorithm using this outlier scheme, making the OIBDP indeed existent for any true dimension q (and therefore, letting it also exists for nonsparse true models). However, from a theoretical perspective, there is no evidence that one cannot result in a nonbroken coefficient since the solution set, i.e., the set of all coefficients that optimize a ranking loss for the data set with the negated responses, does not only consist of broken coefficients but also of nonbroken ones. The argument is the same as in Theorem 1 that sign-inverting some but not all coefficient components may already lead to a perfect ranking prediction on the contaminated sample, so there is no guarantee that all components would be enforced to be sign-inverted.

8.2 Lower bounds for the OIBDP

Lower bounds for the OIBDP in the sense that one asks for example in the situation of Lemma 2 where we assumed that the original data points supply nonbroken coefficients most (i.e., that the data are linearly rankable) if there is any lower OIBDP value that holds with high probability on real data (where linear rankability may not hold) cannot be computed universally due to numerous reasons.

First, the original data contain some noise so even if they would follow some linear model with some true β, the observed response values would differ from the ideal response values so that linear inrankability can occur by chance. However, if one had a model $Y_i = s_{b,\beta}(X_i) + \epsilon_i$ for some stochastic error term ϵ_i, the probability that linear inrankability occurs does not only depend on the error distribution but also on the X_i and on β. For example for $p = 1$, if $\beta > 0$ is very large, the probability

that the errors make the data points linearly inrankable would be smaller than for some smaller $\beta > 0$ if the predictors are kept fixed. For a fixed β, the probability that linear inrankability occurs would also be lower if there are large distances between the predictors since the expected responses then would be better separated.

We also already mentioned in Remark 8 that the underlying numerical algorithm itself may affect the OIBDP. Due to these reasons, we think that if one had a concrete algorithm, a given data set and a good intuition of the error distribution and the true coefficient, one may would be able to compute lower bounds for the OIBDP, but evidently, there is no chance to provide universal results.

Example 6 Consider a very simple artificial situation with $p = 1$ where one has the predictors $X_i = i$, $i = 1, \ldots, 5$, the true coefficient $\beta = 1$ and $Y_i = X_i\beta + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 1)$ i.i.d.. It can clearly happen that $Y_1 \geq Y_2 \geq Y_3 \geq Y_4 \geq Y_5$. In this case, an immediate breakdown is suffered, so the OIBDP is zero on this particular data set. In contrast, if the error distribution is bounded, e.g., a truncated normal distribution on the interval $[-1.1, 1.1]$, it is impossible that this situation occurs. The worst case would be $Y_4 > Y_5 > Y_1 > Y_2 > Y_3$ (alternatively, $Y_3 > Y_4 > Y_5 > Y_1 > Y_2$). Here, one would not have an immediate breakdown as a positive coefficient would make 4 misrankings, while a negative coefficient would make 6 misrankings, but modifying Y_5 so that $Y_5 < Y_3$ would already lead to a breakdown. Here, the lower bound of zero cannot be attained, but only an OIBDP of 0.2 is attainable.

8.3 Practical implications

The results from this work indicate that bounded loss functions lead to more robust ranking problems than unbounded loss functions. This is not surprising and coincides with the well-known results from robust regression and robust classification where redescenders, i.e., loss functions whose gradient in absolute value redescends to zero so that the loss functions asymptotically grow until reaching a constant, are proposed [e.g., Huber and Ronchetti (2009)].

As for ranking losses, we always assumed that $\lim_{u\to\infty}(L(u)) = 0$. Therefore, bounded loss functions with $\lim_{u\to-\infty}(L(u)) = C_l < \infty$ as we assumed in several theorems are in fact redescenders. The problem is that redescenders are nonconvex which makes numerical optimization difficult. In fact, almost every existing ranking algorithm works with convex and therefore unbounded surrogate loss functions. Nevertheless, nonconvex optimization has already been addressed in for example robust regression, so developing a robust ranking algorithm is definitely possible, although, due to the global nature of ranking loss functions, the computational complexity can be assumed to be very high. Therefore, providing a robust ranking algorithm is beyond the scope of this work.

On the other hand, a standard robustification technique is trimming on which many successful machine learning algorithms are based, most prominently the LTS (Rousseeuw, 1984) or the SLTS (Alfons et al., 2013). These trimming techniques are based on the in-sample losses, i.e., one iteratively identifies the relative $(1 - \alpha)$-fraction of instances with the lowest in-sample loss, updates the model by

fitting it on these instances, checks again which instances provide the lowest loss and so forth. We want to point out why this trimming technique is not trivially applicable to ranking.

