Robust Unsupervised Learning via L-Statistic Minimization

Andreas Maurer [1] **Daniela A. Parletta** [1 2] **Andrea Paudice** [1 3] **Massimiliano Pontil** [1 4]

Abstract

Designing learning algorithms that are resistant to perturbations of the underlying data distribution is a problem of wide practical and theoretical importance. We present a general approach to this problem focusing on unsupervised learning. The key assumption is that the perturbing distribution is characterized by larger losses relative to a given class of admissible models. This is exploited by a general descent algorithm which minimizes an L-statistic criterion over the model class, weighting small losses more. Our analysis characterizes the robustness of the method in terms of bounds on the reconstruction error relative to the underlying unperturbed distribution. As a byproduct, we prove uniform convergence bounds with respect to the proposed criterion for several popular models in unsupervised learning, a result which may be of independent interest. Numerical experiments with KMEANS clustering and principal subspace analysis demonstrate the effectiveness of our approach.

1. Introduction

Making learning methods robust is a fundamental problem in machine learning and statistics. In this work we proposes an approach to unsupervised learning which is resistant to unstructured contaminations of the underlying data distribution. As noted by Hampel (Hampel, 2001), “outliers” are an ill-defined concept, and an approach to robust learning, which relies on rules for the rejection of outliers (see Ord, 1996, and references therein) prior to processing may be problematic, since the hypothesis class of the learning process itself may determine which data is to be regarded as structured or unstructured. Instead of the elimination of outliers – quoting Hampel “data that don’t fit the pattern set by the majority of the data” – in this paper we suggest to restrict attention to “a *sufficient portion of the data in good agreement with one of the hypothesized models*”.

To implement the above idea, we propose using L-estimators (Serfling, 1980), which are formed by a weighted average of the order statistics. That is, given a candidate model, we first rank its losses on the empirical data and than take a weighted average which emphasizes small losses more. An important example of this construction is the average of a fraction of the smallest losses. However, our observations apply to general classes of weight functions, which are only restricted to be non-increasing and in some cases Lipschitz continuous.

We highlight that although L-statistics have a long tradition, a key novelty of this paper is to use them as objective functions based on which to search for a robust model. This approach is general in nature and can be applied to robustify any learning method, supervised or unsupervised, based on empirical risk minimization. In this paper we focus on unsupervised learning, and our analysis includes KMEANS clustering, principal subspace analysis and sparse coding, among others.

This paper makes the following contributions:

- A theoretical analysis of the robustness of the proposed method (Theorem 1). Under the assumption that the data-distribution is a mixture of an unperturbed distribution adapted to our model class and a perturbing distribution, we identify conditions under which we can bound the reconstruction error, when the minimizer of the proposed objective trained from the perturbed distribution is tested on the unperturbed distribution.
- An analysis of generalization (Theorems 4–6). We give dimension-free uniform bounds in terms of Rademacher averages as well as a dimension- and variance-dependent uniform bounds in terms of covering numbers which can outperform the dimension-free bounds under favorable conditions.
- A meta-algorithm operating on the empirical objective which can be used whenever there is a descent algorithm for the underlying loss function (Theorem 9).

The paper is organized as follows. In Section 2 we give a brief overview of unsupervised (representation) learning.

[1] Istituto Italiano di Tecnologia, Genoa, Italy [2] University of Genoa, Genoa, Italy [3] University of Milan, Milan, Italy [4] University College London, London, UK. Correspondence to: Andrea Paudice <and.paudice@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

In Sections 3 to 5 we present and analyze our method. In Section 6 we discuss an algorithm optimizing the proposed objective and in Section 7 we present numerical experiments with this algorithm for KMEANS clustering and principal subspace analysis, which indicate that the proposed method is promising. Proofs can be found in the supplementary material.

Previous Work Some elements of our approach have a long tradition. For fixed models the proposed empirical objectives are called L-statistics or L-estimators. They have been used in robust statistics since the middle of the last century (Lloyd, 1952) and their asymptotic properties have been studied by many authors (see Serfling, 1980, and references therein). Although influence functions play a certain role, our approach is somewhat different from the traditions of robust statistics. Similar techniques to ours have been experimentally explored in the context of classification (Han et al., 2018) or latent variable selection (Kumar et al., 2010). (Cuesta-Albertos et al., 1997) give a special case of our method applied to k-means with hard threshold. The method is analyzed with the lenses of different robustness properties in (Garcia-Escudero & Gordaliza, 1999). Finite sample bounds, uniform bounds, the minimization of L-statistics over model classes and the so called risk based-objectives however are more recent developments (Maurer & Pontil, 2018; 2019; Lee et al., 2020), and we are not aware of any other general bounds on the reconstruction error of models trained from perturbed data. A very different line of work for robust statistics are model-independent methods available in high dimensions (Elmore et al., 2006; Fraiman et al., 2019). Although elegant and very general, these depth-related pre-processing methods may perform sub-optimally in practice, as our numerical experiments indicate. Similar data-generating assumption are adopted in Robust estimation, a related line of work where the goal is to identify the parameters of a target distribution up to a small error. In this context, strong parametric assumptions are made on the target distributions and the learning problems involves typically simpler model classes such as, mean and covariance estmations. In constrast, in this paper we focus allows for non-parametric distributions and more complex model classes as singletons, subspaces and linear operators. Please refer to (Diakonikolas & Kane, 2020) for an up-to-date survey on the results and the techniques employed to derive efficient algorithms for robust estimation. Finally, we note that previous work on PAC learning (e.g. Angluin & Laird, 1987) has addressed the problem of learning a good classifier with respect to a target, when the data comes from a perturbed distribution affected by unstructured noise. Similarly to us, they consider that the target distribution is well adapted to the model class.

2. Unsupervised Learning

Let $\mathcal{S}$ be a class of subsets of $\mathbb{R}^d$, which we call the model class. For $S \in \mathcal{S}$ define the distortion function $d_S : \mathbb{R}^d \to [0, \infty)$ by[1]

$$d_S(x) = \min_{y \in S} \|x - y\|^2 \text{ for } x \in \mathbb{R}^d. \tag{1}$$

We assume that the members of $\mathcal{S}$ are either compact sets or subspaces, so the minimum in (1) is always attained. For instance $\mathcal{S}$ could be the class of singletons, a class of subsets of cardinality k, the class of subspaces of dimension k, or a class of compact convex polytopes with k vertices[2].

