

Robust Risk Minimization for Statistical Learning From Corrupted Data

MUHAMMAD OSAMA ¹, DAVE ZACHARIAH ¹, AND PETRE STOICA ¹

¹Division of System and Control, Department of Information Technology, Uppsala University, 751 05 Uppsala, Sweden

CORRESPONDING AUTHOR: MUHAMMAD OSAMA (e-mail: muhammad.osama@it.uu.se)

This work was supported by the Swedish Research Council under Contracts 2017–04610 and 2018–05040.

ABSTRACT We consider a general statistical learning problem where an unknown fraction of the training data is corrupted. We develop a robust learning method that only requires specifying an upper bound on the corrupted data fraction. The method minimizes a risk function defined by a non-parametric distribution with unknown probability weights. We derive and analyse the optimal weights and show how they provide robustness against corrupted data. Furthermore, we give a computationally efficient coordinate descent algorithm to solve the risk minimization problem. We demonstrate the wide range applicability of the method, including regression, classification, unsupervised learning and classic parameter estimation, with state-of-the-art performance.

INDEX TERMS Data corruption, Huber contamination model, risk minimization, robustness.

I. INTRODUCTION

Statistical learning problems encompass regression, classification, unsupervised learning and parameter estimation [1]. The common goal is to find a model, indexed by a parameter θ , that minimizes some loss function $\ell_{\theta}(\mathbf{z})$ on average, using training data $\mathcal{D} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$. The loss function is chosen to target data from a class of distributions, denoted \mathcal{P}_o .

It is commonly assumed that the training data is drawn from some distribution $p_o(\mathbf{z}) \in \mathcal{P}_o$. In practice, however, training data is often corrupted by outliers, systematic mislabeling, or even an adversary. Under such conditions, standard learning methods degrade rapidly [2]–[10]. See Figure 1 for an illustration. Here we consider the Huber contamination model which is capable of modeling the inherent corruption of data and is common in the robust statistics literature [11]–[13]. Specifically, the training data is assumed to be drawn from the unknown mixture distribution

$$p(\mathbf{z}) = (1 - \epsilon)p_o(\mathbf{z}) + \epsilon q(\mathbf{z}), \quad (1)$$

so that roughly ϵn samples come from a corrupting distribution $q(\mathbf{z}) \notin \mathcal{P}_o$. The fraction of outliers, ϵ , may range between 1–10% in routine datasets, but in data collected with less dedicated effort or under time constraints ϵ can easily exceed 10% [32, ch. 1].

In the robust statistics literature, several methods have been developed for various applications. A classical approach is to modify a given loss function $\ell_{\theta}(\mathbf{z})$ so as to be less sensitive to outliers [5], [12], [13]. Some examples of such functions are the Huber and Tukey loss functions [11], [14]. Another approach is to try and identify the corrupted points in the training data based on some criteria and then remove them [15]–[19]. For example, for mean and covariance estimation of $\mathbf{z} \sim p_o(\mathbf{z})$, the method presented in [20] identifies corrupted points by projecting the training data onto an estimated dominant signal subspace and then compares the magnitude of the projected data against some threshold. The main limitation of the above approaches is that they are problem-specific and must be tailored to each learning problem. In addition, even for fairly simple learning problems, such as inferring the mean, the robust estimators can be computationally demanding [21]–[23].

Recent work has been directed towards developing more general and tractable methods for robust statistical learning that is applicable to a wide range of loss functions [24]–[26], [34]. These state-of-the-art methods do, however, exhibit some important limitations relating to the choice of certain tuning parameters. For instance, the two-step method in [26] uses a regularization parameter which depends on the unknown ϵ that the user may not be able to specify precisely.

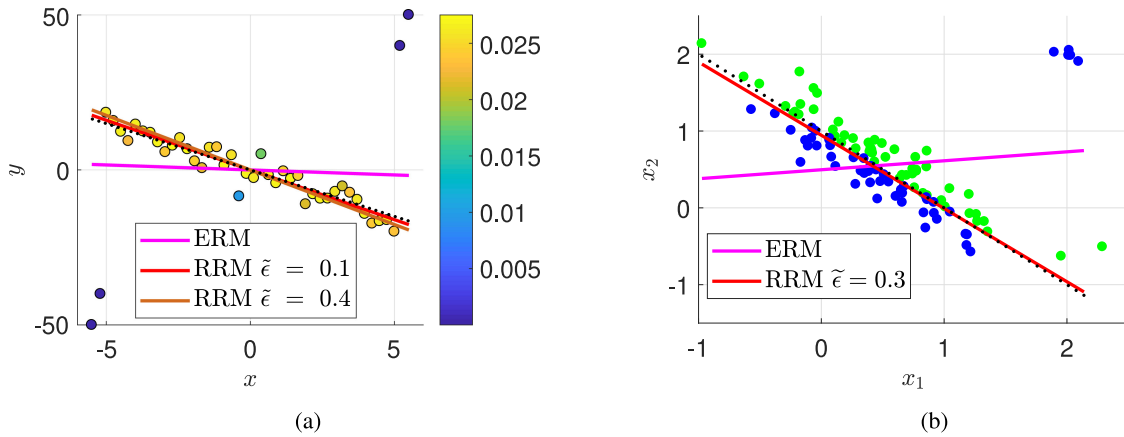


FIGURE 1. Illustration of statistical learning from data with an unknown fraction ϵ of corrupted samples. (a) Regression, where θ parameterizes a linear model. Learned models using standard least squares (ERM) vs. proposed method (RRM) using two different upper bounds $\tilde{\epsilon}$ on ϵ . The target θ^* is illustrated in black dots. The proposed method learns and assigns weights to each observed data point (shown in color scale). The weights of the outlier points are very small and, in turn, contribute marginally to learning the regression line. (b) Binary classification, where θ parameterizes a separating hyperplane. The target θ^* is illustrated in black dots. Learned models using standard logistic regression (ERM) vs. proposed method (RRM) using an upper bound $\tilde{\epsilon}$.