Looking at Fig. 3, there are three instances that contradict a positive ranking coefficient, colored in red. If one would apply trimming with a trimming rate of $\alpha = 0.2$, the first question that arises is which of the three red points should be discarded since the indicator loss would make them indistinguishable in terms of the loss, so one would have to pick two of them randomly. Usually, one considers a surrogate of the indicator loss which would clearly discard the points (6,3) and (10,6) because the loss when comparing these instances with their left neighbors leads to the values $(3-4)(6\beta - 5\beta) = -\beta$ resp. $(6-11)(10\beta - 9\beta) = -5\beta$ as input for the surrogate loss which is negative for all $\beta > 0$, so due to the monotonicity assumption, these pairs lead to the highest losses. Since the comparison of the points (5,4) and (9,11) with their left neighbors does not produce a loss, one would learn that indeed the points (6,3) and (10,6) are problematic from the perspective of ranking. If the trimming rate would be $\alpha = 0.3$, one would discard all three red points.

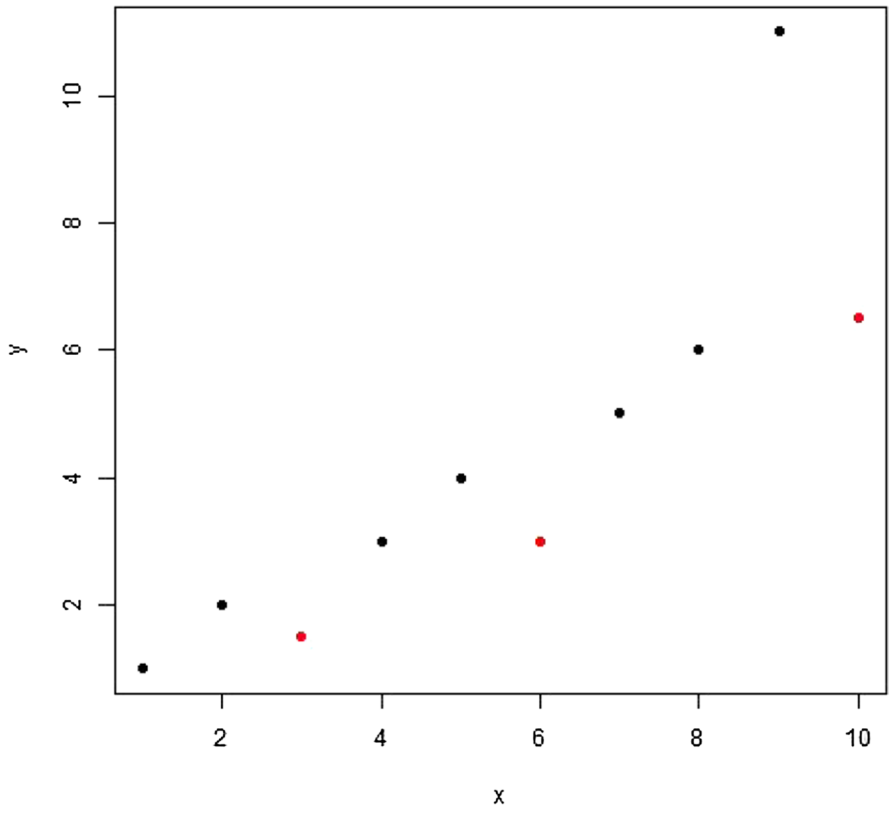

Fig. 3 Example where trimming would be difficult

Although this argumentation seems to be logical, it is in fact highly misleading. Regarding the points themselves, all of them except for (9,11) are likely to have been created by a linear model, so (9,11) would appear as a regression outlier. From the perspective of ranking, this outlier lets however the point (10,6) **appear as an outlier which can be interpreted as a swamping effect** [see, e.g., Rousseeuw and Hubert (2011)]. Therefore, in contrast to regression or classification problems where each instance can be treated individually and where the in-sample losses are instance-specific, the **globality of ranking prevents from applying trimming techniques in the usual way**.

One possible remedy, although not very popular in the ranking community, is to use a plug-in approach, i.e., one applies a regression algorithm and uses the regression predictions for the ranking prediction; in other words, the regression function serves as scoring function. Approaches in this direction have been proposed by Sculley (2010) who however combines a regression and a ranking loss, while Mohan et al. (2011) solely consider the squared loss. In this spirit, robust ranking may be achievable by robust regression, i.e., one could perform algorithms like SLTS and use its predictions for the ranking prediction. This will be an interesting topic for future work.

8.4 Simulations

For the hard continuous ranking problem, we generated B data sets for each of the different configurations of p, n and the signal-to-noise ratio SNR specified in Table 4. We set $B = 100$ if $p = 10$ and $n = 1000$ resp. if $p = 20$ and $n \geq 400$ due to the vast computational complexity, otherwise, $B = 1000$. In the linear model $Y_i = X_i\beta + \epsilon_i$, the X_i i.i.d. follow a $\mathcal{N}_p(1_p, I_p)$-distribution and $\beta_j \sim U([1, 2])$ i.i.d., while the errors ϵ_i are i.i.d. Gaussian so that the respective SNR is attained.