We write $\mathcal{P}(\mathcal{X})$ for the set of Borel probability measures on a locally compact Hausdorff space $\mathcal{X}$. If $\mu \in \mathcal{P}(\mathbb{R}^d)$, define the probability measure $\mu_S \in \mathcal{P}([0, \infty))$ as the push-forward of μ under d_S, that is, $\mu_S(A) = \mu(\{x : d_S(x) \in A\})$ for $A \subseteq [0, \infty)$. Now consider the functional $\Phi : \mathcal{P}([0, \infty)) \to [0, \infty)$ defined by

$$\Phi(\rho) = \int_0^\infty r d\rho(r), \quad \rho \in \mathcal{P}. \tag{2}$$

Then $\Phi(\mu_S) = \mathbb{E}_{X \sim \mu}[d_S(X)]$ is the expected *reconstruction error*, incurred when coding points by the nearest neighbors in S. The measures $\mu_S \in \mathcal{P}([0, \infty))$ and the functional Φ allow the compact and general description of several problems of unsupervised learning as

$$\min_{S \in \mathcal{S}} \Phi(\mu_S) = \min_{S \in \mathcal{S}} \mathbb{E}_{X \sim \mu}[d_S(X)]. \tag{3}$$

Denote with $S^* = S^*(\mu)$ a global minimizer of (3). Returning to the above examples, if $\mathcal{S}$ is the class of singleton sets, then $S^*(\mu)$ is the mean of μ. If it is the class of subsets of cardinality k, then $S^*(\mu)$ is the optimal set of centers for KMEANS clustering. If $\mathcal{S}$ is the class of k-dimensional subspaces, then $S^*(\mu)$ is the principal k-dimensional subspace.

An important drawback of the above formulation is that the functional Φ is very sensitive to perturbing masses at large distortions R. In the tradition of robust statistics (see e.g. Hampel, 1974; Serfling, 1980) this can be expressed in terms of the *influence function*, measuring the effect of an infinitesimal point mass perturbation of the data. Let δ_R be the unit mass at $R > 0$, then the influence function

$$\begin{aligned} \text{IF}(R; \rho, \Phi) &:= \frac{d}{dt} \Phi\big((1-t)\rho + t\delta_R\big)\Big|_{t=0} \\ &= R - \Phi(\rho), \end{aligned} \tag{4}$$

[1] In most parts our analysis applies also to other distortion measures, for example omitting the square in (1). The chosen form is important for generalization bounds, when we want to bound the complexity of the class $\{x \mapsto d_S(x) : S \in \mathcal{S}\}$ for specific cases.

[2] In these cases the set S is the image of a linear operator on a prescribed set of code vectors, see (Maurer & Pontil, 2010). Our setting is more general, e.g. it includes non-linear manifolds.

can be arbitrarily large, indicating that even a single datapoint could already corrupt $S^*(\mu)$. To overcome this problem, in the next section we introduce a class of robust functional based on L-statistics.

3. Proposed Method

Our goal is to minimize the reconstruction error on unperturbed test data, from perturbed training data. Specifically, we assume that the data we observe comes from a perturbed distribution μ that is the mixture of an unperturbed distribution μ^*, which is locally concentrated on the minimizer $S^* = S^*(\mu^*)$, and a perturbing distribution ν which is unstructured in the sense that it does not concentrate on any of our models[3]. Figure 1 depicts such a situation, when $\mathcal{S}$ is the set of singletons and $d = 1$.

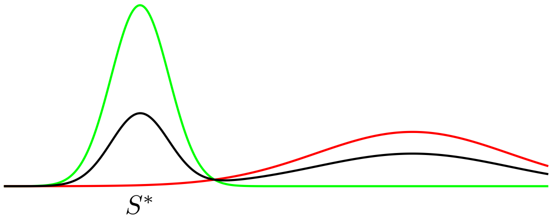

Figure 1. Densities of the unperturbed distribution μ^* (light green) with high local concentration on the optimal model S^*, the perturbing distribution ν (light red) without significant concentration, and the observable mixture $\mu = (1-\lambda)\mu^* + \lambda\nu^*$ (black) at $\lambda = 0.6$.

We wish to train from the available, perturbed data a model $\hat{S} \in \mathcal{S}$, which nearly minimizes the reconstruction error on the unperturbed distribution μ^*. To this end we exploit the assumption that the unperturbed distribution μ^* is much more strongly concentrated at S^* than the mixture $\mu = (1-\lambda)\mu^* + \lambda\nu$ is at models S *away* from S^* in terms of reconstruction error.

The key observation is that if the mixture parameter λ is not too large, the concentration of μ^* causes the cumulative distribution function of the losses for the optimal model $F_{\mu_{S^*}} : r \mapsto \mu_{S^*}[0,r]$ to increase rapidly for small values of r, until it reaches the value $\zeta = F_{\mu_{S^*}}(r^*)$, where r^* is a critical distortion radius depending on S^*. Thus, when searching for a model, we can consider as irrelevant the remaining mass $1-\zeta = \mu_{S^*}(r^*, \infty)$, which can be attributed to ν and may arise from outliers or other contaminating effects. To achieve this, we modify the functional (2) so as to consider only the relevant portion of data, replacing $\Phi(\mu_S)$ by

$$\zeta^{-1} \int_0^{F_{\mu_S}^{-1}(\zeta)} r d\mu_S(r). \tag{5}$$

Intuitively, the minimization of (5) forces the search towards models with the smallest *truncated* expected loss. Among such models there is also S^*, whose losses have the strongest concentration around a *small* value and then leading to a very small value r^* for $F_{\mu_{S^*}}^{-1}(\zeta)$.

More generally, since the choice of the hard quantile-thresholding at ζ is in many ways an ad hoc decision, we might want a more gentle transition of the boundary between relevant and irrelevant data. Let $W : [0,1] \to [0,\infty)$ be a bounded weight function and define, for every $\rho \in \mathcal{P}[0,\infty)$,

$$\Phi_W(\rho) = \int_0^\infty rW(F_\rho(r))\, d\rho(r).$$

We require W to be non-increasing and zero on $[\zeta, 1]$ for some critical mass $\zeta < 1$. The parameter ζ must be chosen on the basis of an estimate of the amount λ of perturbing data. Note that if W is identically 1 then $\Phi_1 = \Phi$ in (2), while if $W = \zeta^{-1}1_{[0,\zeta]}$ then Φ_W is the hard thresholding functional in (5).