Similarly, under certain conditions, there exists a parameter setting for the method in [24] that yields performance guarantees, but there is no practical means or criterion for how to tune this parameter. The cited methods rely on removing data points based on some score function and comparing to a specified threshold. The extent to which the choice of scoring function is problem dependent is unknown and the choice of threshold depends in practice on some user-defined parameter.

The main contribution of this paper is a general robust method with the following properties:

- it is applicable to any statistical learning problem that minimizes an expected loss function,
- it requires only specifying an upper bound on the corrupted data fraction ϵ ,
- it is formulated as a minimization problem that can be solved efficiently using a blockwise algorithm.

We illustrate and evaluate the robust method in several standard statistical learning problems.

II. PROBLEM FORMULATION

Consider a set of models indexed by a parameter $\theta \in \Theta$. The loss of a model θ is denoted $\ell_\theta(z)$, where $z \sim p_o(z)$ is a randomly drawn datapoint. The target model is that which minimizes the expected loss, or *risk*, i.e.,

$$\theta^* = \arg \min_{\theta \in \Theta} \mathbb{E}_{p_o}[\ell_\theta(z)], \quad (2)$$

In lieu of the unknown target distribution $p_o(z)$, we use data from the training distribution $p(z)$ in (1) to approximate θ^* . Specifically, we obtain n independent samples denoted $\mathcal{D} = \{z_i\}_{i=1}^n$ and common learning strategy is then empirical risk minimizing (ERM), which yields the parameter vector

$$\hat{\theta}_{\text{ERM}} = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell_\theta(z_i) \quad (3)$$

Example 1: In regression problems, data consists of features and outcomes, $z = (x, y)$, and θ parameterizes a predictor $\hat{y}_\theta(x)$. The standard loss function $\ell_\theta(z) = (y - \hat{y}_\theta(x))^2$ targets distributions with thin-tailed noise.

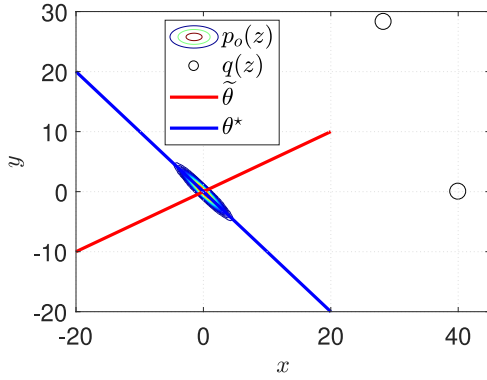
Example 2: In general parameter estimation problems, a standard loss function is $\ell_\theta(z) = -\ln p_\theta(z)$, which targets distributions spanned by $p_\theta(z)$. For this choice of loss function, (3) corresponds to the maximum likelihood estimator.

In real applications, a certain fraction $\epsilon \in [0, 1)$ of the data is *corrupted* such that the minimum risk under the training distribution $p(z)$ is greater than what is achieved under the target distribution. That is,

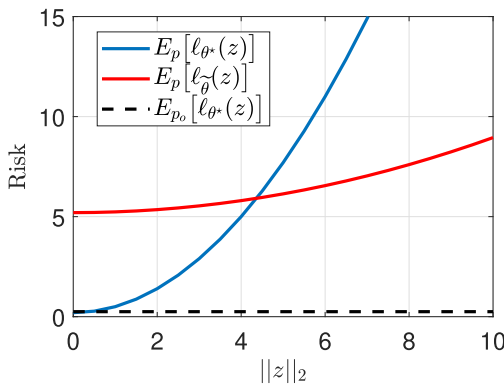
$$\mathbb{E}_p[\ell_\theta(z)] \geq \mathbb{E}_{p_o}[\ell_{\theta^*}(z)], \quad \forall \theta, \quad (4)$$

with equality if and only if $\epsilon = 0$. Under such corrupted data conditions, ERM degrades rapidly as $p(z)$ diverges from $p_o(z)$. Eq. (4) is in fact a necessary condition to identify θ^* from the training distribution (1). If the condition were not satisfied, the fraction $\epsilon > 0$ of data from $q(z)$ would result in a lower minimum risk, such that $\mathbb{E}_p[\ell_\theta(z)] < \mathbb{E}_{p_o}[\ell_{\theta^*}(z)]$ for some $\theta \neq \theta^*$. In this case, by definition, the training distribution would *not* be corrupted with respect to the loss function under consideration.

Example 3: To illustrate the degradation of ERM as $p(z)$ diverges from $p_o(z)$, consider a linear regression problem where $z = (x, y) \sim p(z)$, using the squared-error loss. Fig. 2 (a) illustrates the target distribution $p_o(z)$, which is a zero-mean two-dimensional Gaussian, and $q(z)$ as a point mass distribution generating corrupted leverage points at equal distances to the origin. The figure also illustrates two contrasting regression models: $\hat{\theta}$ and θ^* . Fig. 2(b) shows the risk $\mathbb{E}_p[\ell_\theta(z)]$ for these two models versus the distance of the corrupted points from the origin. Since ERM minimizes this risk in the large-sample case, it will drastically degrade as it opts for



(a) Distributions for data $z = (x, y)$ along with target model θ^* and an alternative model $\tilde{\theta}$.



(b) Risk vs. distance of corrupting points in $q(z)$ from the origin.

FIGURE 2. Comparison of two regression models under corrupted data with varying distance from the target distribution. At a certain distance, $\tilde{\theta}$ attains lower risk than θ^* under training distribution $p(z)$, but not under the target distribution $p_o(z)$, cf. condition (4). (a) Distributions for data $z = (x, y)$ along with target model θ^* and an alternative model $\tilde{\theta}$. (b) Risk vs. distance of corrupting points in $q(z)$ from the origin.