For the hard bipartite ranking problem, we generate the X_i as before, the β_j from uniform distributions according to Table 5a where larger coefficients essentially correspond to stronger signals, although we are not aware of a method how to targetedly attain some SNR in binary settings. We compute $\overline{X_i\beta} := X_i\beta - \text{mean}(X_i\beta)$ and

p	SNR = 0.2			SNR = 0.5			SNR = 1			SNR = 2			SNR = 5			SNR = ∞		
n	50	100	1000	50	100	1000	50	100	1000	50	100	1000	50	100	1000	50	100	1000
1	0.02	0.03	0.075	0.06	0.08	0.115	0.1	0.11	0.146	0.14	0.15	0.18	0.18	0.2	0.215	0.3	0.3	0.293
	0.105	0.103	0.099	0.141	0.138	0.134	0.169	0.167	0.162	0.198	0.195	0.19	0.229	0.226	0.222	0.3	0.3	0.293
	0.18	0.16	0.117	0.2	0.19	0.147	0.22	0.2	0.174	0.24	0.22	0.201	0.26	0.25	0.23	0.3	0.3	0.293
2	0.02	0.07	0.103	0.14	0.13	0.157	0.22	0.19	0.195	0.22	0.21	0.225	0.26	0.23	0.249	0.26	0.25	0.267
	0.246	0.201	0.131	0.271	0.224	0.179	0.288	0.239	0.213	0.302	0.251	0.241	0.314	0.262	0.265	0.326	0.273	0.286
	0.34	0.27	0.161	0.34	0.27	0.207	0.38	0.29	0.237	0.38	0.31	0.259	0.38	0.31	0.281	0.42	0.31	0.303
3	0.14	0.16	0.1	0.26	0.22	0.157	0.26	0.25	0.193	0.32	0.25	0.22	0.32	0.25	0.244	0.32	0.28	0.268
	0.329	0.249	0.138	0.358	0.263	0.19	0.362	0.275	0.227	0.373	0.286	0.257	0.382	0.296	0.281	0.389	0.305	0.304
	0.44	0.37	0.175	0.5	0.34	0.22	0.5	0.34	0.259	0.5	0.37	0.283	0.5	0.37	0.31	0.5	0.37	0.334
4	0.18	0.21	0.109	0.26	0.25	0.161	0.34	0.29	0.209	0.34	0.29	0.233	0.34	0.33	0.261	0.34	0.33	0.293
	0.387	0.3	0.153	0.405	0.32	0.21	0.416	0.336	0.249	0.425	0.35	0.28	0.429	0.363	0.307	0.436	0.374	0.332
	0.5	0.41	0.201	0.5	0.41	0.261	0.58	0.41	0.301	0.58	0.41	0.321	0.5	0.45	0.345	0.48	0.45	0.373
5	0.12	0.26	0.146	0.32	0.26	0.196	0.32	0.31	0.246	0.42	0.31	0.27	0.42	0.31	0.271	0.42	0.36	0.366
	0.444	0.334	0.18	0.464	0.35	0.234	0.477	0.363	0.274	0.495	0.373	0.304	0.506	0.385	0.33	0.513	0.431	0.396
	0.62	0.46	0.221	0.62	0.46	0.296	0.62	0.46	0.321	0.62	0.46	0.346	0.62	0.46	0.371	0.62	0.46	0.431
n	50	200	1000	50	200	1000	50	200	1000	50	200	1000	50	200	1000	50	200	1000
10	0.42	0.355	0.241	0.62	0.405	0.291	0.62	0.455	0.341	0.62	0.455	0.391	0.62	0.505	0.441	0.62	0.505	0.491
	0.656	0.45	0.297	0.685	0.477	0.351	0.728	0.505	0.398	0.777	0.529	0.439	0.808	0.547	0.471	0.818	0.565	0.5
	0.82	0.505	0.341	0.82	0.555	0.391	0.82	0.555	0.441	0.82	0.605	0.491	0.82	0.605	0.491	0.82	0.605	0.541
n	100	400	1000	100	400	1000	100	400	1000	100	400	1000	100	400	1000	100	400	1000
20	0.61	0.453	0.421	0.61	0.453	0.461	0.61	0.503	0.501	0.61	0.553	0.501	0.61	0.553	0.541	0.81	0.6025	0.581
	0.626	0.49	0.447	0.636	0.586	0.483	0.656	0.537	0.505	0.685	0.562	0.538	0.75	0.607	0.551	0.81	0.6515	0.648
	0.81	0.553	0.501	0.81	0.603	0.501	0.81	0.603	0.541	0.81	0.603	0.541	0.81	0.653	0.581	0.81	0.6525	0.681