We now propose to “robustify” unsupervised learning by replacing the original problem (3) by

$$\min_{S \in \mathcal{S}} \Phi_W(\mu_S), \tag{6}$$

and denote a global minimizer by $S^\dagger \equiv S^\dagger(\mu)$.

In practice, μ is unknown and the search for the model $S^\dagger$ has to rely on finite data. If $\hat{\mu}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure induced by an i.i.d. sample $\mathbf{X} = (X_1, ..., X_n) \sim \mu^n$, then the empirical objective is the plug-in estimate

$$\begin{aligned} &\Phi_W(\hat{\mu}(\mathbf{X})_S) \\ &= \frac{1}{n}\sum_{i=1}^n d_S(X_i)\, W\left(\frac{1}{n}\left|\{X_j : d_S(X_j) \le d_S(X_i)\}\right|\right) \\ &= \frac{1}{n}\sum_{i=1}^n d_S(X)_{(i)}\, W\left(\frac{i}{n}\right), \end{aligned} \tag{7}$$

where $d_S(X)_{(i)}$ is the i-th smallest member of $\{d_S(X_1), ..., d_S(X_n)\}$.

The empirical estimate $\Phi_W(\hat{\mu}(\mathbf{X})_S)$ is an L-statistic (Serfling, 1980). We denote a minimizer of this objective by

$$\hat{S}(\mathbf{X}) = \arg\min_{S \in \mathcal{S}} \Phi_W(\hat{\mu}(\mathbf{X})_S). \tag{8}$$

In the sequel we study three questions:

[3] This is in contrast with the assumptions made in adversarial learning, where the goal is to increase robustness against adversarial worst-case perturbations (see e.g. Lee & Raginsky, 2018).

1 If the underlying probability measure is a mixture $\mu = (1-\lambda)\,\mu^* + \lambda\nu$ of an unperturbed measure μ^* and a perturbing measure ν, and $S^\dagger = S^\dagger(\mu)$ is the minimizer of (6), under which assumptions will the reconstruction error $\Phi\left(\mu^*_{S^\dagger}\right)$ incurred by $S^\dagger$ on the unperturbed distribution approximate the minimal reconstruction error $\Phi\left(\mu^*_{S^*}\right)$?

2 When solving (6) for a finite amount of data $\mathbf{X}$, under which conditions can we reproduce the behavior of $S^\dagger$ by the empirical minimizer $\hat{S}(\mathbf{X})$ in (8)?

3 How can the method be implemented and how does it perform in practice?

4. Resilience to Perturbations

Before we address the first question we make a preliminary observation in the tradition of robust statistics and compare the influence functions of the functional Φ_1 to that one of the proposed Φ_W with bounded W, and $W(t) = 0$ for $\zeta \leq t < 1$. While we saw in (4) that for any $\rho \in \mathcal{P}([0,\infty))$ the influence function $\mathrm{IF}(R;\rho,\Phi) = R - \Phi(\rho)$ is unbounded in R, in the case of Φ_W we have, for any $R \in \mathbb{R}^d$, that

$$\begin{aligned}\mathrm{IF}(R;\rho,\Phi_W) &\leq \mathrm{IF}_{\max}(\rho, W)\\ &:= \int_0^{F_\rho^{-1}(\zeta)} W\left(F_\rho(r)\right) F_\rho(r)\, dr.\end{aligned}$$

Notice that the right hand side is always bounded, which already indicates the improved robustness of Φ_W (Hampel, 1974). The upper bound $\mathrm{IF}_{\max}$ on the influence function plays also an important role in the subsequent analysis.

Returning now to the data generating mixture $\mu = (1-\lambda)\,\mu^* + \lambda\nu$, where $\mu^* \in \mathcal{P}\left(\mathbb{R}^d\right)$ is the the ideal, unperturbed distribution and $\nu \in \mathcal{P}\left(\mathbb{R}^d\right)$ the perturbation, we make the following assumption.

Assumption A. There exists $S_0 \in \mathcal{S}$, $\delta > 0$, $\beta \in (0, 1-\lambda)$ and a scale parameter $r^* > 0$ (in units of squared euclidean distance), such that for every model $S \in \mathcal{S}$ satisfying $\Phi\left(\mu^*_S\right) > \Phi\left(\mu^*_{S_0}\right) + \delta$ we have $F_{\mu_S}(r) < \beta F_{\mu^*_{S_0}}(r)$ for all $r \leq r^*$.

Assumption A does not depend so much on the richness of $\mathcal{S}$ (which will be relevant to generalization) but on the concentration properties of μ^* and ν (see the extreme example in the supplement). Loosely speaking the assumption prescribes that, under the perturbed distribution μ, any model S with a large reconstruction error on μ^*, should have its losses far less concentrated than the losses of S_0 for small values of r (any $r \leq r^*$). It generally helps if the perturbing distribution ν has a bounded density with a small bound, so that its contributions to $F_{\mu_S}(r)$ remain small for small values of r. Illustrating examples, which apply to the cases of K-MEANS clustering and principal subspace analysis are given in Figures 1 and 2.

We now state the main result of this section.

Theorem 1. *Let $\mu^*, \nu \in \mathcal{P}\left(\mathbb{R}^d\right)$, $\mu = (1-\lambda)\,\mu^* + \lambda\nu$, and $\lambda \in (0,1)$ and suppose there are S_0, r^*, $\delta > 0$ and $\beta \in (0, 1-\lambda)$, satisfying Assumption A. Let W be nonzero on a set of positive Lebesgue measure, nonincreasing and $W(t) = 0$ for $t \geq \zeta = F_{\mu_{S_0}}(r^*)$. Then $\mathrm{IF}_{\max}(\mu_{S_0}, W) > 0$, and if any $S \in \mathcal{S}$ satisfies*

$$\Phi_W(\mu_S) - \Phi_W(\mu_{S_0}) \leq \left(1 - \frac{\beta}{1-\lambda}\right) \mathrm{IF}_{\max}(\mu_{S_0}, W) \quad (9)$$

then we have that $\Phi\left(\mu^*_S\right) \leq \Phi\left(\mu^*_{S_0}\right) + \delta$. In particular we always have that $\Phi\left(\mu^*_{S^\dagger}\right) \leq \Phi\left(\mu^*_{S_0}\right) + \delta$.