$\tilde{\theta}$ over θ^* at and beyond a certain distance. We note, however, that $\mathbb{E}_p[l_{\tilde{\theta}}(z)] \geq \mathbb{E}_p[l_{\theta^*}(z)]$. That is, by retaining only a $1 - \epsilon$ fraction of uncorrupted data, θ^* has lower risk than an alternative model θ and therefore is identifiable.

The above example illustrates a more general principle which we will exploit in the subsequent section: The minimum risk when excluding the ϵ fraction of corrupted data is lower than when including it. While ϵ in (1) is unknown, it can typically be upper bounded, that is, $\epsilon \leq \tilde{\epsilon}$ [32, ch. 1]. This implies that we expect at least $(1 - \tilde{\epsilon})n$ uncorrupted samples in \mathcal{D} . In principle one could then consider all possible subsets of samples and retain the set which yields the minimum empirical risk. This combinatorial approach is, however, computationally infeasible.

Our goal is to formulate a general and computationally tractable method of risk minimization which learns a model θ from \mathcal{D} that is robust against corrupted training samples using only a specified bound $\tilde{\epsilon}$. We show that it is also possible to calibrate $\tilde{\epsilon}$ using \mathcal{D} .

III. METHOD

We define the risk evaluated at a distribution $p_\pi(z)$ as

$$R(\theta, \pi) = \mathbb{E}_{p_\pi} [\ell_\theta(z)], \quad (5)$$

and consider the following nonparametric class of distributions, indexed by π ,

$$\left\{ p_\pi : p_\pi(z) = \sum_{i=1}^n \pi_i \delta(z - z_i), \pi \in \Pi \right\} \quad (6)$$

where $z_i \in \mathcal{D}$ and the weights belong to the simplex $\Pi = \{\pi \in \mathbb{R}_+^n : \mathbf{1}^\top \pi = 1\}$. Let the entropy of $p_\pi(z)$ be denoted as

$$\mathbb{H}(\pi) \triangleq - \sum_i \pi_i \ln \pi_i \in [0, \ln n],$$

then it is readily seen that ERM in (3) minimizes the risk $R(\theta, \pi)$ under the maximum-entropy distribution $\pi = n^{-1} \mathbf{1}$ [27]. When $\epsilon = 0$, it is well known this choice yields an asymptotically consistent estimate of θ^* under standard regularity conditions.

A. ROBUST RISK MINIMIZATION

If the support of p_π only covers $(1 - \epsilon)n$ samples drawn from p_o , its maximum entropy is at least $\ln[(1 - \tilde{\epsilon})n]$ since $\tilde{\epsilon} \geq \epsilon$. The risk $R(\theta^*, \pi)$ will then tend to be lower than $R(\theta^*, n^{-1} \mathbf{1})$, since the latter includes ϵn corrupted samples, cf. (4). We propose learning θ by utilizing the distribution in (6) that yields the minimum risk, subject to its entropy $\mathbb{H}(\pi)$ being at least $\ln[(1 - \tilde{\epsilon})n]$. That is, the following robust risk minimization (RRM) approach is proposed

$$\hat{\theta}_{\text{RRM}} = \arg \min_{\theta \in \Theta} \min_{\pi \in \Pi : \mathbb{H}(\pi) \geq \ln[(1 - \tilde{\epsilon})n]} R(\theta, \pi) \quad (7)$$

The entropy constraint ensures that θ is learned using an effective sample size of $(1 - \tilde{\epsilon})n$.

We now analyze the optimal weights $\pi^*(\theta)$ of the inner problem in (7) to understand how the proposed method yields robustness against corrupted samples. Note that the objective function in (7), $R(\theta, \pi) = \sum_{i=1}^n \pi_i \ell_\theta(z_i)$, is linear in π , and that the entropy constraint is convex. The inner optimization problem satisfies Slater's condition, since $\pi = n^{-1} \mathbf{1}$ is a strictly feasible point. Hence strong duality holds and the optimal weights π^* can be obtained using the Karush-Kuhn-Tucker (KKT) conditions [28]. The Lagrangian of the optimization problem with respect to π equals

$$L(\pi, \lambda, \nu) = \sum_{i=1}^n \pi_i \ell_\theta(z_i) + \lambda (\ln[(1 - \tilde{\epsilon})n] - \mathbb{H}(\pi)) + \nu (\mathbf{1}^\top \pi - 1)$$

where λ and ν are the dual variables. The KKT conditions can then be expressed as

$$\begin{aligned} \nabla_\pi L(\pi^*, \lambda^*, \nu^*) &= 0, \quad \mathbf{1}^\top \pi^* = 1, \quad \lambda^* \geq 0, \\ \lambda^* (\ln[(1 - \tilde{\epsilon})n] - \mathbb{H}(\pi^*)) &= 0, \\ \mathbb{H}(\pi^*) &\geq \ln[(1 - \tilde{\epsilon})n] \end{aligned} \quad (8)$$

Here (λ^*, ν^*) are the dual optimal solutions. Solving the above KKT conditions leads to the following optimal weights,

$$\pi_i^*(\theta) = c^* \exp \left[-\frac{\ell_\theta(z_i)}{\lambda^*(\tilde{\epsilon})} \right] \geq 0, \quad i = 1, \dots, n, \quad (9)$$

where c^* is a proportionality constant which ensures that the probability weights sum to 1. It is immediately seen that $\pi^*(\theta)$ downweights the data points with high losses for any given θ . The attenuating factor $1/\lambda^*(\tilde{\epsilon})$ increases with the corruption bound $\tilde{\epsilon}$. It goes to zero when $\tilde{\epsilon}$ vanishes, thereby yielding uniform weights $\pi^*(\theta) \rightarrow n^{-1}\mathbf{1}$ as expected.