Fig. 4 Empirical lower bounds (upper rows), averages (middle rows) and upper bounds for different p and n for the hard continuous ranking problem

p	$\beta_j \sim U([0,1])$			$\beta_j \sim U([1,2])$			$\beta_j \sim U([4,5])$		
n	50	100	1000	50	100	1000	50	100	1000
1	0.04	0.02	0.02	0.04	0.06	0.108	0.22	0.22	0.248
	0.094	0.082	0.07	0.187	0.18	0.171	0.279	0.272	0.264
	0.24	0.2	0.154	0.28	0.28	0.218	0.32	0.3	0.276
2	0.02	0.01	0.017	0.18	0.17	0.245	0.26	0.29	0.313
	0.222	0.192	0.176	0.317	0.299	0.287	0.353	0.341	0.329
	0.34	0.37	0.269	0.42	0.37	0.321	0.42	0.37	0.345
3	0.02	0.01	0.073	0.26	0.25	0.295	0.38	0.31	0.343
	0.347	0.274	0.228	0.402	0.359	0.331	0.435	0.388	0.363
	0.5	0.43	0.407	0.5	0.43	0.361	0.5	0.43	0.385
4	0.02	0.17	0.089	0.34	0.33	0.329	0.5	0.33	0.369
	0.48	0.366	0.275	0.508	0.428	0.371	0.514	0.451	0.396
	0.66	0.49	0.345	0.66	0.49	0.401	0.66	0.49	0.417
5	0.42	0.31	0.181	0.42	0.41	0.371	0.62	0.41	0.401
	0.608	0.455	0.308	0.62	0.506	0.395	0.621	0.514	0.417
	0.62	0.61	0.371	0.82	0.61	0.411	0.82	0.61	0.431
n	50	200	1000	50	200	1000	50	200	1000
10	0.41	0.505	0.341	0.41	0.505	0.421	0.41	0.505	0.461
	0.41	0.545	0.401	0.41	0.6	0.463	0.41	0.605	0.477
	0.41	0.605	0.461	0.41	0.605	0.501	0.41	0.705	0.501
n	100	400	1000	100	400	1000	100	400	1000
20	0.81	0.703	0.441	0.81	0.703	0.441	0.81	0.703	0.441
	0.81	0.703	0.499	0.81	0.703	0.514	0.81	0.703	0.51
	0.81	0.703	0.521	0.81	0.703	0.521	0.81	0.703	0.521

(a) Empirical lower bounds (upper rows), averages (middle rows) and upper bounds for different p and n for the hard bipartite ranking problem

p	min	mean	max
1	0.33	0.35	0.37
2	0.37	0.37	0.37
3	0.41	0.41	0.41
4	0.36	0.401	0.41
5	0.49	0.49	0.49
6	0.5	0.5	0.5
7	0.49	0.49	0.49
8	0.55	0.55	0.55

(b) Empirical lower bounds, averages, and upper bounds for different p for the hard continuous ranking problem on the `bodyfat` data set

ab	ac	ad	bc	bd	cd	abc	abd	acd	bac	bad	bcd	cab	cad	cbd	dab	dac	dbc	a	b	c	d
						0.45	0.37	0.27	0.21	0.09	0.07	0.43	0.51	0.45	0.39	0.31	0.47	0.34	0.43	0.52	0.37
0.06	0.32	0.35	0.16	0.18	0.47	0.5	0.42	0.33	0.26	0.16	0.07	0.46	0.53	0.46	0.43	0.35	0.51	0.38	0.48	0.55	0.42
						0.55	0.45	0.37	0.29	0.23	0.09	0.51	0.53	0.49	0.47	0.37	0.53	0.43	0.52	0.58	0.46

Fig. 5 Empirical lower bounds (upper rows), averages (middle rows) and upper bounds for different p and n for the hard continuous ranking problem on the `iris` data set without the `Species` column. a, b, c and d are acronyms for `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`, ab indicates the a is the response and b the regressor, abc that a is the response and b and c are the regressors, and a indicates that a is the response and all other variables are the regressors

generate $Y_i \sim B(1, \eta_i)$ for $\eta_i := \exp(\overline{X_i}\beta)/(1 + \exp(\overline{X_i}\beta))$. Here, we set $B = 1000$ for $p \leq 4$ and $B = 100$ otherwise.

We further investigate all sub data sets of the `iris` data set, disregarding the `Species` column, as specified in Table 5, for the hard continuous ranking problem. Lastly, we consider the `bodyfat` data set from the R-package `TH.data` (Hothorn, 2019) for the hard continuous ranking problem where the variable `DEXfat` is used as response variable. We first applied a best subset regression and used the optimal model suggested by this method for each p in Table 5b.

We want to highlight that we do not want to analyze one specific ranking algorithm from a vast variety of existing ones but try to replicate our universal theoretical findings. Therefore, we approximated the argmin of the loss in Eq. 1 by discretizing the space $[-1, 1]^p$ (note again that scaling the coefficients does not affect the ranking). For each β_j, we discretize the corresponding interval using 50, 20, 10, 10, 6, 4, 4, 4, 2 equidistant points, respectively, for $p = 2, 3, 4, 5, 6, 7, 8, 10, 20$. Of course, this leads to a discretization error; however, there is no ranking algorithm yet that is able to optimize the indicator losses; therefore, this strategy is reasonable in order to empirically check our results that correspond to the indicator function case.