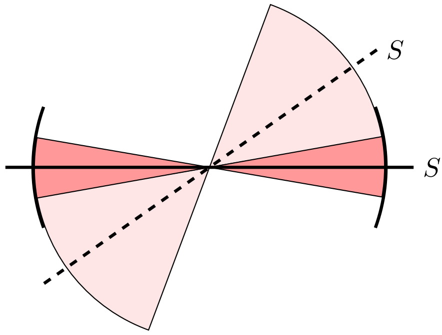

Figure 2. Illustration of Theorem 1 for $d = 2$ and $k = 1$ in the case of PSA. The target distribution (dark gray) is concentrated on the subspace S^*, while the perturbing distribution (light gray) does not concentrate well on any individual subspace.

We close this section by stating some important conclusions of the above theorem.

1. A simplifying illustration of Theorem 1 for principal subspace analysis is provided by Figure 2. The distributions μ^* and ν are assumed to have uniform densities $\rho(\mu^*)$ and $\rho(\nu)$ supported on dark red and light red areas of the unit disk respectively. Suppose $\beta = \rho(\nu)/\rho(\mu^*) < 1 - \lambda$, let $r^* = \sin^2(\pi/\rho(\mu^*))$ and $\delta = 4r^*$. If $\Phi\left(\mu^*_S\right) > \Phi\left(\mu^*_{S^*}\right) + \delta$ then the direction of the subspace S does not intersect the black part of the unit circle and therefore $F_{\mu_S}(r) \leq \beta F_{\mu^*_{S^*}}(r)$ for all $r \leq r^*$. Thus Assumption A is satisfied and consequently, if $W(t) = 0$ for $t \geq F_{\mu_{S^*}}(r^*)$, then $S^\dagger$ must intersect the black part of the unit circle and $\Phi\left(\mu^*_{S^\dagger}\right) \leq \Phi\left(\mu^*_{S^*}\right) + \delta$. Refer also to Figure 2 for an illustration.

2. The generic application of this result assumes that $S_0 = S^*(\mu^*)$, but this is not required. Suppose $\mathcal{S}$

is the set of singletons and μ^* is bimodal, say the mixture of distant standard normal distributions, and $\lambda = 0$ for simplicity. Clearly there is no local concentration on the midpoint $S^*(\mu^*)$, but there is on each of the modes. If S_0 is the mean of the first mode and ζ is sufficiently small, then $S^\dagger$ can be near the mean of the other mode, because it has comparable reconstruction error. In this way the result also explains the astonishing behavior of our algorithm in clustering experiments with mis-specified number of clusters. Refer also to Figure 3 (top right) for an illustration.

3. The conditions on W prescribe an upper bound on the cutoff parameter ζ. If the cutoff parameter ζ is chosen smaller (so that $W(t) = 0$ for $t \geq \zeta \ll F_{\mu_{S^*}}(r^*)$), the required upper bound in (9) decreases and it becomes more difficult to find S satisfying the upper bound. This problem becomes even worse in practice, because the bounds on the estimation error also increase with ζ, as we will see in the next section.

5. Generalization Analysis

Up to this point we were working with distributions and essentially infinite data. In practice we only have samples $\mathbf{X} \sim \mu^n$ and then it is important to understand to which extend we can obtain the conclusion of Theorem 1, when S is the minimizer of the empirical robust functional $\Phi_W(\hat{\mu}(\mathbf{X})_S)$. This can be settled by a uniform bound on the estimation error for Φ_W.

Proposition 2. *Under the conditions of Theorem 1 with $\mathbf{X} \sim \mu^n$ we have that*

$$\begin{aligned}&\Pr\left\{\Phi\left(\mu^*_{\hat{S}(\mathbf{X})}\right) \leq \Phi\left(\mu^*_{S^*}\right) + \delta\right\} \\ &\quad\geq \Pr\left\{2\sup_{S\in\mathcal{S}}\left|\Phi_W(\mu_S) - \Phi_W(\hat{\mu}_S(\mathbf{X}))\right|\right. \\ &\qquad\qquad\left.\leq \left(1 - \frac{\beta}{1-\lambda}\right)\mathrm{IF}_{\max}(\mu_{S^*}, W)\right\}.\end{aligned}$$

The left hand side is the probability that the minimization of our robust L-statistic objective returns a δ-optimal model for the target distribution μ^*. The right hand side goes to 1 as n grows. As we show next, this is due to the fact that the class $\{\mu_S\}$ enjoys a uniform convergence property with respect to the functional Φ_W. Particularly, we present three uniform bounds that control the rate of decay of the same estimation error $|\Phi_W(\mu_S) - \Phi_W(\hat{\mu}_S(\mathbf{X}))|$.

The first two bounds are dimension-free and rely on Rademacher and Gaussian averages of the function class $\{x \mapsto d(x,S) : S \in \mathcal{S}\}$. Bounds for these complexity measures in the practical cases considered can be found in (Maurer & Pontil, 2010; Lee & Raginsky, 2019; Vainsencher et al., 2011). Our last bound is dimension dependent but may outperform the other two if the variance of the robust objective is small under its minimizer. All three bounds require special properties of the weight function W.

For this section we assume $\mu \in \mathcal{P}(\mathbb{R}^d)$ to have compact support, write $\mathcal{X}$ =support(μ) and let $\mathcal{F}$ be the function class

$$\mathcal{F} = \{x \in \mathcal{X} \mapsto d(x,S) : S \in \mathcal{S}\}.$$

We also set $R_{\max} = \sup_{f\in\mathcal{F}}\|f\|_\infty$.

The first bound is tailored to the hard-threshold $\zeta^{-1}1_{[0,\zeta]}$. It follows directly from the elegant recent results of (Lee et al., 2020). For the benefit of the reader we give a proof in the appendix, without any claim of originality and only slightly improved constants.

