Neurocomputing 429 (2021) 174–186

Contents lists available at ScienceDirect

Neurocomputing

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

Robust kernels for robust location estimation

Joseph A. Gallego [a], Fabio A. González [a,*], Olfa Nasraoui [b]

[a] MindLab Research Group, Computing Systems and Industrial Engineering, Universidad Nacional de Colombia, Bogotá, Colombia
[b] Knowledge Discovery and Web Mining Lab, CECS Department, University of Louisville, Louisville, KY 40292, United States

ARTICLE INFO

Article history:
Received 10 April 2020
Revised 1 September 2020
Accepted 25 October 2020
Available online 21 November 2020
Communicated by Steven Hoi

Keywords:
Robust statistics
M-estimators
Kernel methods
Kernel clustering
Kernel matrix factorization

ABSTRACT

This paper shows that least-square estimation (mean calculation) in a reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ corresponds to different M-estimators in the original space depending on the kernel function associated with $\mathcal{F}$. In particular, we present a proof of the correspondence of mean estimation in an RKHS for the Gaussian kernel with robust estimation in the original space performed with the Welsch M-estimator. This result is generalized to other types of M-estimators. This generalization facilitates the definition of new robust kernels associated to Huber, Tukey, Cauchy and Andrews M-estimators. The new kernels are empirically evaluated in different clustering tasks where state-of-the-art robust clustering methods are compared to kernel-based clustering using robust kernels. The results show that some robust kernels perform on a par with the best state-of-the-art robust clustering methods.

1. Introduction

In the context of statistical estimation, robustness refers to the ability of a method to deal with contamination, i.e. outliers, noise, and, in general, departures from model assumptions. Classical methods suffer from problems such as masking effect (a method does not detect outliers or deviating points) and swapping (good data points seem like outliers). A robust method needs to safeguard against deviations from the assumptions and identify highly influential data points [1]. Robust statistics study the development of robust methods, i.e. reasonably efficient methods in the neighborhood of the assumed statistical model [2]. Examples of robust estimation methods include various robust extensions of the Maximum Likelihood Estimation (MLE) for Gaussian and other known distributions, such as the ϵ-contamination model, M-estimators, and robust clustering [3,4].

Robustness is also an important issue in machine learning, and various efforts have been done to design robust versions of unsupervised and supervised methods. Some examples include: different robust versions of principal component analysis (PCA), where robustness is introduced by improving the computation of the covariance matrix using a robust scale function or by centering the data around the L_1-median [5–9]; robust classification where loss functions are modified using, in some cases, an M-estimator [10–15]; and, robust clustering where robustness relies on using robust statistics techniques such as trimmed mean [16–20]. In the particular case of kernel methods, there are few works that deal with robustness, some examples include: robust kernel density estimation where robustness depends on changing the kernelized loss function with a M-estimator function [21] and robust support vector machines where robustness relies on changing the Euclidean distance with a more robust function such as an M-estimators [22].

The goal of this paper is to study the theoretical and empirical robustness of kernel-based algorithms within the framework of robust statistical estimation and, as a followup, to use this framework to design new kernels that can deal with noise and outliers, thus qualifying as robust kernels. In particular, we show that a classic kernel such as the Gaussian Kernel has intrinsically robustness built in. Additionally, the paper extends this result to new kernels that are derived form known M-estimators. M-estimators are a class of estimators obtained by the maximization of a loss function ρ, calculated over the data. Depending on the particular function ρ the M-estimator may be more or less robust.

In pursuing the above goal, this paper presents: (1) a unified view of two families of methods (robust estimation and kernel-based methods) that have had an immense impact on data analysis and machine learning; (2) methods which can be shown to have some classical statistical robustness mechanisms naturally built-in, although they were not conceived to be robust; (3) a new

* Corresponding author at. Universidad Nacional de Colombia, Computing Systems and Industrial Engineering, Of 101 Edif 453, Universidad Nacional de Colombia, Ciudad Universitaria, Bogota, Colombia.
E-mail addresses: jagallegom@unal.edu.co (J.A. Gallego), fagonzalezo@unal.edu.co (F.A. González), olfa.nasraoui@louisville.edu (O. Nasraoui).

https://doi.org/10.1016/j.neucom.2020.10.090
0925-2312/

framework for building new robust kernels; (4) four new robust kernels associated with M-estimators; (5) theoretical discussion of each new robust kernel; and (6) an empirical systematic comparison between new robust kernels and state-of-the-art methods in the context of unsupervised learning.

One interesting result is showing that least-square estimation (mean calculation) in a reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ corresponds to different M-estimators in the original space depending on the kernel function associated with $\mathcal{F}$. This new finding opened an avenue to extend kernel based learning by proposing four new robust kernels associated to Huber, Tukey, Cauchy and Andrews M-estimators. Our evaluations further show that these new robust kernels exhibit good performance compared to state-of-the-art methods.

Kernel-based methods such as kernel-based clustering, Gaussian Processes, and support vector machines have been significant players in machine learning. Therefore, our findings can have a significant impact, in particular on studying the theoretical robustness properties of kernel based machine learning methods and for designing extended robust kernel based learning algorithms with desired robustness properties. This paper shows that using robust statistics, kernel methods can be improved in clustering tasks making them competitive when compared to state-of-the-art algorithms.

The rest of this paper is organized as follows. Section 2 reviews robust statistical estimators and related work. Section 3 presents a formal proof that mean estimation in the feature space with a Gaussian kernel is equivalent to robust mean estimation with the Welsch M-estimator in the data space. Section 4 presents the definition and theoretical discussions of four new kernels using the ideas presented in Section 3. Section 5 presents an empirical comparison between several state-of-the-art clustering algorithms and kernel-based clustering using the new kernels presented in Section 4. Finally, Section 6 presents our conclusion and future work.

2. Background on M estimators and kernels

2.1. M-estimators and robust statistics

Robust statistics emerged as a family of theories and techniques for estimating the parameters of a parametric model while dealing with deviations from idealized assumptions [23,24,4,25]. Examples of deviations include contamination of data by gross errors, rounding and grouping errors, and departure from an assumed sample distribution. Gross errors or outliers are data severely deviating from the pattern set by the majority of the data. This type of error usually occurs due to mistakes in copying or computation. They can also be due to part of the data not fitting the same model, as in the case of data with multiple clusters. Gross errors are often the most dangerous type of errors. In fact, a single outlier can completely spoil the Least Squares estimate, causing it to break down. Rounding and grouping errors result from the inherent inaccuracy in the collection and recording of data which is usually rounded, grouped, or even coarsely classified. The most common and practical robust estimators are M and W-estimators. Other estimators include L, Least Trimmed Squares, and Reweighted Least Squares estimators. Below, we review several M and W-estimators [4,26].