Plugging $\pi^*(\theta)$ back into (7) yields the equivalent concentrated problem

$$\min_{\theta \in \Theta} \sum_{i=1}^n \exp \left[-\frac{\ell_\theta(z_i)}{\lambda^*(\tilde{\epsilon})} \right] \ell_\theta(z_i), \quad (10)$$

which focuses the learning of θ on the set of $(1 - \tilde{\epsilon})n$ samples with lowest losses. By downweighting corrupted samples that increase the risk at a given θ , problem (10) and hence (7) provides robustness against outliers in \mathcal{D} . This is achieved without tailoring a new robustified loss function or tuning a loss-specific user parameter to a given problem. Instead, the user needs only specify an upper bound on the corruption fraction, $\tilde{\epsilon}$. While (10) provides insight into the properties of RRM, it does not lend itself to a tractable method.

Remark: The proposed method is readily extendable to expected loss minimization problems with regularization for θ . In that case, the loss $\ell_\theta(z)$ in (5) is simply replaced by

$$\tilde{\ell}_\theta(z) = \ell_\theta(z) + r(\theta),$$

where $r(\theta)$ is the regularization term.

B. BLOCKWISE MINIMIZATION ALGORITHM

We now propose an efficient computational method of finding a solution of (7). Given a fixed parameter $\tilde{\theta}$, we define the weights

$$\hat{\pi}(\tilde{\theta}) = \arg \min_{\pi \in \Pi : \mathbb{H}(\pi) \geq \ln[(1-\tilde{\epsilon})n]} R(\tilde{\theta}, \pi), \quad (11)$$

which is the solution to a convex optimization problem and can be computed efficiently using standard numerical packages, e.g. barrier methods [29] which have polynomial-time complexity. For a given $\tilde{\pi}$, the minimizer

$$\hat{\theta}(\tilde{\pi}) = \arg \min_{\theta \in \Theta} R(\theta, \tilde{\pi}), \quad (12)$$

is the solution to a standard weighted risk minimization problem. Solving both problems in a cyclic manner constitute a blockwise coordinate descent method which we summarize in Algorithm 1. When the parameter set Θ is closed and convex, the algorithm is guaranteed to converge to a critical point of (7), see [30].

The general form of the proposed method renders it applicable to a diverse range of learning problems in which ERM is conventionally used. In the next section, we illustrate the

Algorithm 1: Robust Risk Minimization (RRM).

- 1: Input: \mathcal{D} and $\tilde{\epsilon}$
 - 2: Set $k := 0$ and $\pi^{(0)} = n^{-1}\mathbf{1}$
 - 3: **repeat**
 - 4: $\theta^{(k+1)} = \hat{\theta}(\pi^{(k)})$
 - 5: $\pi^{(k+1)} = \hat{\pi}(\theta^{(k+1)})$
 - 6: $k := k + 1$
 - 7: **until convergence**
 - 8: Output: $\hat{\theta} = \theta^{(k)}, \hat{\pi} = \pi^{(k)}$
-

performance and generality of the proposed method using numerical experiments for different supervised and unsupervised machine learning problems.

IV. NUMERICAL EXPERIMENTS

We illustrate the generality of our framework by addressing four common problems in regression, classification, unsupervised learning and parameter estimation. For the sake of comparison, we also evaluate the recently proposed robust SEVER method [24], which was derived on very different grounds as a means of augmenting gradient-based learning algorithms with outlier rejection capabilities. We use the same threshold settings for the SEVER algorithm as were used in the experiments in [24], with $\tilde{\epsilon}$ in lieu of the unknown fraction ϵ . Code for experiments is available at <https://github.com/Muhammad-Osama/robustRisk>.

A. LINEAR REGRESSION

Consider data $z = (x, y)$, where $x \in \mathbb{R}^{10}$ and $y \in \mathbb{R}$ denote feature vectors and outcomes, respectively. We consider a class of predictors $\hat{y} = x^\top \theta$, where $\Theta = \mathbb{R}^{10}$, and a squared-error predictive loss $\ell_\theta(x, y) = (y - x^\top \theta)^2$. This loss function targets thin-tailed distributions with a linear conditional mean function.

We learn θ using $n = 40$ i.i.d. training samples drawn from

$$p(x, y) = (1 - \epsilon)p(x)p_o(y|x) + \epsilon p(x)q(y|x), \quad (13)$$

where

$$p_o(y|x) = \mathcal{N}(x^\top \theta^*, \sigma^2), \quad q(y|x) = t(x^\top \theta^*, \nu), \quad (14)$$

and $p(x) = \mathcal{U}([-5, 5]^{10})$.¹ The above data generator yields observations concentrated around a hyperplane, where roughly ϵ observations are corrupted by heavy-tailed t -distributed noise. Data is generated with $\theta^* = \mathbf{1}$ and noise standard deviation $\sigma = 0.25$.

We evaluate the distribution of estimation errors $\|\theta^* - \hat{\theta}\|$ relative to $\|\theta^*\|$ using 100 Monte Carlo runs. In the first experiment, we set ϵ to 20% and $\nu = 1.5$, in which case the tails of $q(y|x)$ are so heavy that the variance is undefined. We apply RRM with $\tilde{\epsilon} = 0.40$, which is a conservative upper bound. Note that (12) is a weighted least-squares problem

¹Symbols $\mathcal{N}(\cdot)$, $t(\cdot)$ and $\mathcal{U}(\cdot)$ represent Gaussian-, t - and uniform distributions, respectively.