Theorem 3. *Let $W = \zeta^{-1}1_{[0,\zeta]}$ and $\eta > 0$. With probability at least $1-\eta$ in $\mathbf{X} \sim \mu^n$ we have that*

$$\begin{aligned}&\sup_{S\in\mathcal{S}}\left|\Phi_W(\mu_S) - \Phi_W(\hat{\mu}_S(\mathbf{X}))\right| \\ &\quad\leq \frac{2}{\zeta n}\mathbb{E}_{\mathbf{X}}\mathcal{R}(\mathcal{F},\mathbf{X}) + \frac{R_{\max}}{\zeta\sqrt{n}}\left(2 + \sqrt{\frac{\ln(2/\eta)}{2}}\right),\end{aligned}$$

where $\mathcal{R}(\mathcal{F},\mathbf{X})$ is the Rademacher average

$$\mathcal{R}(\mathcal{F},\mathbf{X}) = \mathbb{E}_\epsilon\left[\sup_{S\in\mathcal{S}}\sum_{i=1}^n \epsilon_i d(X_i,S)\right]$$

with independent Rademacher variables $\epsilon = (\epsilon_1,\ldots,\epsilon_n)$.

The next bound requires boundedness and a Lipschitz property for the weight function W which can otherwise be arbitrary. We define the norm $\|W\|_\infty = \sup_{t\in[0,1]}|W(t)|$ and seminorm $\|W\|_{\mathrm{Lip}} = \inf\{L : \forall t,s \in [0,1],\ W(t) - W(s) \leq L|t-s|\}$.

Theorem 4. *For any $\eta > 0$*

$$\begin{aligned}&\sup_{S\in\mathcal{S}}\left|\Phi_W(\mu_S) - \Phi_W(\hat{\mu}_S(\mathbf{X}))\right| \\ &\quad\leq \frac{2\sqrt{\pi}\left(R_{\max}\|W\|_\infty + \|W\|_{\mathrm{Lip}}\right)}{n}\mathbb{E}_{\mathbf{X}}\mathcal{G}(\mathcal{F},\mathbf{X}) \\ &\qquad + R_{\max}\|W\|_\infty\sqrt{\frac{2\ln(2/\eta)}{n}}\end{aligned}$$

where $\mathcal{G}(\mathcal{F},\mathbf{X})$ is the Gaussian average

$$\mathcal{G}(\mathcal{F},\mathbf{X}) = \mathbb{E}_\gamma\left[\sup_{S\in\mathcal{S}}\sum_{i=1}^n \gamma_i d(X_i,S)\right],$$

with independent standard normal variables $\gamma_1,\ldots,\gamma_n$.

Our last result also requires a Lipschitz property for W and uses a classical counting argument with covering numbers for a variance-dependent bound.

Theorem 5. *Under the conditions of the previous theorem, with probability at least $1-\eta$ in $\mathbf{X}\sim\mu^n$ we have that for all $S\in\mathcal{S}$*

$$\begin{aligned}&\left|\Phi_W\left(\mu_S\right)-\Phi_W\left(\hat{\mu}_S\left(\mathbf{X}\right)\right)\right|\\&\quad\leq\sqrt{2V_SC}+\frac{6R_{\max}\left(\left\|W\right\|_\infty+\left\|W\right\|_{\mathrm{Lip}}\right)C}{n}\\&\qquad\qquad+\frac{\left\|W\right\|_\infty R_{\max}}{\sqrt{n}},\end{aligned}$$

where V_S is the variance of the random variable $\Phi_W\left(\hat{\mu}_S\left(\mathbf{X}\right)\right)$, and C is the complexity term

$$C=kd\ln\left(16n\left\|\mathcal{S}\right\|^2/\eta\right)$$

if $\mathcal{S}$ is the set of sets with k elements, or convex polytopes with k vertices and $\|\mathcal{S}\|=\sup_{x\in S\in\mathcal{S}}\|x\|$, or

$$C=kd\ln\left(16nR_{\max}^2/\eta\right)$$

if $\mathcal{S}$ is the set of set of k-dimensional subspaces.

We state two important conclusion from the above theorems.

1. Our bounds decrease at least as quickly as $n^{-1/2}\ln n$. However, the bound in the last theorem may be considerably smaller than the previous two if n is large and the unperturbed distribution is very concentrated. The last term, which is of order $n^{-1/2}$ does not carry the burden of the complexity measure and decays quickly. The second term contains the complexity, but it decreases as n^{-1}. It can be shown from the Efron-Stein inequality (see e.g. Boucheron et al., 2013, Theorem 3.1) that the variance V_S of our L-statistic estimator is at most of order n^{-1}, so the entire bound is at most of order $n^{-1/2}\ln n$. On the other hand V_s can be very small. For example, if the unperturbed distribution is completely concentrated at S^* and ζ is chosen appropriately $V_{S^*}=0$ and, apart from the complexity-free last term the decay is as $n^{-1}\ln n$.

2. The above bounds implies that, by equating the estimation error to $\frac{1}{2}\left(1-\frac{\beta}{1-\lambda}\right)\mathrm{IF}_{\max}\left(\mu_{S^*},W\right)$ and solving for η, our method recovers a δ-optimal (w.r.t. μ^*) model with probability at least equal to $1-\exp(-n)$.

Finally, we highlight that the above uniform bounds may be of independent interest. For example, consider the case that the test data also come from the perturbed distribution. In such a situation one might be interested in evaluating the performance of the learned model only on data that fit the model class, i.e. $\Phi_W(\mu_S)$. These bounds guarantee that by minimizing the empirical robust functional, one also get good performances on future data from the same distribution.

6. Algorithms

In this section we present our algorithm for (approximately) minimizing the robust L-statistic $\Phi_W(\hat{\mu}(\mathbf{X})_S)$ w.r.t. model $S\in\mathcal{S}$. Throughout we assume W non-increasing and fixed, and to simplify the notation we use the shorthand $\hat{\Phi}_S(\mathbf{X})\equiv\Phi_W(\hat{\mu}(\mathbf{X})_S)$.

6.1. General Algorithm

Let $\mathbf{x}=(x_1,\ldots,x_n)$ be a realization of $\mathbf{X}\sim\mu^n$, consider the following function of $S\in\mathcal{S}$

$$\hat{\Phi}_S(\mathbf{x})=\frac{1}{n}\sum_{i=1}^{n}W\left(\frac{\pi(i)}{n}\right)d_S(x_i)\qquad(10)$$

where π is the ascending ordering of the $d_S(x)_{(i)}$ and notice that minimizing (10) is equivalent to minimize (7). Let p any fixed element in Sym_n[4] and let

$$\phi_S(\mathbf{x},p)=\frac{1}{n}\sum_{i=1}^{n}W\left(\frac{p(i)}{n}\right)d_S(x_i).$$

In the following we will leverage the following property of ϕ_S.