The ordinary Least Squares (LS) method to estimate parameters is not robust because its objective function, $\sum_{j=1}^{N} x_j^2$, increases indefinitely with the residuals x_j between the j^{th} data point and the estimated fit, with N being the total number of data points in a data set. Hence, extreme outliers with arbitrarily large residuals can have an infinitely large influence on the resulting estimate. M-estimators [24] attempt to limit the influence of outliers by replacing the square of the residuals with a less rapidly increasing loss function of the data value, x, and parameter estimate, t, $\rho(x;t)$. This function is usually called contrast function. The M-estimator, $T(x_1,\cdots,x_N)$ for the function ρ and the sample $x_1,\cdots,x_N$, is the value that minimizes the following objective

$$T = \min_t \quad \{J = \sum_{j=1}^{N} \rho(x_j;t)\}. \tag{2.1}$$

The optimal parameter, T, is determined by solving

$$\frac{\partial J}{\partial t} = \sum_{j=1}^{N} \psi(x_j;t) = 0 \tag{2.2}$$

where, except for a multiplicative constant,

$$\psi(x_j;t) = \frac{\partial \rho(x_j;t)}{\partial t}. \tag{2.3}$$

When the M-estimator is equivariant, i. e., $T(x_1+a,\cdots, x_N+a) = T(x_1,\cdots,x_N)+a$ for any real constant a, we can write ψ and ρ in terms of the residuals $x-t$. Also, in general, an auxiliary scale estimate, S is used to obtain the scaled residuals $r = \frac{x-t}{S}$. Hence, we can write

$$\psi(r) = \psi\left(\frac{x-t}{S}\right) = \psi(x;t),$$

and

$$\rho(r) = \rho\left(\frac{x-t}{S}\right) = \rho(x;t).$$

The M-estimator can be written as a weighted average of the samples:

$$T = \sum_{j=1}^{n} w(x_j;T)x_j$$

where the weight function is defined as:

$$w(x_j;t) = \begin{cases} \frac{\psi(x_j;t)}{x_j}, & \text{if } x_j \neq 0 \\ \psi(0), & \text{if } x_j = 0 \end{cases} \tag{2.4}$$

The ρ-functions for some familiar M-estimators are listed in Table 1. Note that LS can be considered an M-estimator, even though it is not a *robust* M-estimator. As seen in this table, M-estimators rely on both an accurate estimate of scale and a fixed tuning constant, c. Most M-estimators use a multiple of the Median of Absolute Deviations (MAD) as a scale estimate which implicitly assumes that the noise contamination rate is 50%. MAD is defined as follows:

$$MAD(x_i) = med_i\{|x_i - med_j(x_j)|\}$$

The most common scale estimate used is $1.483 \times MAD$ where the 1.483 factor adjusts the scale for maximum efficiency when the data samples come from a Gaussian distribution. Fig. 1.

2.2. Robust machine learning

Robustness to outliers, noisy samples, and heavy-tailed distributions is an important issue for machine learning methods. However, this is not necessarily an issue which is directly addressed when developing new methods. Nevertheless, there are some works that apply robust statistical concepts and methods to machine learning techniques.

In some works, robust statistical techniques, such as influence curves and breaking point analysis, have been applied to analyze the robustness of some machine learning methods. Kim et al. [21] propose a novel robust kernel density estimation method, where robustness depends on changing the kernelized loss

Table 1
Different familiar M-estimators with $\rho(r), \psi(r)$ and $w(r)$ functions. Some functions are defined piece-wise, so the scale parameter c defines the breakpoint where the principal residual function is replaced with a less rapidly increasing loss function the *range* of r defines the range of residual r for the $\rho(r)$ function is defined.

Type	$\rho(r)$	$\psi(r)$	w(r)	Range of r
L_2 (mean)	$\frac{1}{2}r^2$	r	1	IR
L_1 (median)	$\lvert r\rvert$	sgn(r)	$\frac{sgn(r)}{r}$	IR
Huber	$\frac{1}{2}r^2$	r	1	$\lvert r\rvert \leqslant c$
	$c\lvert r\rvert - \frac{1}{2}c^2$	c sgn(r)	$\frac{csgn(r)}{r}$	$\lvert r\rvert > c$
Cauchy	$\frac{c^2}{2} log\left[1 + (\frac{r}{c})^2\right]$	$\frac{r}{1+(\frac{r}{c})^2}$	$\frac{1}{1+(\frac{r}{c})^2}$	IR
Welsch	$\frac{c^2}{2}\left[1 - exp(-(\frac{r}{c})^2)\right]$	$r\ exp(-(\frac{r}{c})^2)$	$exp(-(\frac{r}{c})^2)$	IR
Tukey's	$1 - \left[1 - (\frac{r}{c})^2\right]^3$	$r(1 - (\frac{r}{c})^2)^2$	$\left[1 - (\frac{r}{c})^2\right]^2$	$\lvert r\rvert \leqslant c$
Biweight	1	0	0	$\lvert r\rvert > c$
Andrews	$c\left[1 - cos(\frac{r}{c})\right]$	$sin(\frac{r}{c})$	$sin(\frac{r}{c})/r$	$\lvert r\rvert \leqslant c\pi$
	$2c$	0	0	$\lvert r\rvert > c\pi$

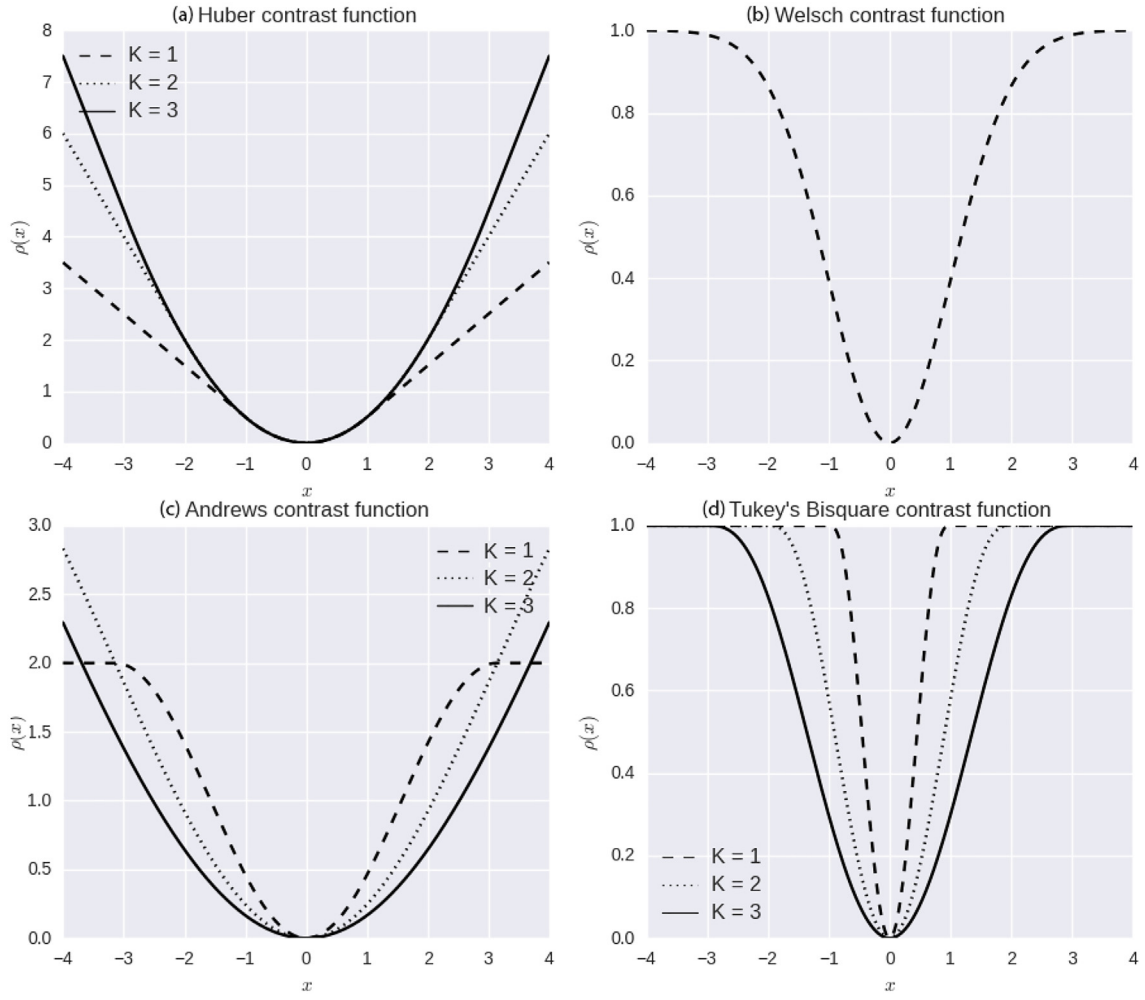

Fig. 1. Different examples of robust M-contrast functions using $c := 1$: (a) Huber contrast function for different k parameters, note that parameter k is chosen arbitrary and modifying the weight of the residuals , (b) Welsch contrast function, (c) Andrews contrast function for different k parameters, note that a higher value of k decrease the power of the residual in the estimation of ρ, (d) Tukey's Bisquare contrast function for different k parameters 2.1.