Our results, shown in Tables 4, 5a, 5 and 5b are not surprising as they confirm that the OIBDP grows with the dimension p and with the SNR. For $p = 1$, the results for SNR $= \infty$ exactly reproduce our theoretical results, up to rounding issues as, e.g., for $n = 100$, a breakdown is achieved for 30 outliers since the theoretical number $100(1 - \sqrt{0.5})$ is not attainable. For $p \geq 2$, note that we assumed in the Proof of Theorem 2 the worst case that comparisons of original points and outliers do not contribute to the breakdown which however does not need to hold. Therefore, even

noiseless cases lead to smaller OIBDPs since there are such pairs that indeed contribute. Further note that the empirical OIBDP generally decreases with increasing n due to more crowded data points as discussed in Sect. 8.2. Cases with low n but high p only allow for little variation as we generate chunks of p outliers, leading to cases where the minimum empirical OIBDP equals the maximum one. Note that on the real data, the only variation comes from randomly picking original instances for contamination for $p \geq 2$.

9 Outlook: SVM-type approaches

A large class of ranking algorithms are of SVM-type which potentially operate in infinite-dimensional reproducing kernel Hilbert spaces (RKHS). At the first glance, such methods would be problematic for a ranking BDP since even finite-dimensional RKHSs like the ones induced by polynomial kernels would seemingly be prone to hurt the condition $p < n - 1$. The angular BDP from Zhao et al. (2018) has already been extended to kernel-based classification methods where they require the angle between the linear expansion [due to the representer theorem, e.g., Schölkopf et al. (2001)] of the true function resp. the solution computed on the contaminated data set, measured by the norm in the corresponding RKHS, to be nonpositive. Similarly, due to the component-wise nature of our OIBDP and the representer theorem, we can propose a reasonable definition of a BDP for kernel-based ranking estimators which is similar as the angular BDP from Zhao et al.(2018, Def. 2+2').

Definition 7 *(Order-inversal breakdown point for kernel-based ranking)* Assume that the true model has the form

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_{i,j} - \alpha_{i,j}^*)(K(X_i, x) - K(X_j, x)) + b \tag{15}$$

for some kernel K, an intercept term b with $|b| < \infty$ and coefficients $\alpha_{i,j}, \alpha_{i,j}^* \geq 0$. The **population order-inversal breakdown point for kernel-based ranking** is defined by

$$\check{\epsilon}(f, Z_n) := \min \left\{ \frac{m}{n} \,\middle|\, \hat{f}(Z_n^m) \in S_\cap^- \right\}, \quad S_\cap^- := \bigcap_{k: f_k \neq 0} \{\tilde{f}_k \mid \langle \tilde{f}_k, f_k \rangle_{\mathcal{H}} < 0\}$$

where $\mathcal{H}$ is the RKHS corresponding to K and where f_k is the k-th component of f.

Standard SVM classification solutions do not invoke the α^*-coefficients. For ranking algorithms that solely invoke α-coefficients, w.l.o.g. set $\alpha_{i,j}^* = 0$ for all i to consistently cover both cases with the definition of the OIBDP for kernel-based ranking. This general assumption covers the SVM-type ranking approaches like (Joachims, 2002; Cao et al., 2006; Brefeld and Scheffer, 2005; Pahikkala et al., 2007; Tian et al., 2011) where for example Herbrich et al. (1999a) let the class label enter as factor and Rakotomamonjy (2004) let the indices i and j run through

all positive resp. negative instances. Additional constraints for the coefficients like upper bounds as considered in Rakotomamonjy (2004) are not relevant in our BDP setting, while particular index sets are covered by setting the coefficients of the remaining summands to zero. As for the α^*-coefficients, many of the existing ranking algorithms are tailored to bipartite ranking and essentially approximate the conditional probability $\eta(x) := P(Y = 1|X = x)$ which relates ranking problems and regression algorithm like support vector regression (SVR).

There already exist sparse SVMs for ranking, see Tian et al. (2011), Pahikkala et al. (2010), Lai et al. (2013) and Laporte et al. (2014), but there is no guarantee that the selected number of features would be smaller than n. However, considering SVR techniques, due to the requirement that the coefficients α_i and α_i^* have to be nonnegative, we can conclude that for SVR-type algorithms, we need to enforce that sign $(\alpha_i - \alpha_i^*)$ switches for every $i = 1, \dots, n$ while preserving the sign of the differences of the features or kernelized features or vice versa. Studying the quantitative robustness of SVMs and SVRs in terms of the OIBDP which requires a thorough investigation of the corresponding dual problems for the α- (and α^*-)coefficients would exceed the scope of this work. However, we can state an enlightening result regarding standard SVR. Note that the proposed outlier scheme is the same as in Zhao et al. (2018).