Lemma 6. *For any $S\in\mathcal{S}$ and any $p\in\mathrm{Sym}_n$, if π is the ascending ordering of the $d_S(x_i)$s, then $\phi_S(\mathbf{x},p)\geq\phi_S(\mathbf{x},\pi)=\hat{\Phi}_S(\mathbf{x})$.*

We need also the following definition.

Definition 7. *A mapping $\mathcal{D}:\mathcal{S}\times S_n\to\mathcal{S}$ is a Descent Oracle for ϕ_S iff for any $S\in\mathcal{S}$ and any $p\in\mathrm{Sym}_n$, $\phi_{\mathcal{D}(S,p)}(\mathbf{x},p)\leq\phi_S(\mathbf{x},p)$.*

The algorithm attempts to minimize (10) via alternating minimization of ϕ_S. At the beginning, it picks an initial model S_0 and sort the induced losses in ascending order, i.e. pick the optimal permutation π_0. Then it starts iterating this two steps by first calling the descent oracle $\mathcal{D}(S_t,\pi_t)$ and then sorting the induced losses. At each step either the permutation π_t or the model S_t are fixed. Pseudocode is given in Algorithm 1. Indeed, at each step the algorithm first finds a descending iteration S_{t+1} of $\phi_{S_t}(\mathbf{x},\pi_t)$ and then sort the losses according to π_{t+1}, an operation that by Lemma 6 cannot increase the value of $\phi_{S_{t+1}}$. Thus the following holds.

Theorem 8. *Algorithm 1 is a descent algorithm for the problem of minimizing (10), i.e. for any t, $\hat{\Phi}_{S_{t+1}}(\mathbf{x})\leq\hat{\Phi}_{S_t}(\mathbf{x})$.*

This algorithm is general and to apply it to a specific learning problem an implementation of the descent oracle is needed.

[4] Here Sym_n denotes the set of all $n!$ permutations over n objects.

Algorithm 1

```
Pick any $S_0 \in \mathcal{S}$
$\pi_0 \leftarrow \arg\min_{p \in \mathrm{Sym}_n} \phi_{S_0}(\mathbf{x}, p)$
for $t = 1, \ldots, T$ do
    $S_t \leftarrow \mathcal{D}(S_{t-1}, \pi_{t-1})$
    $\pi_t \leftarrow \arg\min_{p \in S_n} \phi_{S_t}(\mathbf{x}, p)$
end for
return $S_T$
```

The efficiency of Algorithm 1 depends upon such oracle. In the following we show two descent oracles for the cases of KMEANS and PSA. On the negative side notice that in the case of KMEANS when W is the identity, the problem reduces to finding the optimal KMEANS solution, a problem which is known to be hard (further hardness evidence are provided in the supplement). Thus, in the general case, it is not possible to solve our empirical problem optimally. Our algorithms, are a first step towards the design of methods with provable approximation guarantees.

k-Means Clustering (KMEANS). In this case $\mathcal{S}$ is the set of all possible k-tuples of centers in $\mathbb{R}^d$ and $d_S(x) = \min_{c \in S} \|x - c\|_2^2$. Keeping fixed the permutation p, we consider as descent oracle the following Lloyd-like update for the centers. Each center $c \in S$ induces a cluster formed by a subset of training points x_i, $i \in \mathcal{I}$ which are closer to c than every other center (breaking ties arbitrarily). The overall *loss* of representing point in $\mathcal{I}$ with c is

$$\sum_{i \in \mathcal{I}} W\left(\frac{p(i)}{n}\right) \|x_i - c\|_2^2.$$

This loss is minimized at

$$\hat{c} = \frac{1}{\sum_{i \in \mathcal{I}} W\left(\frac{p(i)}{n}\right)} \sum_{i \in \mathcal{I}} W\left(\frac{p(i)}{n}\right) x_i,$$

so the following holds.

Proposition 9. *Given S and p, the mapping that for every $c \in S$ returns the $\hat{c}$ defined above is a descent oracle for* KMEANS *and its runtime is $O(nkd)$.*

The resulting algorithm can is a generalization of the method proposed in (Chawla & Gionis, 2013).

Principal Subspace Analysis (PSA). In this case $\mathcal{S}$ is the set of all possible $d \times k$ matrices U such that $U^\top U = I_d$, $d_S(x) = \|x - UU^\top x\|_2^2$ and

$$\phi_U(\mathbf{x}, p) = \sum_{i=1}^{n} W\left(\frac{p(i)}{n}\right) \|x - UU^\top x_i\|_2^2.$$

Given p, it is easy to see that the above function is minimized at the matrix $\hat{U}$ formed by stacking as columns the k eigenvectors of $\sum_i^n W\left(\frac{p(i)}{n}\right) x_i x_i^\top$ associated to the top k eigenvalues, so the following holds.

Proposition 10. *Given U and p, the mapping that returns the $\hat{U}$ defined above is a descent oracle for* PSA *and its runtime is $O(\min\{d^3 + nd^2, n^3 + n^2 d\})$.*

7. Experiments

The purpose of the numerical experiments is to show that:

- Our algorithms for PSA and KMEANS outperform standard SVD, KMEANS++ and the *Spherical Depth* method (SD) in presence of outliers, while obtain similar performances on clean data.
- Our algorithms on real data are not too sensitive to the parameters of the weight function. In particular, we show that there exist a wide-range of ζ values such that using the hard-threshold function leads to good results.
- In the case of KMEANS our method is able to accurately reconstruct some of the true centers even when the value of k is miss-specified. This matches the second remark after Theorem 1.

Implemented Algorithms. For KMEANS++ we used the *sklearn* implementation fed with the same parameters for the maximum number of iterations T and the initializations r we used for our method. Notice that T is only an upper bound to the number of iterations, the algorithms stop when the difference between the current objective value and the previous one is smaller than 10^{-7}. To set r we used the largest value before diminishing returns were observed. For standard PSA we compute the SVD of $\sum_i x_i x_i^\top$. The SD method is a general purpose pre-processing technique that is applied on the data before performing KMEANS and PSA (see e.g. Elmore et al., 2006; Fraiman et al., 2019). This method computes a score for each point in the dataset by counting in how many balls, whose antipodes are pairs of points in the data, it is contained. The $1 - \zeta n$ points with the smallest scores are discarded. If the data contain n points, the methods needs to check $O(n^2)$ balls for each of the n point resulting in a runtime of $O(n^3)$. For scalability on real data, we implemented a randomized version of this method that for each point only check M balls picked uniformly at random from the set of all possible balls and used $M = O(n)$; the resulting runtime is $O(n^2)$. In the following we refers to our methods as RKM and RPSA respectively. All experiments have been run on an standard laptop equipped with an Intel i9 with 8 cores each working at 2,4 GHz and 16 GB of RAM DDR4 working at 2,6 GHz.