function with an M-estimator function, which showed good behavior for different datasets according to the influence curve and the breakdown point. Among the robust statistical techniques, the most widely used are M-estimators, see Section 2, which has been applied to robust regression, robust estimation, and clustering, among other tasks [21,22,27,28].

In the case of PCA [5], the first efforts to robustify it were based on finding a robust estimate of the covariance matrix. One of the drawbacks of this approach is its high computational demand [29]. An important step for applying PCA is to previously center the data. One approach to robustify this step is to use a robust location estimator such as the median[9]. In the case of the use of the

sum of squared errors cost function as an optimization function, some authors have used regularization of parameters and weighting errors to make the optimization more robust against outliers [30]. Reformulating the reconstruction function has been one of the ways to robustify PCA. In [31], de la Torre et al. showed that the loss function can be changed by a related Geman-McLure M-estimator. The idea is to define the reconstruction error as $\arg\min_{Z,\mu,\sigma}\rho(X-\mu-Z,\sigma)$ where σ is a scale parameter of the Geman-MClure function, defined as $\rho(x;\sigma)=\frac{x^2}{x^2+\sigma^2}$. In this case, σ is responsible for the outlier proportion. In [32], Huang and Yeh proposed an iterative kernel principal component analysis (KPCA) using a relaxed optimization function. This new optimization function enables the use of m-estimators such as the Geman-McClure m-estimator. This approach shows good performance and convergence when contamination is added to real datasets. In [33], Svensén and Bishop proposed a new approach to Bayesian mixture modeling. They used a heavy-tailed t-student distribution providing a more robust algorithm when the data has outliers. However, there is no relationship with robust estimation.

In the case of *k*-Means, which generally uses the Euclidean distance, the centroid estimation may be biased by outliers [34]. In this sense, several strategies have been proposed in order to robustify it. One approach is to create a new cluster where all the outliers are added so that every point in the data set will be equidistant from that cluster. Other strategies use fuzzy membership functions instead of the minimum Euclidean distance-based hard membership assignments [35,15,19]. Krishnapuram et al. [36] use possibilistic memberships in a version known as Possibilistic c-means. Del barrio et al. [37] propose to use the Wasserstein space to obtain the trimmed *k*-barycenters that enable parallel computation. The method proposed by Zhou et al. [38] estimates a local density using non-parametric density estimation and assign each point to the nearest neighbor with higher density.

Other problems, such as manifold learning and matrix factorization has been addressed with robust methods. A good example is Robust Manifold Non-Negative Matrix Factorization (RMNMF) [39] which uses the norm $l_{2,1}$ instead of the Frobenius' norm. The objective function of RMNMF, measures the difference between matrix X and factorization FG^T in a robust way. The authors propose a regularization with the Laplacian graph, to obtain a spectral clustering. Another example of non-negative matrix factorization method claiming to be robust is non-negative matrix factorization random walks (NMFR) [40], which uses the random walks notion in order to improve clustering results in spectral clustering.

2.3. Kernels and kernels methods

In the case of kernel methods, some works involved robust estimation. For example, in [41], the authors define a new way to center the data in the feature space. They use the L_1 norm in order to center the data in a robust way. Instead of removing the outliers in the input space, the outliers can be deleted in the feature space. This is achieved through the reconstruction error which can be found in the feature space and by doing so, the points that have a large deviation with respect to the normal values can be identified. The authors claim that in the case of Kernel PCA, the kernel chosen does not have any significance [42]. In [17], Chen et al. propose a new optimization function for robust kernel c-means where the robustness depends on a kernelized version of probabilistic fuzzy clustering. The optimization problem involves a multi-step solution where the Gaussian kernel is used.

Only a few works are found that linked M-estimators with kernels. Chen [22] proposed a new kernel for support vector machines. Basically, the Euclidean distance between the samples is exchanged for a more robust distance given by the M-estimator defined as $\sum_{j=1}^{d}\rho(|x_j-x'_j|,\gamma)$. Liao et al. [27] proposed a robust kernel inspired by a robust M-estimator to achieve robust machine learning.

The previously discussed papers [27,42] showed a robust behavior of some kernel methods. However, they did not attempt to explain why this happens. In this paper, we show that there is an intrinsic relationship between some types of special kernels and M-estimators. This relationship is not only interesting for its own sake, but it further builds a new more general framework that allows to create new robust kernels, which, to our knowledge, had not been previously discovered.

In [43], Liang et al. proposed a method called robust linear discriminant analysis for dimensionality reduction where the squared of the projection distance using a kernel is maximized. The word robust in the name is related to the use of kernel embedding. In [44], Kang et al. proposed a new method for sparse similarity learning. This method used the addition of positive definite kernels to solve a sparse representation of the original matrix. Besides in [45], Kang et al. proposed a clustering method using the intrinsic structure of a learning graph. This method shows good performance in state-of-the-art datasets. However, none of the above results methods are related to robust estimation.

3. The robustness of kernel estimation

In [46], we showed, empirically, that the use of a Gaussian kernel makes clustering techniques more robust to outliers compared to a linear kernel. These results showed that when the contamination is increased, the linear kernel had a higher bias measure between the location estimates (centroids) and the real parameters compared to the results when using the Gaussian kernel. Taking this empirical result into account, in this section, we formally establish the intrinsic relationship that exists between kernels and robust M-estimators. In particular, we propose a formal proof that doing mean estimation in a feature space, induced by some kind of kernel, is equivalent to doing robust mean estimation in the original space. The overall process for doing robust location estimation using a kernel is as follows: first, we start with data represented in the original problem space; second, a kernel function is used to implicitly map every data point in the data space to an induced feature space; third, the centroid of the images of the data points is calculated in the feature space (note that in the feature space, the centroid is not necessary robust); finally, with the help of an approximate inverse function P_Φ, the centroid in the feature space is mapped back to the original problem space. We will show that the inverse image of the centroid in the feature space is a robust estimator of the mean of the original data points. In particular, it corresponds to the Welsch-estimator which is a robust M-estimator. This relationship is proved in Proposition 2.

3.1. Approximate pre-image definition

Let $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a kernel (not necessarily positive semidefinite) and let $\Phi:\mathcal{X}\to\mathcal{F}$ be a mapping from the original space to the corresponding reproducing kernel Hilbert space (RKHS), i.e, $k(x,y)=\langle\phi(x),\phi(y)\rangle_F$ [47]. Since Φ is not onto, in general, there are elements $\phi\in\mathcal{F}$ which do not have a pre-image, i.e., $\nexists x\in\mathcal{X},\Phi(x)=\phi$. The next definition is motivated by this fact.