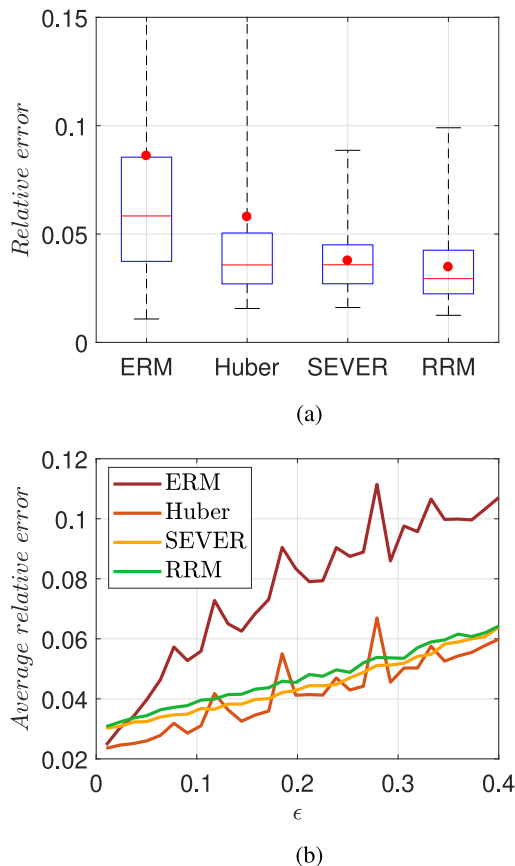


FIGURE 3. Linear regression. (a) Box plot of distribution of relative error $\|\theta^* - \hat{\theta}\|/\|\theta^*\|$ when $\epsilon = 20\%$. Each box spans the 25th to 75th quantiles and the red dots show the means. The whiskers extend to the minimum and maximum values of error. (b) Expected relative error versus percentage of corrupted samples ϵ . Throughout we use the upper bound $\tilde{\epsilon} = 40\%$.

with a closed-form solution. The distribution of errors for ERM, SEVER and RRM are summarized in Figure 3(a). We also include the Huber method, which is tailored specifically for linear regression [5, ch. 2.6.2]. Both RRM and SEVER perform similarly in this case and are substantially better than ERM, reducing the errors by almost a half.

Next, we study the performance as the percentage of corrupted data ϵ increases from 0% to 40%. We set $\nu = 2.5$ so that the variance of the corrupting distribution is defined. Figure 3(b) shows the expected relative error versus ϵ for the different methods, where the robust methods, once again, perform similarly to one another, and much better than ERM.

B. LOGISTIC REGRESSION

Consider data $z = (x, y)$ where $x \in \mathbb{R}^2$ is a feature vector and $y \in \{0, 1\}$ an associated class label. We consider the cross-entropy loss

$$\ell_{\theta}(x, y) = -y \ln(\sigma_{\theta}(x)) - (1 - y) \ln(1 - \sigma_{\theta}(x)), \quad (15)$$

where

$$\sigma_{\theta}(x) = (1 + \exp(-\phi^{\top}(x)\theta))^{-1}$$

and $\phi(x) = [1, x]^{\top}$. Thus the loss function targets distributions with linearly separable classes.

We learn $\theta \in \Theta = \mathbb{R}^3$ using $n = 100$ i.i.d points drawn from

$$p(x, y) = (1 - \epsilon)p_o(x)p_o(y|x) + \epsilon q(x)q(y|x), \quad (16)$$

where $p_o(x) = \mathcal{N}\left(0.5, \begin{pmatrix} 0.25 & -0.25\rho \\ -0.25\rho & 0.25 \end{pmatrix}\right)$ with $\rho = 0.99$. An illustration of $p_o(x, y)$ is given in Figure 4(a), where the separating hyperplane corresponds to $\theta^* = [-1, 1, 1]$. The corrupting distribution is given by $q(x) = \mathcal{N}\left(0.5, \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}\right)$ and $q(y = 0|x) \equiv 1$ as illustrated in Figure 4(b).

Data is generated according to (16) with ϵ equal to 5%.

We apply RRM with $\tilde{\epsilon} = 0.30$. Note that $\hat{\theta}(\tilde{\pi})$ is readily computed using the standard iterative re-weighted least square or majorization-minimization algorithms [1], with minor modifications to take into account the fact that the data points are weighted by π . Figure 4(b) shows the learned separating planes, parameterized by θ , for a single realization. We observe that the plane learned by ERM and SEVER is shifted towards the outliers. By contrast, the proposed RRM method is marginally affected by the corrupting distribution. Figure 4(c) summarizes the distribution of angles between θ^* and $\hat{\theta}$, i.e., $\arccos \frac{\hat{\theta}^{\top} \theta^*}{\|\hat{\theta}\| \|\theta^*\|}$, using 100 Monte Carlo simulations. RRM outperforms the other two methods in this case.

C. PRINCIPAL COMPONENT ANALYSIS

Consider data $z \in \mathbb{R}^2$ where we assume z to have zero mean. Our goal is to approximate z by projecting it onto a subspace. We consider the loss $\ell_{\theta}(z) = \|z - P_{\theta}z\|_2^2$ where P_{θ} is an orthogonal projection matrix. The loss function targets distributions where the data is concentrated around a linear subspace. In the case of a one-dimensional subspace considered here, $P_{\theta} = \theta\theta^{\top}$ where $\Theta = \{\theta \in \mathbb{R}^2 : \|\theta\| = 1\}$.

We learn θ using $n = 40$ i.i.d datapoints drawn from

$$p(z) = (1 - \epsilon) \underbrace{p_o(z_2|z_1)p_o(z_1)}_{p_o(z)} + \epsilon q(z), \quad (17)$$

where

$$p_o(z_2|z_1) = \mathcal{N}(2z_1, \sigma^2), \quad p_o(z_1) = \mathcal{N}(0, 1) \quad (18)$$

and $q(z) = t(\mathbf{0}, \mathbf{I}, \nu)$ for outliers. Note that $p_o(z)$ in (17) corresponds to a subspace parameterized by $\theta^* = [\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}]^{\top}$.