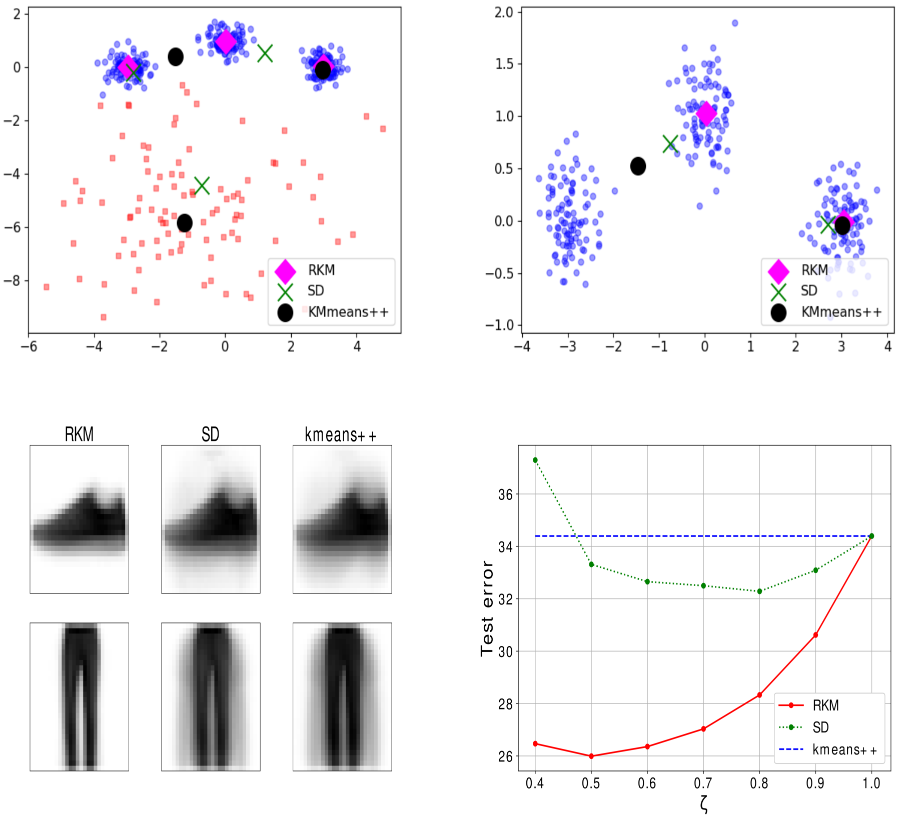

Figure 3. Experiments for KMEANS on synthetic data (top row) and real data (bottom row).

7.1. KMEANS Clustering

Synthetic Data. We run two experiments with artificial data in $\mathbb{R}^2$. In the first experiment, we generated 300 inliers from 3 isotropic truncated Gaussians (100 points each) with variance 0.1 along both axis and mean $(-3, 0)$, $(0, 1)$ and $(3, 0)$ respectively. We then corrupt the data adding 100 points from a fourth isotropic truncated Gaussian centered at $(-1, -5)$ with variance 5 along both axis. For both RKM and KMEANS++ we $T = 10$ and $r = 30$. We initialized RKM with uniform centers and set $\zeta = 0.75$, the same ζ is used for SD. Results are shown in Figure 3 top left, where it is possible to see that while RKM recovers the true centers, SD and KMEANS++ both fail badly placing one centers in the middle of the two clusters and the other close to the mean of the perturbing distribution. In the second experiment, we generated 300 points from the same 3 inliers Gaussians and set the algorithms with $k = 2$ and $\zeta = 0.6$, while T and r are as above. Results are shown in the top right of Figure 3, where it is possible to see that KMEANS++ and SD – although to a lesser extend – wasted a center to merge 2 clusters, while RKM correctly recovers 2 out of the 3 centers.

Real Data. In the synthetic experiments we choose ζ according to the exact fraction of outliers, a quantity which is usually unknown in practice. Here we show that there is a wide range of values for ζ such that RKM performs better than KMEANS++. We used the Fashion-MNIST dataset which consists of about 70000 28×28 images of various types of clothes splitted in a training set of 60000 images and a test set of 10000 images. Specifically, there are 10 classes in the dataset: t-shirts, trousers, pullover, dresses, coats, sandals, shirts, sneakers, bags and ankle boots. The training data were generated by sampling 1000 points, from the training set, each from the sneakers and the trousers classes as inliears, and 250 points from each other class as outliers. The resulting fraction of outliers is about 0.5. The test data consist of all the sneakers and the trousers in the test set and has size of about 2000. We run the algorithms with $T = 50, r = 30, M = 4000, k = 2$ and ζ in the range $[0.4, 1]$. Results are shown in the bottom row of Figure 3. In the lower left, it is possible to see that the centers learned by RKM at the optimal threshold value $\zeta = 0.5$ look good, while the centers found by SD and KMEANS++ are affected by the outliers. Specifically, the such centers arise from the overlap of multiple classes. One center suffers from the effect of the other two shoes classes (sandald and boots)

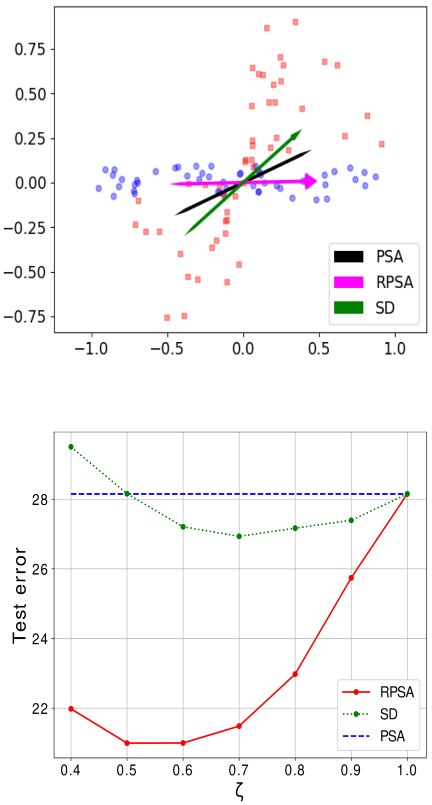

Figure 4. Experiments for PSA on synthetic data (top) and real data (bottom).

as witnessed by the elongated background area, while the other is affected by the clothes classes (most noticeably, the coats) as suggested by background shadow. As for the reconstruction error, RKM outperforms SD uniformly over the range of considered values of ζ.