Definition 1. An Approximate pre-image P_Φ is defined as

$$\begin{aligned} P_\Phi:\ \mathcal{F} &\to \mathcal{P}(\mathcal{X}) \\ \phi &\mapsto \operatorname*{argmin}_{x\in X}||\Phi(x)-\phi||^2_{\mathcal{F}} \end{aligned} \tag{3.1}$$

where $||f||_{\mathcal{F}}^2 = \langle f, f \rangle_{\mathcal{F}}$ is the squared norm in the corresponding Hilbert space.

3.2. Kernel robust estimation

In the following propositions, we show that if we find the centroid's pre-image–from a set of points projected in a feature space, then these centroids will correspond to robust location estimators if the appropriate kernel is used [48]. In this case, *appropriate* means that the kernel is isotropic. An isotropic kernel is a kernel that only depends on the distance, in the input space between its arguments, i.e., $k(x, y) = g(||x - y||)$ [49]. We will indistinctly use the notation $k(||x - y||)$ and $k(x, y)$ to refer to the application of the kernel, hoping the meaning will be clear from the context.

Proposition 1. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be an isotropic kernel with $\Phi : \mathcal{X} \to \mathcal{F}$ the associated mapping to the induced feature space and $\{x_1, \cdots, x_n\} \subseteq \mathcal{X}$ a set of samples, then

$$P_\Phi(\mu) = \underset{y \in \mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} ||\Phi(x_i) - \Phi(y)||_{\mathcal{F}}^2, \text{ with } \mu = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)$$

Proof. By Definition 1,

$$\begin{aligned}
P_\Phi(\mu) &= \underset{y \in \mathcal{X}}{\operatorname{argmin}} ||\Phi(y) - \mu||_{\mathcal{F}}^2 \\
&= \underset{y \in \mathcal{X}}{\operatorname{argmin}} ||\Phi(y) - \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)||_{\mathcal{F}}^2 \\
&= \underset{y \in \mathcal{X}}{\operatorname{argmin}} k(||y - y||) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(||x_i - x_j||) - \frac{2}{n}\sum_{i=1}^{n} k(||y - x_i||) \\
&= \underset{y \in \mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} k(||y - y||) + \sum_{i=1}^{n} k(||x_i - x_i||) - 2\sum_{i=1}^{n} k(||y - x_i||) \\
&= \underset{y \in \mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} (k(||y - y||) + k(||x_i - x_i||) - 2k(||y - x_i||)) \\
&= \underset{y \in \mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} (\langle \Phi(y), \Phi(y) \rangle_{\mathcal{F}} + \langle \Phi(x_i), \Phi(x_i) \rangle_{\mathcal{F}} - 2\langle \Phi(y), \Phi(x_i) \rangle_{\mathcal{F}}) \\
&= \underset{y \in \mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} ||\Phi(x_i) - \Phi(y)||_{\mathcal{F}}^2
\end{aligned}$$

The equality in the fourth line follows from the fact that the first and second terms on the right side of the third equality do not depend on y, since $k(||y - y||) = k(0)$; hence they can be substituted by an arbitrary constant (an expression that does not depend on y). For the same reason, the $\frac{1}{n}$ coefficient of the third term of the right side of the third equality can be eliminated.

Proposition 2. *Given a set of points $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, the approximate pre-image of its centroid in a feature space, $\mathcal{F}$, induced by a Gaussian kernel, k, corresponds to the Welsch location M-estimator. In other words:*

$$P_\Phi(\mu) = P_\Phi\left(\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\right) = \arg\min_{y \in \mathcal{X}} \sum_{i=1}^{n} \rho_{\text{welsch}}(||x_i - y||)$$

Proof. Let $P_\Phi(\mu)$ be the approximate pre-image of μ, defined as in Proposition 1,

$$\begin{aligned}
P_\Phi(\mu) &= \underset{y \in \mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} ||\Phi(x_i) - \Phi(y)||_{\mathcal{F}}^2 \\
&= \underset{y \in \mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} (\langle \Phi(x_i), \Phi(x_i) \rangle_{\mathcal{F}} + \langle \Phi(y), \Phi(y) \rangle_{\mathcal{F}} - 2\langle \Phi(x_i), \Phi(y) \rangle_{\mathcal{F}}) \\
&= \underset{y \in \mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} (k(x_i, x_i) + k(y, y) - 2k(x_i, y)) \\
&= \underset{y \in \mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} (2 - 2e^{-\left(\frac{||x_i - y||^2}{2\sigma^2}\right)}) \\
&= \underset{y \in \mathcal{X}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_{\text{welsch}}(||x_i - y||)
\end{aligned}$$

with c = $\sqrt{2}\sigma$, where the first equality follows by the Proposition 1.

4. Robust kernels

Section 3 showed that location estimation in a feature space induced by a Gaussian kernel is equivalent to doing robust estimation in the original space using a robust Welsch estimator. Proposition 1 is a general result and it can be used as a framework to build new robust kernels. Consequently, we propose four new robust kernels, Tukey, Andrew, Cauchy and Huber kernels, which are motivated by their corresponding robust M-estimators.

Before presenting the proposed robust kernels, it is necessary to point out the following. In the case of isotropic and radial kernels, we will use the residual r indistinct of the euclidean distance $||x - y||$. Besides, the constant c is always positive $\mathbb{R}^+$. We present two definitions: compactly supported univariate function and conditional positive definite kernel [50].

Definition 2. $\Phi : \mathbb{R} \to \mathbb{R}$ is a compactly supported univariate function if $\exists p : \mathbb{R} \to \mathbb{R}$ such that:

$$\Phi(r) = \left\{ \begin{array}{cc} p(r), & \text{if} r \in [0, 1] \\ 0, & \text{o.c.} \end{array} \right\}$$

Definition 3. A kernel K is conditionally positive definite if and only if it satisfies:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geqslant 0$$

where $x_1, \cdots, x_n \in \mathbf{X}, n \geqslant 1, c_1, \cdots, c_n \in \mathbb{R}$ with $\sum_{i=1}^{n} c_i = 0$ and kernel K is a symmetric function, i.e. $K(x, y) = K(y, x)$.

4.1. Tukey robust kernel

Definition 4. The Tukey Robust kernel is defined as follows:

$$k_{\text{Tukey,c}}(r) = \left\{ \begin{array}{cc} \frac{1}{2}(1 - (\frac{r}{c})^2)^3, & \text{if} \frac{r}{c} \leq 1 \\ 0, & \text{if} \frac{r}{c} > 1 \end{array} \right\} \tag{4.1}$$

Proposition 3. *Given a set of points $S = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, the preimage of its centroid in a feature space, $\mathcal{F}$, induced by a Tukey kernel, k, corresponds to Tukey's Bisquare location M-estimator. In other words:*

$$P_\Phi(\mu) = P_\Phi\left(\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\right) = \arg\min_{y\in\mathcal{X}}\sum_{i=1}^{n}\rho_{\text{tukey}}(||x_i - y||)$$

