Pattern Recognition Letters 112 (2018) 317–323

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

On the exact minimization of saturated loss functions for robust regression and subspace estimation

Fabien Lauer*

Université de Lorraine, CNRS, LORIA, Nancy, F-54000, France

ARTICLE INFO

Article history:
Received 17 January 2017
Available online 6 August 2018

ABSTRACT

This paper deals with robust regression and subspace estimation and more precisely with the problem of minimizing a saturated loss function. In particular, we focus on computational complexity issues and show that an exact algorithm with polynomial time-complexity with respect to the number of data can be devised for robust regression and subspace estimation. This result is obtained by adopting a classification point of view and relating the problems to the search for a linear model that can approximate the maximal number of points with a given error. Approximate variants of the algorithms based on ramdom sampling are also discussed and experiments show that they offer an accuracy gain over the traditional RANSAC for a similar algorithmic simplicity.

1. Introduction

Robust estimation is a classical problem raised by the presence of outliers in the data. Such outliers are points that do not coincide with the underlying data distribution being learned and that must be rejected in order to estimate an accurate model. A standard approach, entering the statistical framework of redescending M-estimators [13,14], relies on the minimization of a saturated loss function. Indeed, this saturation ensures that outliers yielding gross errors have a very limited influence on the estimation as the gradient of the loss at these points is zero. However, saturated loss functions are inherently nonconvex and their minimization is a highly nontrivial task. For some applications, suboptimal solutions or other heuristics such as the RANdom SAmple Consensus (RANSAC) [6] can provide satisfactory models. Yet, robust estimation problems also appear for instance iteratively in a bounded-error framework for problems where the data is assumed to be generated by a collection of models with unknown assignments of the data points to the models, such as in switching linear regression [1,4,9,10] or subspace clustering [2,12,15]. In such applications, the models are often estimated one by one while considering the data assigned to other models as outliers. In this context, relying on suboptimal solutions can lead to highly unsatisfactory results with many misclassifications of data points. Robust methods based on convex relaxations [2,3,12] or iteratively hard-thresholding [5] offer some guarantees but are only optimal under particular conditions on the data.

Instead, in this paper, we aim at unconditional optimality and discuss the computational complexity of globally minimizing a saturated loss function for the robust estimation of linear models, let it be regression ones or subspaces. In particular, the paper focuses on the question of the existence of an algorithm with a polynomial time-complexity with respect to the number of data, N. To this end, we devise an algorithm by enumerating all classifications of the points into two categories: those for which saturation of the loss occurs and those for which it does not. This classification point of view is also motivated by the equivalent formulation of the problem as the maximization of the number of points approximated by a linear model with a bounded error combined with the minimization of a standard (non-saturated) loss over these points only. Indeed, this leads to the distinction between points with error less than and greater than a predefined threshold. Since there are 2^N binary classifications of N points, such a combinatorial approach based on the enumeration of all of them yields an algorithm with exponential complexity in $\mathcal{O}(2^N)$. Yet, we adopt its classification viewpoint and show that the number of classifications, and thus the complexity, can be reduced to a polynomial function of N. From this classification viewpoint, the minimization of a saturated squared loss for regression can be related to the least trimmed squares estimator [13], for which exact algorithms with polynomial complexity wrt. N have been proposed in [7,11]. However, these are restricted to problems with a single variable (one-dimensional data) and work with a fixed number of inliers rather than an error threshold.

* Corresponding author.
E-mail address: fabien.lauer@loria.fr

https://doi.org/10.1016/j.patrec.2018.08.004
0167-8655/

While a polynomial complexity appears convenient, the degree of the polynomials can limit the applicability of the exact algorithms. Therefore, we also briefly discuss approximate variants of the algorithms devised to leverage the computational load by avoiding the complete enumeration of the classifications through random sampling.

Notation. We write vectors in lowercase bold letters and matrices in uppercase bold letters. We define $\text{sign}(u)$ as taking value $+1$ if and only if $u \geq 0$ and -1 otherwise. $\text{sign}_0(u)$ is defined similarly except that $\text{sign}_0(0) = 0$. The indicator function $\mathbf{1}_A$ is 1 when the Boolean expression A is true and 0 otherwise.

Paper organization. Section 2 gives the precise formulations of the regression and subspace estimation problems we consider. Then, Section 3 shows how these can be solved in polynomial time with respect to N. Section 4 discusses the approximate variants of the algorithms and Section 5 provides a few numerical results. Finally, Section 6 gives concluding remarks.

2. Problem formulation

In general terms, in an estimation problem, one can fit a model to the data by minimizing a loss function of the error between the model output and the data.[1] For instance, standard loss functions include the ℓ_p-losses defined for $p \geq 0$ and all values of the error $e \in \mathbb{R}$ as

$$\ell_p(e) = \begin{cases} \mathbf{1}_{|e|>0}, & \text{if } p = 0 \\ |e|^p, & \text{if } p \in (0, +\infty). \end{cases} \tag{1}$$

Here, we concentrate on robust estimation in the presence of outliers and formulate the problem in terms of a *saturated* loss function $\ell_{p,\epsilon} : \mathbb{R} \to \mathbb{R}^+$, defined for $p \in \{0, 1, 2\}$ by

$$\forall \epsilon > 0, \quad \ell_{p,\epsilon}(e) = \begin{cases} \mathbf{1}_{|e|>\epsilon}, & \text{if } p = 0 \\ (\min(|e|, \epsilon))^p, & \text{if } p \in \{1, 2\}. \end{cases} \tag{2}$$

Indeed, saturating the loss function limits the influence of outliers in the overall cost function to be minimized and thus on the resulting estimate. The statistical properties of these types of loss functions have been studied in the framework of redescending M-estimators, see e.g., [13]. For $p = 0$, this approach is also related to bounded-error estimation. Indeed, we can equivalently view it as the maximization of the number of points for which the error is small and below the threshold ϵ. For $p > 0$, a similar viewpoint can be taken with the additional feature that the small errors are measured by a standard ℓ_p-loss function and further minimized.