Lemma 5 *If* $(\hat{\alpha}, \hat{\alpha}^*)$ *is the solution to the standard SVR problem* (*see. e.g. Friedman et al.* (2001))

$$\min_{\alpha,\alpha^*} \left(\epsilon \sum_i (\alpha_i^* + \alpha_i) - \sum_i y_i(\alpha_i^* - \alpha_i) + \frac{1}{2} \sum_i \sum_j (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\langle x_i, x_j \rangle \right)$$
$$0 \le \alpha_i, \alpha_i^* \le C, \quad \sum_i (\alpha_i^* - \alpha_i) = 0, \quad \alpha_i \alpha_i^* = 0 \ \forall i$$

for some cost parameter C *and the cutoff* ϵ *from the* ϵ*-insensitive loss function,* $(\hat{\alpha}^*, \hat{\alpha})$ *is the solution of the SVR problem on the data where the signs of all responses were switched.*

This statement is of major importance since it already proves the astounding fact that **there is no "blessing of dimensionality" for support vector regression regarding our OIBDP** since the same statement is true when $\langle X_i, X_j \rangle$ is replaced by $K(X_i, X_j)$, so even infinite-dimensional feature spaces do not prevent the OIBDP from existing. This is no contradiction to Remark 11 since Lemma 5 is tailored to the special case of SVR, so the statement does not transfer to other machine learning algorithms. We will not extensively study all existing SVM-type ranking algorithms, but we state the following for one of the most important and pioneering ranking algorithms.

Corollary 9 *The OIBDP of the ranking SVM algorithm from Herbrich et al.* (1999*a*), *Herbrich et al.* (1999*b*) *always exists.*

As for a general statement of the OIBDP for kernel-based ranking, we refer to the results from Zhao et al. (2018, Thm. 3+Prop. 3+Prop. 4) who proved upper bounds

for their angular BDP for kernel-based classification if unbounded loss functions and unbounded kernels are considered. Although their angular BDP is not identical to our OIBDP, they essentially sign-revert every summand in the corresponding representer theorem expansion by keeping the predictor values but by switching the sign of the respective responses. Interestingly, they derive very similar results as we did for the linear ranking setting for unbounded loss functions, i.e., the upper bound for the BDP is given by $\tilde{p}/n$ if $\tilde{p}$ is the dimension of the RKHS induced by the kernel, so the same problems concerning BDPs greater than 0.5 or even nonexistent BDPs occur here. As for unbounded RKHS's, the idea of Zhao et al. (2018) is to consider the effective dimension, i.e., the dimension of the finite-dimensional subspace of the RKHS in which the true scoring function f can be represented. The resulting upper bound is then again given by this number divided by n.

We postulate that the OIBDP for the SVM-type ranking algorithms always exists and takes a value lower than 1. This assumption is motivated by the results in Zhao et al. (2018, Ch. 4) and by Lemma 5. However, due to the huge variety of SVM-type ranking algorithms, we leave rigorous results about their OIBDPs, both regarding upper and possible lower bounds and for bounded resp. unbounded kernels, open for future research.

10 Conclusion

We introduced the order-inversal breakdown point for ranking and argued why neither the classical regression breakdown point nor the angular breakdown point for classification is appropriate for this setting. We then systematically studied the breakdown points for different types of ranking problems that we carefully distinguished. Our contribution includes least favorable outlier configurations and corresponding characterizations of the OIBDP as well as sharp asymptotic upper bounds, respecting all types of ranking problems that are appropriate for this setting combined with the extreme cases of unbounded loss functions and noncontinuous indicator loss functions. Our results are illustrated by empirical evaluations on simulated and real data.

One could argue that our BDPs may not be reasonable since cases with asymptotic BDPs of 1 or even cases where the BDP does not even exist arise. However, these problems are directly related to the sparsity of the underlying true model. Since a sparsity assumption is always encouraged in high-dimensional settings, relatively mild conditions on the growing behavior of the predictor dimension allow for an OIBDP smaller than 1.

Our results imply that robust ranking can be achieved by optimizing (nonconvex) redescending surrogate losses, but we leave the derivation of a concrete algorithm of this type as well as studying the plug-in approach based on robust regression open for future research. We also shortly discussed an extension of our OIBDP for ranking for the case of SVM-type scoring functions and proved the existence of this BDP, even for infinite-dimensional feature spaces.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-022-00847-1.

References

Agarwal, S. (2010). Learning to rank on graphs. *Machine Learning, 81*(3), 333–357.

Agarwal, S., Sengupta, S. (2009). Ranking genes by relevance to a disease. *Proceedings of the 8th annual international conference on computational systems bioinformatics,* 37–46.

Alfons, A., Croux, C., Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics, 7*(1), 226–248.

Alqallaf, F., Van Aelst, S., Yohai, V. J., et al. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics, 37*(1), 311–331.

Averbukh, V., Smolyanov, O. (1967). The theory of differentiation in linear topological spaces. *Russian Mathematical Surveys, 22*(6), 201–258.

Becker, C., Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association, 94*(447), 947–955.

Brefeld, U., Scheffer, T. (2005). AUC maximizing support vector learning. *Proceedings of the ICML 2005 workshop on ROC analysis in machine learning*, 92–99.

Bühlmann, P., Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22*(4), 477–505.

Bühlmann, P., Van De Geer, S. (2011). *Statistics for high-dimensional data*: *Methods, theory and applications*. Berlin, Heidelberg: Springer Science & Business Media.

Cao, Y., Xu, J., Liu, T.Y., Li, H., Huang, Y., Hon, H. W. (2006). Adapting ranking SVM to document retrieval. *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, 186–193. ACM.

Chu, L. Y., Nazerzadeh, H., Zhang, H. (2020). Position ranking and auctions for online marketplaces. *Management Science, 66*(8), 3617–3634.