Proof. Lets $P_\Phi(\mu)$ be the approximate pre-image of μ, defined as in Proposition 1,

$$\begin{aligned}
P_\Phi(\mu) &= \operatorname{argmin}_{y\in\mathcal{X}}\sum_{i=1}^{n}||\Phi(x_i) - \Phi(y)||^2_{\mathcal{F}} \\
&= \operatorname{argmin}_{y\in\mathcal{X}}\sum_{i=1}^{n}(\langle\Phi(x_i),\Phi(x_i)\rangle_{\mathcal{F}} + \langle\Phi(y),\Phi(y)\rangle_{\mathcal{F}} - 2\langle\Phi(x_i),\Phi(y)\rangle_{\mathcal{F}}) \\
&= \operatorname{argmin}_{y\in\mathcal{X}}\sum_{i=1}^{n}(k(x_i,x_i) + k(y,y) - 2k(x_i,y)) \\
&= \operatorname{argmin}_{y\in\mathcal{X}}\sum_{i=1}^{n}\left\{\begin{array}{cc} \frac{1}{2}+\frac{1}{2}+\frac{2}{2}(1-\left(\frac{||x_i-y||}{c}\right)^2)^3, & \text{if } \frac{||x_i-y||}{c}\le 1 \\ \frac{1}{2}+\frac{1}{2}+0, & \text{if } \frac{||x_i-y||}{c} > 1 \end{array}\right\} \\
&= \operatorname{argmin}_{y\in\mathcal{X}}\sum_{i=1}^{n}\left\{\begin{array}{cc} (1-(1-\left(\frac{||x_i-y||}{c}\right)^2)^3) & \text{if } \frac{||x_i-y||}{c}\le 1 \\ 1, & \text{if } \frac{||x_i-y||}{c} > 1 \end{array}\right\} \\
&= \operatorname{argmin}_{y\in\mathcal{X}}\sum_{i=1}^{n}\rho_{\text{tukey}}(||x_i - y||)
\end{aligned}$$

A desirable feature for a kernel is to be positive definite. This guarantees that there is a Hilbert space, the feature space, for which the kernel corresponds to its dot product [51]. Nevertheless, undefined kernels may be useful for different applications [47]. Tukey Robust kernel is not positive definite in $\mathbb{R}^s$ for every s because it is a radial function with compact support [52], however, it is strictly positive definite for some s as the following proposition show.

Proposition 4. *The Tukey Robust kernel is definitive positive in a space with dimensions less or equal to* $\mathbb{R}^5$.

Proof. Following Bernstein's representation theorem in monotonous functions [52], compactly supported univariate functions, as defined in 2, are not (conditional) positive definite in $\mathbb{R}^d$ for all $d \geq 1$ [53].

The function

$$\phi(r) = (1-r^2)^l_+ \quad l \in \mathbb{N} \tag{4.2}$$

where $(x)_+ = \{x, \text{ if } x \geqslant 0; 0 \text{ otherwise}\}$ is strictly positive definite and radial on $\mathbb{R}^{2l-1}$ as shown by Wu [54].

The multiplication by $1/2$ and the positive constant c in 4.2 are positive definite [55,51] so that we can remove them. If we replace 3 with l in Definition 4 and change the variables x and y with the residual r, we get the following equation:

$$k(r) = \left\{\begin{array}{cc} (1-r^2)^l, & \text{if } 1-r^2 \geqslant 0 \\ 0, & \text{if } 1-r^2 < 0 \end{array}\right\} \tag{4.3}$$

we can deduce then that Tukey Robust kernel is conditional positive definite in a space with dimensions less or equal than R^5.

One of the remarkable findings is the relationship found between Tukey robust kernel and the Epanechnikov kernel defined by Ong Cheng in [47] as follows,

$$k(x_i,y) = \left\{\begin{array}{cc} \left(1-\left(\frac{||x_i-y||^2}{\sigma}\right)\right)^p, & \text{if } \frac{||x_i-y||}{\sigma}\le 1 \\ 0, & \text{if } \frac{||x_i-y||}{\sigma} > 1 \end{array}\right\}$$

where he showed that if principal eigenvalues are calculated, several of them will be negative.

4.2. Andrews robust kernel

Definition 5. The Andrews Robust kernel is defined as:

$$k_{\text{Andrews,c}}(r) = \left\{\begin{array}{cc} -\frac{c}{2}\left[1-\cos\left(\frac{r}{c}\right)\right], & \text{if } \frac{r}{c}\leqslant\pi \\ -c, & \text{if } \frac{r}{c} > \pi \end{array}\right\} \tag{4.4}$$

Proposition 5. *Given a set of points $S = \{x_1,\ldots,x_n\} \subseteq \mathcal{X}$, the preimage of its centroid in a feature space, $\mathcal{F}$, induced by an Andrews kernel, k, corresponds to the Andrews location M-estimator. In other words:*

$$P_\Phi(\mu) = P_\Phi\left(\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\right) = \arg\min_{y\in\mathcal{X}}\sum_{i=1}^{n}\rho_{\text{Andrews}}(||x_i - y||)$$

The Andrew's Kernel is not (conditional) positive definite as the following proposition shows.

Proposition 6. *The Andrews kernel is not (conditional) positive definite.*

Proof. Suppose that $k_{\text{Andrews}}(r)$ is (conditional) positive definite. Define

$$\phi(r) = k_{\text{Andrews,c}}(r) + c$$

A (conditional) positive kernel plus a positive constant is (conditional) positive definite [55,52]. Therefore, $\phi(r)$ is (conditional) positive definite. However, $\phi(r)$ has compact support and using the same argument as in Tukey Robust kernel, compactly supported univariated functions are not (conditional) positive definite in $\mathbb{R}^d$ for all $d \geqslant 1$.

4.3. Huber Robust Kernel

Definition 6. The Huber Robust kernel is defined as follows:

$$k_{\text{Huber,c}}(r) = \left\{\begin{array}{cc} -\frac{1}{4}r^2, & \text{if } \frac{r}{c}\le 1 \\ -\frac{c}{2}r + \frac{c^2}{4}, & \text{if } \frac{r}{c} > 1 \end{array}\right\} \tag{4.5}$$

Proposition 7. *Given a set of points $S = \{x_1,\ldots,x_n\} \subseteq \mathcal{X}$, the preimage of its centroid in a feature space, $\mathcal{F}$, induced by a Huber kernel, k, corresponds to the Huber location M-estimator. In other words:*

$$P_\Phi(\mu) = P_\Phi\left(\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\right) = \arg\min_{y\in\mathcal{X}}\sum_{i=1}^{n}\rho_{\text{Huber}}(||x_i - y||)$$

[52] proved that a non trivial positive definite function cannot have zeros for $\mathbb{R}^n$. Therefore, Huber robust kernel is not positive definite for $\mathbb{R}^n$. In [56], it is proved that $-||x-y||^{\mathcal{B}}$ for $0 \leqslant \mathcal{B} \leqslant 2$ is conditional positive kernel. Huber robust kernel is conditional positive definite and is proved in Proposition 9. First, We need to define a relationship between conditional positive definite and completely monotone on $(0,\infty)$:

Proposition 8 (*Michelli*). *Let $\varphi \in \mathbb{C}[0,\infty]$. $(-1)^m\varphi^{(m)}$ is completely monotonous, i.e., $(-1)^m\varphi^{(m)} \geqslant 0$ for $m = 1,2,3,\cdots$, if and only if $\Phi = \varphi(||\cdot||^2)$ is conditionally positive definite of order m and radial on $\mathbf{R}^n$ for all n [52].*

Proposition 9. *The Huber Robust kernel is conditional positive definite.*

Proof. Huber robust kernel can be written as:

$$\varphi_{\text{Huber,c}}(r^2) = \left\{ \begin{array}{ll} -r, & \text{if } r^{\frac{1}{2}} \leqslant c \\ -\frac{c}{2} r^{\frac{1}{2}} + \frac{c^2}{4}, & \text{if } r^{\frac{1}{2}} > c \end{array} \right\} \tag{4.6}$$

The first derivative of the Eq. 4.6 is positive and is defined as follows:

$$\varphi^{(1)}_{\text{Huber,c}}(r^2) = \left\{ \begin{array}{ll} -1, & \text{if } r^{\frac{1}{2}} \leqslant c \\ -\frac{c}{2} \frac{1}{2} r^{-\frac{1}{2}}, & \text{if } r^{\frac{1}{2}} > c \end{array} \right\} \tag{4.7}$$

In general, the $(-1)^m \varphi^{(m)}$ for all m greater than two of Huber kernel is positive and is defined as follows:

$$\varphi^{(m)}_{\text{Huber,c}}(r^2) = \left\{ \begin{array}{ll} 0, & \text{if } r^{\frac{1}{2}} \leqslant c \\ \frac{c}{2}\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\cdots\left(-\frac{2(m-1)-1}{2}\right) r^{-\frac{2(m-1)-1}{2}}, & \text{if } r^{\frac{1}{2}} > c \end{array} \right\} \tag{4.8}$$

According to Definition 8, Huber Robust kernel is a completely monotonous function, therefore, it is a conditionally positive definite.