Data is generated with $\sigma = 0.25$, $\nu = 1.5$ and ϵ is set to 20%. We apply RRM with $\tilde{\epsilon} = 0.40$. Note that $\hat{\theta}(\tilde{\pi})$ can be obtained as

$$\hat{\theta}(\tilde{\pi}) = \arg \max_{\theta \in \Theta} \theta^{\top} R \theta, \quad (19)$$

which is equivalent to maximizing a Rayleigh quotient and the solution is simply the dominant eigenvector of the covariance

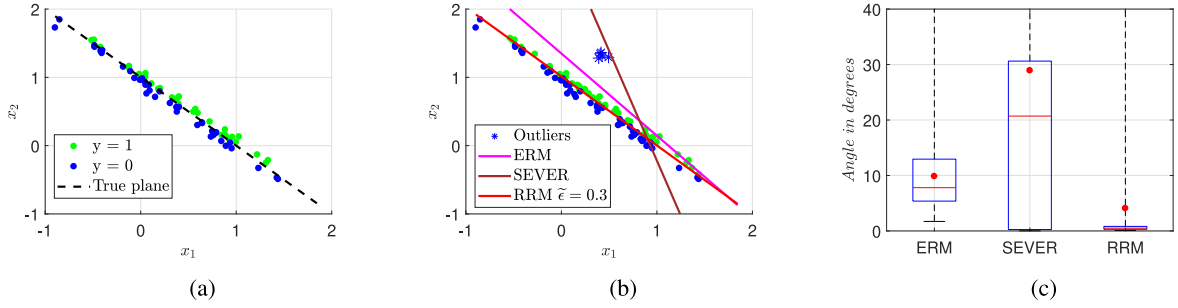


FIGURE 4. Logistic regression: data points with labels 0 and 1 are shown in blue and green, respectively. (a) Single realization from target distribution $p_o(z)$ for linearly separable classes, also shown is the true hyperplane θ^* , (b) Samples from corrupting distribution $q(z)$ (denoted by stars) along with estimated separating hyperplanes $\hat{\theta}$ using ERM and SEVER and RRM methods. (c) Box plot of angle (in degrees) between the true hyperplane θ^* and estimated hyperplanes $\hat{\theta}$ for different methods.

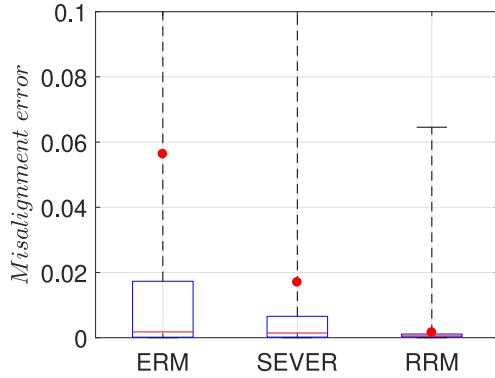


FIGURE 5. Principal component analysis. Box plot of subspace misalignment error $1 - |\cos(\hat{\theta}^\top \theta^*)|$.

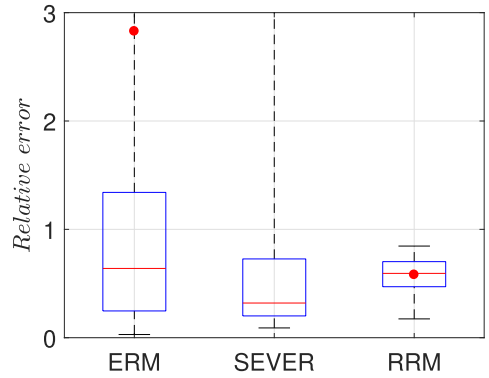


FIGURE 6. Covariance estimation. Box plot of distribution of relative errors $\|\Sigma^* - \hat{\Sigma}\|_F / \|\Sigma^*\|_F$. Note that the expected relative error for SEVER is too large to be contained in the given plot.

matrix

$$\mathbf{R} = \sum_{i=1}^n \tilde{\pi}_i \mathbf{z}_i \mathbf{z}_i^\top. \quad (20)$$

We quantify the misalignment of the subspaces using the metric $1 - |\cos(\hat{\theta}^\top \theta^*)|$ evaluated over 100 Monte Carlo simulations. Figure 5 summarizes the distribution of errors for the three different methods. For this problem, RRM outperforms both ERM and SEVER.

D. COVARIANCE ESTIMATION

Consider data $\mathbf{z} \in \mathbb{R}^2$ with an unknown mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We consider the loss function

$$\ell_{\theta}(\mathbf{z}) = -(\mathbf{z} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) + \ln |\boldsymbol{\Sigma}|$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This loss function targets sub-Gaussian distributions.

We learn $\boldsymbol{\theta}$ using $n = 50$ i.i.d samples drawn from

$$p(\mathbf{z}) = (1 - \epsilon)p_o(\mathbf{z}) + \epsilon q(\mathbf{z}) \quad (21)$$

where $p_o(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*)$ and $q(\mathbf{z}) = t(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu)$. Data is generated using (21) with $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma}^* = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$ and with $\epsilon = 20\%$. We set $\nu = 1.5$, which means that the corrupting distribution $q(\mathbf{z})$ has no finite covariance matrix.

We apply RRM with upper bound $\tilde{\epsilon} = 0.30$. Note that $\hat{\boldsymbol{\theta}}(\tilde{\pi})$ has a closed-form solution, given by the weighted sample mean and covariance matrix with the weight vector equal to $\tilde{\pi}$. We evaluate the error $\|\Sigma^* - \hat{\Sigma}\|_F$ relative to $\|\Sigma^*\|_F$ over 100 Monte Carlo simulations and show it in Figure 6. We see that SEVER is prone to breaking down due to the heavy-tailed outliers, whereas RRM is stable.

E. SENSITIVITY TO UNDER/OVER-ESTIMATION OF ϵ

In this subsection, we study the sensitivity of RRM with a fixed upper bound $\tilde{\epsilon}$, when varying ϵ . We fix $\tilde{\epsilon} = 50\%$ throughout all experiments below.