Clémençon, S., Achab, M. (2017). Ranking data with continuous labels through oriented recursive partitions. *Advances in neural information processing systems,* 4603–4611.

Clémençon, S., Vayatis, N. (2007). Ranking the best instances. *Journal of Machine Learning Research, 8*(Dec), 2671–2699.

Clémençon, S., Vayatis, N. (2008). Tree-structured ranking rules and approximation of the optimal ROC curve. *Proceedings of the 2008 conference on algorithmic learning theory. Lecture Notes in Artificial Intelligence,* Vol. 5254, 22–37.

Clémençon, S., Vayatis, N. (2010). Overlaying classifiers: a practical approach to optimal scoring. *Constructive Approximation, 32*(3), 619–648.

Clémençon, S., Lugosi, G., Vayatis, N. (2008). Ranking and empirical minimization of U-statistics. *The Annals of Statistics, 36*(2), 844–874.

Clémençon, S., Depecker, M., Vayatis, N. (2013a). Ranking forests. *Journal of Machine Learning Research, 14*(Jan), 39–73.

Clémençon, S., Depecker, M., Vayatis, N. (2013b). An empirical comparison of learning algorithms for nonparametric scoring: the TreeRank algorithm and other methods. *Pattern Analysis and Applications, 16*(4), 475–496.

Clémençon, S., Robbiano, S., Vayatis, N. (2013c). Ranking data with ordinal labels: Optimality and pairwise aggregation. *Machine Learning, 91*(1), 67–104.

Davies, P. L. (1993). Aspects of robust linear regression. *The Annals of Statistics, 21*(4), 1843–1899.

Davies, P. L., Gather, U. (2005). Breakdown and groups. *The Annals of Statistics, 33*(3), 977–1035.

Donoho, D. L. (2006). High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete & Computational Geometry, 35*(4), 617–652.

Donoho, D. L., Huber, P. J. (1983). The notion of breakdown point. *A Festschrift for Erich L. Lehmann,* 157–184.

Donoho, D. L., Stodden, V. (2006). Breakdown point of model selection when the number of variables exceeds the number of observations. *The 2006 IEEE international joint conference on neural network proceedings,* 1916–1921. IEEE.

Freund, Y., Iyer, R., Schapire, R. E., et al. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research, 4*(Nov), 933–969.

Friedman, J., Hastie, T., Tibshirani, R. (2001). *The elements of statistical learning*. Springer Series in Statistics, Vol. 1. New York, NY: Springer New York.

Fürnkranz, J., Hüllermeier, E. (2011). *Preference learning,* Vol. 19. 01 ISBN 978-3-642-14124-9. https://doi.org/10.1007/978-3-642-14125-6.

Fürnkranz, J., Hüllermeier, E., Vanderlooy, S. (2009). Binary decomposition methods for multipartite ranking. *Joint European conference on machine learning and knowledge discovery in databases*, 359–374. Berlin, Heidelberg: Springer.

Gather, U., Hilker, T. (1997). A note on Tyler's modification of the mad for the stahel-donoho estimator. *Annals of Statistics, 25*(5), 2024–2026.

Genton, M. G. (1998). Spatial breakdown point of variogram estimators. *Mathematical Geology, 30*(7), 853–871.

Genton, M. G. (2003). Breakdown-point for spatially and temporally correlated observations. *Developments in robust statistics*, 148–159. Heidelberg: Springer.

Genton, M. G., & Lucas, A. (2003). Comprehensive definitions of breakdown points for independent and dependent observations. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *65*(1), 81–94.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics, 42*(6), 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association, 69*(346), 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., et al. (1986). *Robust statistics*: *The approach based on influence functions*. New York: Wiley-Interscience.

He, X. (2005). Discussion of "breakdown and groups" by P.L. Davies and U. Gather. arXiv: math/0508501.

Hennig, C. (2008). Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis, 99*(6), 1154–1176.

Herbrich, R., Graepel, T., Obermayer, K. (1999a). Support vector learning for ordinal regression. *9th international conference on artificial neural networks*: *ICANN '99*, 97–102. IET.

Herbrich, R., Graepel, T., Obermayer, K. (1999b). *Regression models for ordinal data*: *A machine learning approach*. Citeseer.

Hodges, J. L., Jr. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* Vol. 1, 163–186.

Hothorn, T. (2019). *TH.data*: *TH's data archive*, URL https://CRAN.R-project.org/package=TH.data. R package version 1.0-10.

Huber, P. J., Ronchetti, E. (2009). *Robust statistics*. New Jersey: John Wiley & Sons.

Hubert, M. (1997). The breakdown value of the L_1 estimator in contingency tables. *Statistics & Probability Letters, 33*(4), 419–425.

Hubert, M., Rousseeuw, P. J., Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science, 23*(1), 92–119.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining,* 133–142. ACM.

Kanamori, T., Takenouchi, T., Eguchi, S., et al. (2004). The most robust loss function for boosting. *Neural information processing*, 496–501. Berlin, Heidelberg: Springer.