4.4. Cauchy robust kernel

Definition 7. The Cauchy Robust kernel is defined as follows:

$$k_{\text{Cauchy}}(r) = \left\{ -\frac{c^2}{2} \log\left[1 + \left(\frac{r}{c}\right)^2\right] \right\} \tag{4.9}$$

Proposition 10. *Given a set of points $S = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, the preimage of its centroid in a feature space, $\mathcal{F}$, induced by a Cauchy kernel, k, corresponds to the Cauchy location M-estimator. In other words:*

$$P_\Phi(\mu) = P_\Phi\left(\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\right) = \arg\min_{y\in\mathcal{X}} \sum_{i=1}^{n} \rho_{\text{Cauchy}}(||x_i - y||)$$

Using the results in [55], it can be proved that Cauchy Robust Kernel is a conditional positive definite kernel.

Proposition 11 (*Berg*). *If $K_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is conditionally positive definite and satisfies $K_1(x,x) \leqslant 0$ for $x \in \mathcal{X}$, then any new built kernel defined as $K_2(x,y) = -(-K_1(x,y))^{\mathcal{B}}$ for $0 \leqslant \mathcal{B} \leqslant 1$or $K_3(x,y) = -ln(1 - K_1(x,y))$ is also conditional positive definite for $x, y \in \mathcal{X}$.*

Proposition 12. The *Cauchy Robust kernel is conditional positive definitive.*

Proof. The multiplication by $c^2/2$ and the fraction c are positive definite [55,51] so that we can remove them. Cauchy Robust kernel can be expressed as $-ln(1 - K)$ where K is the euclidean distance. According to the Definition 11, Cauchy Robust kernel is conditional positive definite.

The computational cost of these new robust kernels is similar to classic kernels as Gaussian Kernel and Polynomial kernel. There is a new cost due to the kernel being piecewise definite, however, the cost is very low so it can be ignored.

5. Empirical evaluation of robust kernels

The use of kernels in unsupervised learning has been well studied [13,10,22,15]. In this section, we evaluate the new robust kernels, presented in the previous section, using them as part of kernel clustering algorithms and show that the performance of these algorithms is on par with state-of-the-art robust clustering algorithms. Nevertheless, in this section, we will show, in a systematic way, that Tukey Kernel outperforms the results given by Linear and Gaussian Kernels and other non-kernel-based robust clustering algorithms, such as Robust Manifold Nonnegative Matrix Factorization (RMNMF) or Nonnegative Matrix Factorization Random Walks (NMFR) [39,40]. We evaluate different kernel-based algorithms that use linear, Gaussian, Tukey, Andrews, Huber, and Cauchy kernels. In our experiments (see Table 2).

5.1. Datasets

We used thirteen data sets for evaluation. The datasets are described below:

- Abalone is a dataset that predicts the age of an abalone from eight features including sex, length, diameter, and height. The age is discretized in three different classes [57].
- AR is a dataset of cropped images of 100 faces of 50 men, and 50 women. These images are frontal view faces that were taken in two different sessions with different facial expressions, illumination conditions, and occlusions such as by use of glasses or scarf. To manage the dimensionality of the images, we use an image resizing of one over five [58].
- AT&T is a dataset of 40 distinct subjects, where each subject has 10 different images. These images were taken at different times, varying the lighting, facial expressions, and facial details. To manage the dimensionality of the images, we use an image resizing of one over eight [59].
- Balance Scale is a dataset that models psychological experimental results where each example is classified as having the balance scale tip to the right, tip to the left, or be balanced [57].
- Coil is a dataset that consists of images of 20 objects that contains both the object and the background [60].
- Steel Plates Faults is a dataset of steel plates, classified into 7 different types[57].
- Jaffe is a dataset of images with 10 different Japanese females, where each set of images has 7 facial expressions. To manage the dimensionality of the images, we use an image resize of 1 over 25 [61].

Table 2
Data set description

Data set	Number of features	Number of samples	Number of classes
Abalone	8	4177	3
AR	792	2600	100
AT&T	168	400	40
Balance Scale (BS)	4	625	3
Coil	484	1440	20
Fault	27	1941	7
Jaffe	100	213	10
Movement Libras (ML)	90	360	15
Orl	1024	400	40
Segment	19	2310	7
Umist	644	575	20
Wineq	11	4898	3
Yale	1024	2414	38

- Movement Libras is a dataset that contains references to a hand movement type in LIBRAS which is the name of the official Brazilian signal language [57].
- The Orl database of faces is a dataset of face images from 40 distinct subjects. The images were taken at different times, varying light and different facial expression, among others. This database is the same as AT&T but with no reduction of the dimensionality [57].
- Image Segmentation (Segment) is a dataset of high-level numeric-valued attributes. The images were drawn randomly from a database of 7 outdoor images [57].
- The Sheffield (Umist) is an image dataset of 20 individuals (mixed race/gender/appearance). The photographs of the individuals show a range of poses from profile to frontal views [62].
- Wineq is a dataset of chemical analysis to determine the origin of different wines [57].
- Yale is a grayscale image dataset of 15 individuals, with 11 images per individual. Each image has a different configuration or expression.

5.2. Clustering methods

Clustering methods are used to identify groups of data where elements have some similarity [63–65]. Kernel-based clustering uses kernels as the similarity function in order to identify those groups. We use two kernel-based clustering methods for our evaluation of robust kernel: Convex nonnegative matrix factorization (CNMF) and Kernel K-Means (KKM).

- Kernel-Convex NMF (KCNMF) is a generalization of the Convex NMF algorithm. KCNMF solves the following optimization problem:$\arg min$-
  $||\phi(x) - \phi(x)WB||^2 = Tr(I - B^TW^T)\langle\phi(x), \phi(x)\rangle(I - B^TW^T)$, where $X \in \mathbb{R}^{p\times n}, B \in \mathbb{R}^{k\times n}$ and $W \in \mathbb{R}^{n\times k}$ [66].
- Kernel K-Means (KKM) is a generalization of the K-Means clustering algorithm. Basically, KKM performs k-means in a feature space implicitly defined by a kernel. The most important characteristic of KKM is that it is possible to carry it out without explicitly using the mapping Φ [67].