We first consider the case of linear regression with the same experimental settings as in Section IV-A except number of datapoints $n = 100$. We vary ϵ from 1% to 90%. For each value of ϵ , we generate 200 Monte Carlo data samples and compute the average relative error $\frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|}{\|\boldsymbol{\theta}^*\|}$. Figure 7 shows the average relative error against ϵ for RRM and ERM. Note that when $\epsilon < \tilde{\epsilon}$, i.e., the fraction of corrupted data is overestimated, RRM still gives smaller relative error. As ϵ increases beyond $\tilde{\epsilon}$, the corruption fraction becomes underestimated but the resulting error only increases gradually. This is partly a result of the exponential weighting in (9) which assigns lower

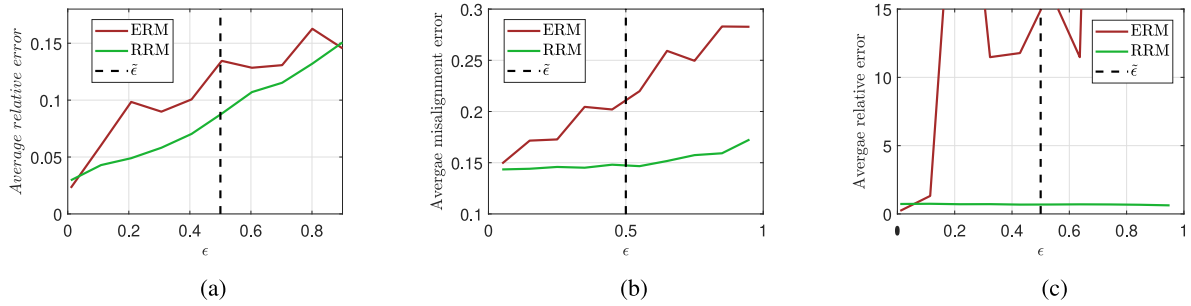


FIGURE 7. Sensitivity to under/over-estimation of ϵ : Estimation errors versus ϵ for (a) Linear regression. (b) Principal component analysis. (c) Covariance estimations. For all experiments $\tilde{\epsilon} = 50\%$.

weights to the corrupted samples and higher weights to the uncorrupted samples.

Figures 7(b) and 7(c) show the average misalignment error $1 - |\cos(\hat{\theta}^\top \theta^*)|$ and the average relative error $\frac{\|\hat{\Sigma} - \Sigma^*\|_F}{\|\Sigma^*\|_F}$ for the principle component and covariance estimation problems of Sections IV-C and IV-D respectively. For both cases, we vary ϵ from 5% to 95%. The average of the error is computed over 50 Monte Carlo simulations. For the principal components, the error increases gracefully with ϵ whereas for covariance estimation it remains more or less the same.

F. CALIBRATING THE UPPER BOUND $\tilde{\epsilon}$

Our experiments have shown that RRM performs well even with a conservative value of $\tilde{\epsilon}$. In this subsection, we show that it is also possible to calibrate the value of $\tilde{\epsilon}$.

Recall that the method estimates a set of probability weights $\hat{\pi}$. A very small weight is thus indicative of a corrupted sample and we propose to estimate ϵ by

$$\hat{\epsilon} = \frac{1}{n} \sum_{i=1}^n 1(\hat{\pi}_i \leq \pi_{\min}), \quad (22)$$

using an appropriately chosen threshold $\pi_{\min} \ll n^{-1}$. By initializing with a high value for $\tilde{\epsilon}$, we obtain $\hat{\pi}$ and estimate $\hat{\epsilon}$ using (22). Then we can calibrate the corruption fraction, setting $\tilde{\epsilon} = \hat{\epsilon}$. We illustrate the idea for linear regression.

We generate $n = 100$ datapoints according to the experimental settings in Section IV-A. We initialize $\tilde{\epsilon}$ by a rather conservative value of 70%. We vary ϵ from 10% to 50%, and for each value of ϵ , we estimate $\hat{\epsilon}$ according to (22) with a threshold $\pi_{\min} = 10^{-4}$. In Figure 8, we compare the estimates $\hat{\epsilon}$ with ϵ , using 50 Monte Carlo runs. It can be seen that the estimates tend to follow ϵ and do not systematically underestimate the unknown corruption fraction. Thus the initial conservative $\tilde{\epsilon}$ can be adjusted downward in a data adaptive manner.

V. REAL DATA

Finally, we test the performance of RRM on real data. We use the Wisconsin breast cancer dataset from the UCI repository [31]. The dataset consists of $n = 683$ points, with features $\mathbf{x} \in \mathbb{R}^9$ and labels $y \in \{0, 1\}$. The class labels 0 and 1

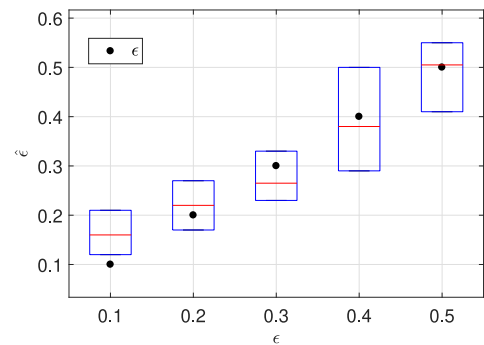


FIGURE 8. $\hat{\epsilon}$ versus ϵ , using 50 Monte Carlo runs in the linear regression example. $\tilde{\epsilon}$ was set to a high value of 70% throughout.