Kayala, M. A., Azencott, C.-A., Chen, J. H., et al. (2011). Learning to predict chemical reactions. *Journal of Chemical Information and Modeling, 51*(9), 2209–2222.

Lai, H., Pan, Y., Liu, C., et al. (2013). Sparse learning-to-rank via an efficient primal-dual algorithm. *IEEE Transactions on Computers, 62*(6), 1221–1233.

Laporte, L., Flamary, R., Canu, S., et al. (2014). Nonconvex regularizations for feature selection in ranking with sparse SVM. *IEEE Transactions on Neural Networks and Learning Systems, 25*(6), 1118–1130.

Maronna, R. A., Martin, R. D., Yohai, V. J., et al. (2019). *Robust statistics*: *theory and methods* (*with R*). Chichester, England: John Wiley & Sons.

Meinshausen, N., Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *72*(4), 417–473.

Mohan, A., Chen, Z., Weinberger, K. (2011). Web-search ranking with initialized gradient boosted regression trees. *Proceedings of the learning to rank challenge,* 77–89. PMLR.

Morrison, J. L., Breitling, R., Higham, D. J., et al. (2005). Generank: Using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics, 6*(1), 1–14.

Page, L., Brin, S., Motwani, R., et al. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report Nr. 1999-66, Stanford InfoLab, November URL http://ilpubs.stanford.edu:8090/422/. Previous number = SIDL-WP-1999-0120.

Pahikkala, T., Tsivtsivadze, E., Airola, A. et al. (2007). Learning to rank with pairwise regularized least-squares. *SIGIR 2007 workshop on learning to rank for information retrieval,* Vol. 80, 27–33.

Pahikkala, T., Airola, A., Naula, P. et al. (2010). Greedy RankRLS: A linear time algorithm for learning sparse ranking models. *SIGIR 2010 workshop on feature generation and selection for information retrieval*, 11–18. ACM.
Pickett, K. S. (2006). *Audit planning*: *A risk-based approach*. New Jersey: John Wiley & Sons.
Qian, C., Tran-Dinh, Q., Fu, S., et al. (2019). Robust multicategory support matrix machines. *Mathematical Programming, 176*(1–2), 429–463.
Rakotomamonjy, A. (2004). Optimizing area under Roc curve with SVMs. *Proceedings of the ECAI-2004 workshop on ROC analysis in AI,* 71–80.
Rieder, H. (1994). *Robust Asymptotic Statistics*, Vol. 1. New York: Springer Verlag.
Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871–880.
Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications, 8*(37), 283–297.
Rousseeuw, P. J., Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*, *1*(1), 73–79.
Rousseeuw, P. J., Leroy, A. M. (2005). *Robust regression and outlier detection*, Vol. 589. Hoboken, New Jersey: John Wiley & Sons.
Rousseeuw, P. J., Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery, 12*(1), 29–45.
Ruckdeschel, P., Horbenko, N. (2012). Yet another breakdown point notion: EFSBP. *Metrika, 75*(8), 1025–1047.
Rudin, C. (2009). The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research, 10*(Oct), 2233–2271.
Sakata, S., White, H. (1995). An alternative definition of finite-sample breakdown point with applications to regression model estimators. *Journal of the American Statistical Association, 90*(431), 1099–1106.
Sakata, S., White, H. (1998). High breakdown point conditional dispersion estimation with application to S & P 500 daily returns volatility. *Econometrica*, 529–567.
Schölkopf, B., Herbrich, R., Smola, A. (2001). A generalized representer theorem. *Computational Learning Theory*, 416–426. Berlin, Heidelberg: Springer.
Sculley, D. (2010). Combined regression and ranking. *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining,* 979–988.
Stromberg, A. J., Ruppert, D. (1992). Breakdown in nonlinear regression. *Journal of the American Statistical Association, 87*(420), 991–997.
Tian, Y., Shi, Y., Chen, X., et al. (2011). AUC maximizing support vector machines with feature selection. *Procedia Computer Science, 4*, 1691–1698.
Torgo, L., Ribeiro, R. (2007). Utility-based regression. *European conference on principles of data mining and knowledge discovery,* 597–604. Berlin, Heidelberg: Springer.
Von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics, 18*(3), 309–348.
Wang, S., Nan, B., Rosset, S., et al. (2011). Random lasso. *The Annals of Applied Statistics, 5*(1), 468.
Werner, D. (2006). *Funktionalanalysis*. Berlin, Heidelberg: Springer.
Werner, T. (2021a). A review on instance ranking problems in statistical learning. *Machine Learning, 111*(2), 415–463.
Werner, T. (2021b). Trimming stability selection increases variable selection robustness. arXiv:2111.11818.
Werner, T. (2022). Elicitability of instance and object ranking. *Decision Analysis, 19*(2), 123–140.
Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science, 66*(3), 1045–1070.
Zhao, J., Yu, G., Liu, Y. (2018). Assessing robustness of classification using angular breakdown point. *Annals of Statistics, 46*(6B), 3362.