Ten baseline algorithms for clustering were selected, Several are classic methods such as K-Means while others are state-of-the-art methods:

- The K-Means algorithm represents the clusters by a set of centroids $\{C_1, \ldots, C_k\}$. These centroids are a disjoint partition of the input data set X, such that $X = \bigcup_{i=1}^{k} C_i$. The minimization is accomplished by an optimization process that iteratively reassigns data points to clusters while refining the centroid estimations in each iteration [68].
- Nonnegative matrix factoriazation: NMF attempts to factorize X as: $X_+ \approx F_+G_+^T$, where positive symbols means $X, F, G > 0$ and $X \in \mathbb{R}^{p\times n}, F \in \mathbb{R}^{p\times k}$ and $G \in \mathbb{R}^{n\times k}$. This type of factorization may be used to perform clustering. The input data points are the columns of X (p n-dimensional data points). The columns of F correspond to the coordinates of the centroids. The columns of G indicate to which cluster each sample belongs, specifically if x_j belongs to C_i, then $G_{i,j} = 1$, otherwise $G_{i,j} = 0$. With this interpretation, this function is equivalent to K-Means with the Euclidean distance. An important advantage of this approach is that values in the matrix G are not required to be binary; in fact, they can be continuous values. These values can be interpreted as soft membership values of data samples into clusters, i.e. NMF can produce a soft clustering of the input data [69,70].
- Convex nonnegative matrix factorization (CNMF): CNMF attempts to factorize X as: $X_\pm \approx X_\pm W_+ G_+^T$, where $X \in \mathbb{R}^{pxn}$, $W_+ \in \mathbb{R}^{pxk}$ and $G \in \mathbb{R}^{nxk}$[66].
- Projective NMF (PNMF): PNMF attempts to factorize X as: $X_+ \approx W_+W_+^TX^T, X \in \mathbb{R}^{p\times n}$ and $W \in \mathbb{R}^{p\times k}$ [71].
- Orthogonal Nonnegative matrix factorization (ONMF). ONMF is a matrix factorization method like NMF that attempts to find a subspace where the data lie, but this time the basis vectors F and G are restricted to be orthogonal, i.e., $F^TF = I$ and$G^TG = I$ [72].
- Normalized cuts (NCUT): NCUT is used and it is defined as follows: $Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}$, where A, B and V be two subsets of vertices; $assoc(A,V) = \sum_{u\in A, t\in v} w(u,t)$ is the sum of connection weights from nodes in A to all nodes in the graph; and $assoc(B,V)$ is similarly defined [73].
- NSC: The principal idea in Nonnegative Spectral Cut is to impose a nonnegative restriction in the optimization of Normalized Cut [67].
- Left-stochastic matrix factorization (LSD): The principal idea in left-stochastic matrix factorization is to produce a probability matrix $P \in \mathbb{R}^{n\times k}$ where each column i indicates the probability that each row sample j belongs to the i^{th} cluster. This method deals with a similarity matrix that is a more general matrix compared to the kernel matrix where the similarity matrix needs not be positive semi-definite (PSD) and it can be an indefinite kernel matrix [74].
- Robust Manifold NMF (RMNMF) uses the $l_{2,1}$ norm instead of Frobenius. The objective function of the RMNMF algorithm, which is based on the error between matrix X and factorization FG^T, needs to be robust. This is achieved by using a robust distance for the values with noise in matrix X. A regularization with the Laplacian graph is also used in order to obtain a spectral clustering. For further reference, see [39].
- Nonnegative matrix factorization random walks (NMFR) uses the random walks notion in order to improve clustering results in Spectral Clustering. To do so, a W regularization parameter is added, thus: $\min_{W\geq 0} - Tr(W^TAW) + \lambda\sum_i(\sum_k W_{ik}^2)^2$ $s.t. W^TW = I$. In so doing, the author claims that the trace is minimized since by augmenting parameter λ, optimization will tend to give lesser values to its diagonal [40].

5.3. Experimental setup

The goal of the experimental evaluation in this section is to show the robustness of kernel algorithms in unsupervised learning. In the case of kernel algorithms, four different kernels will be used; linear kernel, Gaussian kernel, Tukey kernel, Andrews kernel, Huber kernel and Cauchy kernel.

As evaluation measure, we use *clustering accuracy* that requires a one to one mapping between clusters and classes, which is known to be an NP-hard problem for $k \geqslant 3$. The confusion matrix has a dimensionality of k clusters by k' classes where the$(i,j)^th$ entry of the matrix correspond to the number of data points from cluster i that are classified in class j. Clustering accuracy is computed with the following equation $Acc = \frac{\max Trace(confusion\_matrix())}{N}$. To solve this maximization problem the Hungarian algorithm is used [75].

5.4. Hyperparameters optimization

We performed an exploration of hyperparameters for each of the proposed and baseline methods. The exploration was performed

with the appropriate scale for the type of parameter. In the particular case of the bandwidths of the different kernels used, a logarithmic scale was used, varying them from two to the power of minus fifth power to two to the fifth power.

5.5. Results

Fig. 2 shows the sensibility of clustering accuracy for each robust kernel with different bandwidth parameter. Tukey, Andrew's and RBF show good results between two to the minus second power to two to the second power for the Movement Libras dataset. Both RBF and Andrews robust kernels have poor clustering accuracy for high values of the bandwidth parameter. Tukey robust kernel has better results for both datasets when the bandwidth parameter is greater.

Fig. 3 shows the clustering accuracy results for thirteen datasets and CNMF as the main algorithm with five robust kernels and one linear kernel. It can be seen that Tukey KCNMF is better in six of the thirteen datasets. Also, Andrews KCNMF is better in three of the thirteen datasets. It is worth nothing that CNMF without any Robust Kernels is the worst in three of the thirteen datasets and also it is not the best in any of the thirteen datasets.

Fig. 4 shows clustering accuracy results for thirteen datasets and KMeans as the main algorithm with five robust kernels and one linear kernel. It can be seen that RBF KMeans is better in three of the thirteen datasets. Also, Tukey KMeans is better in three of the thirteen datasets. It is worth nothing that KMeans with the linear kernel is not the best in any of the thirteen datasets.

Table 3 presents the findings after running each algorithm thirty (30) times for each algorithm in every data set. Each cell of the table presents the mean accuracy for the column algorithm with the row data set. It can be observed that none of the algorithms is the best in every data set, also that the Gaussian, Tukey, Andrews, Huber and Cauchy kernels together are the best in five of the 13 datasets. The NMFR state-of-the-art algorithm performs quite well, being the best in five of the 13 datasets but with poor performance on datasets like Abalone, AR and WineQ.

Given that none of the algorithms is the best with every data set, it is necessary to do a non-parametric test to verify if a significant difference exists among algorithms. A variance analysis non-parametric Friedman' test was used in order to test the behavior of the different algorithms. It is sought to determine whether or not there is a significant difference between the different algorithms in Table 3. The test is implemented as follows: Based on $\{x_{ij}\}_{m\times n}$ - a data table where m is the number of datasets and n is the number of algorithms. The rank of the algorithms is defined for each of the datasets, organizing them from highest to lowest in accordance with the clustering accuracy presented in Table 3. Finally, the average rank, among the datasets, is calculated for each algorithm, as reported in the last column of Table 4. A small average rank for an algorithm means that the algorithm performs better than other algorithms, so it is ranked close to the top, for several datasets.

The null hypothesis of the Friedman's test is that all algorithms are equivalent. To perform the test, we calculate:

$$\chi_{F^2} = \frac{12n}{k(k+1)}\left(\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right)$$

where R_j is the rank of the j-th algorithm, and it is distributed according to a χ_{F^2} with $k-1$ degrees of freedom. A F_F statistic is found with

$$F_F = \frac{(n-1)\chi_{F^2}}{n(k-1)-\chi_{F^2}} = 3.6524$$

With this, we rule out the null hypothesis that states that there is no significant statistical difference in the accuracy of the methods. In light of the above, a pairwise Nemenyi post hoc test is performed in order to attempt to determine whether there is a significant difference between two algorithms. To this effect, two algorithms are significantly different if the range average differs by at least one critical difference

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6n}}$$

Where $k = 19$ and $n = 13$. q_α is obtained as the t standardized distribution, divided by $\sqrt{2}$. Fig. 5 illustrate the differences between methods according to their average rank and the critical differences. The length of the bars for each method correspond to the critical difference (CD). According to the Nemenyi post hoc test, the average rank of two methods is significantly different if the corresponding bars do not intercept.