7.2. Principal Subspace Analysis

Synthetic Data. We run a synthetic experiment with artificial data in $\mathbb{R}^2$. We generate 50 points from the uniform distribution over $[-1, 1] \times [-0.1, 0.1]$ as inliers and 50 points for the uniform distribution over $\mathbb{R}_{++} \cup \mathbb{R}_{--} \cap B(0, 1)$[5] as outliers. We run RPSA with $T = 50$, $r = 30$, $\zeta = 0.5$ and initialize U as a normalized Gaussian matrix. We set $k = 1$ for all algorithms. Results are shown in the top plot of Figure 4 where it is possible to see that the principal subspace learned by RPSA is not affected by the outliers, as opposed to SD and PSA.

Real Data. Similarly to the case of KMEANS, we tested our method on real data for a range of values of ζ. We used again the same setting as before on the Fashion-MNIST dataset. We run the algorithms we $T = 50$, $r = 5$, $M = 4000$, $k = 2$ and ζ in the range $[0.4, 1]$. Results are shown in the bottom plot of Figure 4, where it is possible our algorithm outperforms both PSA and does better than SD.

[5] Here with $\mathbb{R}_{++}$ and $\mathbb{R}_{--}$ we denote the top right and the bottom left orthant of $\mathbb{R}^2$.

8. Conclusions and Future Works

In this work, we address the important problem of designing robust methods for unsupervised learning. We proposed a novel general framework, based on the minimization of an L-statistic, to design algorithms that are resilient to the presence of outliers and/or to model miss-specification. Our method has strong statistical guarantees, is flexible enough to incorporate many problems in unsupervised learning and is effective in practice as the experiments reveal. On the other hand, several extensions can be considered. First, here we studied in details KMEANS and PSA, but our theory also covers the cases of KMEDIAN, sparse coding or non-negative matrix factorization. A related improvement also regards the design of methods for the choice of ζ which do not require an estimate of the fraction of outliers. Second, we believe that this framework can be extended to supervised learning problems such us canonical correlation analysis and partial least squares. Third, our algorithm has only a descent property, and it would be interesting to design algorithms with stronger guarantees such as provable approximation properties.

References

Angluin, D. and Laird, P. D. Learning from noisy examples. *Mach. Learn.*, 2(4):343–370, 1987. doi: 10.1007/BF00116829. URL https://doi.org/10.1007/BF00116829.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Chawla, S. and Gionis, A. k-means-: A unified approach to clustering and outlier detection. In *Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA*, pp. 189–197. SIAM, 2013. doi: 10.1137/1.9781611972832.21. URL https://doi.org/10.1137/1.9781611972832.21.

Cuesta-Albertos, J. A., Gordaliza, A., Matrán, C., et al. Trimmed k-means: An attempt to robustify quantizers. *Annals of Statistics*, 25(2):553–576, 1997.

Diakonikolas, I. and Kane, D. M. Robust high-dimensional statistics. In Roughgarden, T. (ed.), *Beyond the Worst-Case Analysis of Algorithms*, pp. 382–402. Cambridge University Press, 2020. doi: 10.1017/9781108637435.023. URL https://doi.org/10.1017/9781108637435.023.

Elmore, R. T., Hettmansperger, T. P., and Xuan, F. Spherical data depth and a multivariate median. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72:87, 2006.

Fraiman, R., Gamboa, F., and Moreno, L. Connecting pairwise geodesic spheres by depth: DCOPS. *J. Multivar. Anal.*, 169:81–94, 2019. doi: 10.1016/j.jmva.2018.08.008. URL https://doi.org/10.1016/j.jmva.2018.08.008.

Garcia-Escudero, L. A. and Gordaliza, A. Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94(447):956–969, 1999.

Hampel, F. R. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

Hampel, F. R. Robust statistics: A brief introduction and overview. In *Research report/Seminar für Statistik, Eidgenössische Technische Hochschule (ETH)*, volume 94. Seminar für Statistik, Eidgenössische Technische Hochschule, 2001.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pp. 8527–8537, 2018.

Kumar, M. P., Packer, B., and Koller, D. Self-paced learning for latent variable models. In *Advances in neural information processing systems*, pp. 1189–1197, 2010.

Lee, J. and Raginsky, M. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2687–2696, 2018.

Lee, J. and Raginsky, M. Learning finite-dimensional coding schemes with nonlinear reconstruction maps. *SIAM Journal on Mathematics of Data Science*, 1(3):617–642, 2019.

Lee, J., Park, S., and Shin, J. Learning bounds for risk-sensitive learning. *arXiv preprint arXiv:2006.08138*, 2020.

Lloyd, E. Least-squares estimation of location and scale parameters using order statistics. *Biometrika*, 39(1/2):88–95, 1952.

Maurer, A. and Pontil, M. k-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

Maurer, A. and Pontil, M. Empirical bounds for functions with weak interactions. *arXiv preprint arXiv:1803.03934*, 2018.

Maurer, A. and Pontil, M. Uniform concentration and symmetrization for weak interactions. *arXiv preprint arXiv:1902.01911*, 2019.

Ord, K. Outliers in statistical data : V. Barnett and T. Lewis, 1994, 3rd edition, (John Wiley & Sons, Chichester), 584 pp., [UK pound]55.00, ISBN 0-471-93094-6. *International Journal of Forecasting*, 12(1):175–176, March 1996. URL https://ideas.repec.org/a/eee/intfor/v12y1996i1p175-176.html.

Serfling, R. J. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 1980.

Vainsencher, D., Mannor, S., and Bruckstein, A. M. The sample complexity of dictionary learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 773–788. JMLR Workshop and Conference Proceedings, 2011.