TABLE 1 Confusion Matrix for ERM. Classification Accuracy 89.42%

| | | |
|-----------|-------------|-------------|
| $n = 274$ | Predicted 1 | Predicted 0 |
| Actual 1 | 69 | 28 |
| Actual 0 | 1 | 176 |

TABLE 2 Confusion Matrix for SEVER. Classification Accuracy 89.42%

| | | |
|-----------|-------------|-------------|
| $n = 274$ | Predicted 1 | Predicted 0 |
| Actual 1 | 71 | 26 |
| Actual 0 | 3 | 174 |

TABLE 3 Confusion Matrix for RRM. Classification Accuracy 91.24%

| | | |
|-----------|-------------|-------------|
| $n = 274$ | Predicted 1 | Predicted 0 |
| Actual 1 | 76 | 21 |
| Actual 0 | 3 | 174 |

correspond to ‘benign’ and ‘malignant’ cancers, respectively. 60% of the data was used for training, which was subsequently corrupted by flipping the labels of 40 class 1 datapoints to 0 ($\epsilon \approx 10\%$). The goal is to estimate a linear separating plane to predict the class labels of test data. We use the cross-entropy loss function $\ell_\theta(\mathbf{z})$ in (15) and apply the proposed RRM method with $\tilde{\epsilon} = 0.15$. For comparison, we also use the standard ERM and SEVER methods.

Tables 1 for ERM, 2 for SEVER and 3 for RRM summarize the results using the confusion matrix as the metric. The classification accuracy for the RRM method is visibly higher than that of ERM and SEVER for class 1.

VI. CONCLUSION

We proposed a general risk minimization approach which provides robustness in a wide range of statistical learning problems in cases where a fraction of the observed data comes from a corrupting distribution. Unlike existing general robust methods, our approach does not depend on any problem-specific thresholding techniques to remove the corrupted data points, nor does it rely on an exactly specified corruption fraction ϵ . We illustrated the wide applicability and performance of our method by testing it on several classical supervised and unsupervised statistical learning problems using both simulated and real data.

REFERENCES

- [1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.
- [2] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 284–293.
- [3] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015, *arXiv:1412.6572v3*.
- [4] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," 2019, *arXiv:1708.06733v2*.
- [5] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust Statistics for Signal Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2018.
- [6] S. A. Kassam and V. H. Poor, "Robust techniques for signal processing: A survey," *Proc. IEEE*, vol. 73, no. 3, pp. 433–481, Mar. 1985.
- [7] B. Biggio, G. Fumera, and F. Roli, "Security evaluation of pattern classifiers under attack," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 4, pp. 984–996, Apr. 2014.
- [8] J.-G. Hsieh, Y.-L. Lin, and J.-H. Jeng, "Preliminary study on Wilcoxon learning machines," *IEEE Trans. Neural Netw.*, vol. 19, no. 2, pp. 201–211, Feb. 2008.
- [9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.
- [10] X. Zhang, L. Zhao, A. P. Boedihardjo, and C.-T. Lu, "Online and distributed robust regressions under adversarial data corruption," in *Proc. IEEE Int. Conf. Data Mining*, 2017, pp. 625–634.
- [11] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in Statistics*, Berlin, Germany: Springer, 1992, pp. 492–518.
- [12] P. J. Huber, *Robust Statistics*. Berlin, Germany: Springer, 2011.
- [13] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust Statistics: Theory and Methods (with R)*. Hoboken, NJ, USA: Wiley, 2019.
- [14] J. Tukey, *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley, 1977.
- [15] A. R. Klivans, P. M. Long, and R. A. Servedio, "Learning halfspaces with malicious noise," *J. Mach. Learn. Res.*, vol. 10, pp. 2715–2740, 2009.
- [16] K. Bhatia, P. Jain, and P. Kar, "Robust regression via hard thresholding," *Adv. Neural Inf. Process. Syst.*, pp. 721–729, 2015.
- [17] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar, "Consistent robust regression," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 2110–2119.
- [18] P. Awasthi, M. F. Balcan, and P. M. Long, "The power of localization for efficiently learning linear separators with noise," *J. ACM*, vol. 63, pp. 50:1–50:27, 2017.
- [19] A. Paudice, L. Muñoz-González, A. Gyorgy, and E. C. Lupu, "Detection of adversarial training examples in poisoning attacks through anomaly detection," 2018, *arXiv:1802.03041v1*.
- [20] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Being robust (in high dimensions) can be practical," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017.
- [21] I. Diakonikolas and D. M. Kane, "Recent advances in algorithmic high-dimensional robust statistics," 2019, *arXiv:1911.05911*.
- [22] T. Bernholt, "Robust estimators are hard to compute," Tech. Rep. 2005.52, 2006.
- [23] M. Hardt and A. Moitra, "Algorithms and hardness for robust subspace recovery," in *Proc. Conf. Learn. Theory*, 2013, pp. 354–375.
- [24] I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and A. Stewart, "Sever: A robust meta-algorithm for stochastic optimization," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 1596–1606.
- [25] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar, "Robust estimation via robust gradient estimation," 2018, *arXiv:1802.06485*.
- [26] M. Charikar, J. Steinhardt, and G. Valiant, "Learning from untrusted data," in *Proc. 49th Annu. ACM SIGACT Symp. Theory Comput.*, 2017, pp. 47–60.
- [27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.
- [28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [29] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar. 2014. [Online]. Available: <http://cvxr.com/cvx>
- [30] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss–Seidel method under convex constraints," *Oper. Res. Lett.*, vol. 26, no. 3, pp. 127–136, 2000.
- [31] UCI, "Breast cancer wisconsin UCI repository," [Online]. Available: [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(original\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original))
- [32] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. Hoboken, NJ, USA: Wiley, Sep. 2011.
- [33] Z. M. Shibzukhov, *On the Principle of Empirical Risk Minimization Based on Averaging Aggregation Functions*. Berlin, Germany: Springer, 2017.
- [34] T. Li, A. Beirami, M. Sanjabi, and V. Smith, "Tilted empirical risk minimization," 2020, *arXiv:2007.01162v1*.