In this paper, we will focus the discussion on the corresponding optimization problem whose difficulty comes from the nonconvexity of the saturated losses.

The computation of the argument e as a function of the model parameters and the precise form of the optimization problem depends on the specific problem considered and will be detailed next for regression and subspace estimation.

2.1. Robust regression via saturated loss minimization

The aim of linear regression is to estimate a linear model $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$ from a data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ of regression vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ and target outputs $y_i \in \mathbb{R}$. Here, we adopt an error-minimizing approach and more precisely focus on saturated loss functions as defined above in order to confine the influence of outliers on the global cost. Let us define the index sets $I = \{1, \dots, N\}$ and

$$I_1(\boldsymbol{w}) = \{i \in I : |y_i - \boldsymbol{w}^T\boldsymbol{x}_i| < \epsilon\}, \tag{3}$$

before formally stating the robust regression problem we consider.

Problem 1 ($\ell_{p,\epsilon}$-linear regression)**.** Given a data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ and a threshold $\epsilon > 0$, find a global solution to

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} J_p(\boldsymbol{w}), \tag{4}$$

where

$$\begin{aligned} J_p(\boldsymbol{w}) &= \sum_{i=1}^N \ell_{p,\epsilon}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i) \\ &= \begin{cases} N - |I_1(\boldsymbol{w})|, & \text{if } p = 0 \\ \displaystyle\sum_{i \in I_1(\boldsymbol{w})} |y_i - \boldsymbol{w}^T\boldsymbol{x}_i|^p + \epsilon^p(N - |I_1(\boldsymbol{w})|), & \text{if } p \in \{1, 2\}. \end{cases} \end{aligned} \tag{5}$$

The formulation of Problem 1 emphasizes the connection between saturated loss minimization and bounded-error estimation, i.e., the maximization of the number of points approximated with a bounded error that are here marked with index in $I_1(\boldsymbol{w})$.

This also draws a connection with the classification problem of separating between points that are approximated with a bounded error by an optimal model and those that are not. In particular, given the solution to this classification problem, i.e., $I_1(\boldsymbol{w}^*)$ for some global minimizer $\boldsymbol{w}^*$ of $J_p(\boldsymbol{w})$, a (perhaps different[2]) global solution $\hat{\boldsymbol{w}}$ can be recovered by solving Problem 1 under the constraint $I_1(\boldsymbol{w}) = I_1(\boldsymbol{w}^*)$. Then, for $p = 0$, $J_p(\boldsymbol{w})$ is a mere constant and it suffices to find a $\boldsymbol{w}$ such that $|y_i - \boldsymbol{w}^T\boldsymbol{x}_i| < \epsilon$ for all $i \in I_1(\boldsymbol{w}^*)$ to satisfy the constraint. Conversely, for other values of p, the cost function $J_p(\boldsymbol{w})$ simplifies to a sum of error terms over a fixed set of points plus a constant. Hence, its minimization amounts to a standard regression problem with a non-saturated loss and we can compute $\hat{\boldsymbol{w}}$ by solving

$$\hat{\boldsymbol{w}} \in \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \begin{cases} \displaystyle\max_{i \in I_1(\boldsymbol{w}^*)} |y_i - \boldsymbol{w}^T\boldsymbol{x}_i|, & \text{if } p = 0 \\ \displaystyle\sum_{i \in I_1(\boldsymbol{w}^*)} |y_i - \boldsymbol{w}^T\boldsymbol{x}_i|^p, & \text{otherwise.} \end{cases} \tag{6}$$

Such standard problems have polynomial complexities in $\mathcal{O}(d^2N)$ for $p = 2$ and $\mathcal{O}(d^4N^4)$ for $p \in \{0, 1\}$.

2.2. Robust subspace estimation via saturated loss minimization

A d_s-dimensional subspace of $\mathbb{R}^d$ can be thought of as the column space of a $d \times d_s$ matrix $\boldsymbol{B}$ with orthonormal columns. In this case, the projection of a vector $\boldsymbol{x} \in \mathbb{R}^d$ onto the subspace can be written as $\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{x}$ and the corresponding scalar approximation error as $\|(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}^T)\boldsymbol{x}\|$.

Therefore, subspace estimation from a data set $\{\boldsymbol{x}_i\}_{i=1}^N$ with a fixed subspace dimension equal to d_s can be set as the search for a matrix $\boldsymbol{B} \in \mathbb{R}^{d \times d_s}$ such that $\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I}$ and that the approximation error is minimized over the data set. In the presence of outliers, a robust estimation can be obtained from the minimization of a saturated loss function (as defined in (2)) of this approximation error.

For any $\boldsymbol{B} \in \mathbb{R}^{d \times d_s}$, we define the index set

$$I_1(\boldsymbol{B}) = \{i \in I : \|(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}^T)\boldsymbol{x}_i\| < \epsilon\}, \tag{7}$$

in order to state the problem of robust subspace estimation as follows.

[1] Note that we focus on problems where the dimensionality is significantly smaller than the number of data and where regularization of linear models might not be necessary. However, given the nature of the proposed approach, introducing a convex regularizer should not raise difficulties.

[2] Problem 1 may have multiple global solutions, especially when $p = 0$.