5.6. Discussion

In this section, we presented a systematic comparison between state-of-the-art algorithms and robust kernels built under Proposition 1. It was found at first glance that the kernel algorithms, Kernel KMeans and Kernel CNMF, perform better when a robust kernel, such as Gaussian or Tukey Kernel, is used compared to the use of a linear kernel. Furthermore, it was found that Tukey kernel improves the results obtained by the Gaussian kernel; this would imply that this kernel could be used in other domains where the Gaussian kernel has been successful, notably in conjunction

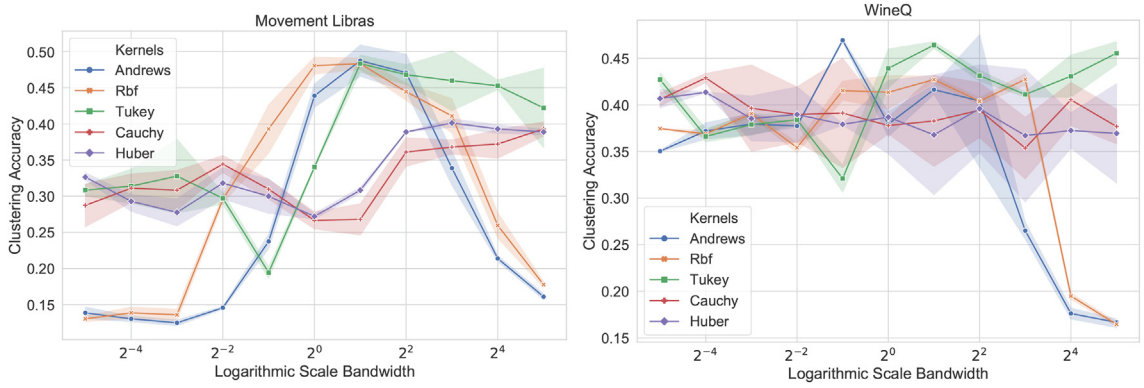

Fig. 2. Sensibility study of bandwidth parameters (c or σ) for each robust kernel in two datasets using kernel convex nonnegative matrix factorization (Kernel CNMF). The x-axis represents a logarithmic scale from two to the minus fifth power to two to the fifth power. The hue represents the standard deviation of the clustering accuracy for a given robust kernel with a certain bandwidth number.

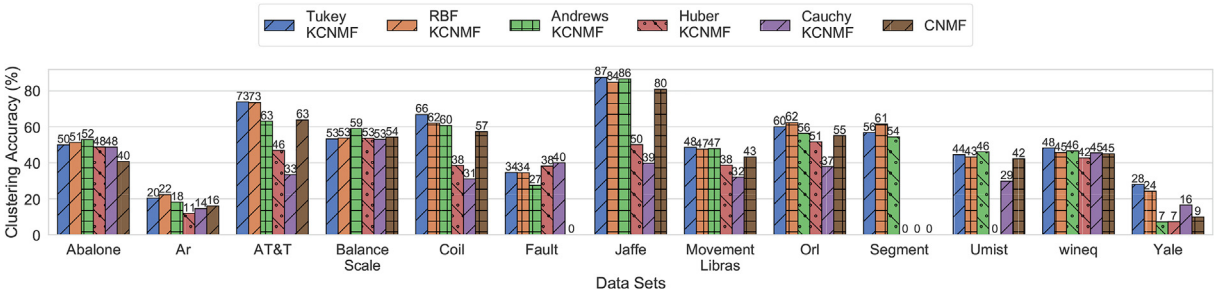

Fig. 3. Clustering accuracy results for Convex Non Negative Matrix Factorization (CNMF) with different robust kernels.

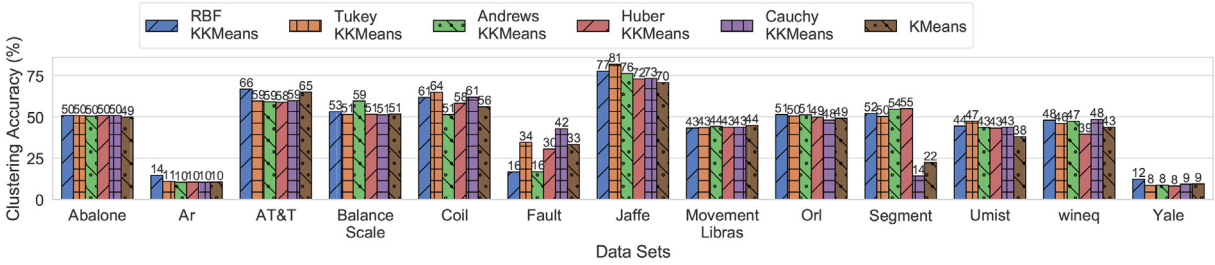

Fig. 4. Clustering accuracy results for Kernel KMeans algorithm with different kernels.

Table 3
Clustering Accuracy Results for different robust kernel and non-kernel methods.

	Abalone	AR	AT&T	Scale Balance	Coil	Fault	Jaffe	Libras Movement	ORL	Segment	Umist	Wineq	Yale
Tukey KCNMF	0.500	0.203	0.738	0.533	0.668	0.346	0.875	0.486	0.600	0.568	0.445	0.482	0.280
RBF KCNMF	0.513	**0.223**	0.735	0.538	0.620	0.345	0.848	0.4756	0.623	0.615	0.432	0.457	0.242
Andrews KCNMF	**0.529**	0.181	0.631	0.590	0.607	0.275	0.866	0.479	0.564	0.544	0.461	0.466	0.072
Huber KCNMF	0.488	0.119	0.469	0.537	0.385	0.382	0.501	0.387	0.517	0	0	0.427	0.073
Cauchy KCNMF	0.487	0.145	0.334	0.5301	0.310	0.400	0.398	0.32	0.379	0	0.298	0.455	0.165
RBF KKMeans	0.508	0.146	0.668	0.531	0.617	0.167	0.776	0.433	0.514	0.520	0.445	0.480	0.123
Tukey KKMeans	0.507	0.110	0.596	0.515	0.649	0.346	0.818	0.435	0.505	0.503	0.474	0.460	0.085
Andrews KKMeans	0.505	0.104	0.591	**0.596**	0.514	0.168	0.762	0.442	0.512	0.545	0.436	0.473	0.087
Huber KKMeans	0.508	0.106	0.588	0.516	0.582	0.306	0.729	0.437	0.499	0.550	0.432	0.395	0.082
Cauchy KKMeans	0.508	0.105	0.598	0.512	0.619	**0.428**	0.732	0.437	0.481	0.143	0.437	**0.484**	0.093
NSC	0.388	0.183	0.799	0.562	0.806	0.279	**0.910**	0.470	0.661	0.611	0.615	0.445	**0.314**
NCUT	0.387	0.181	0.801	0.559	0.804	0.262	**0.910**	0.467	0.671	0.610	0.588	0.456	0.308
LSD	0.388	0.171	0.810	0.546	0.754	0.252	**0.910**	0.496	**0.676**	0.641	0.613	0.400	0.300
NMFR	0.372	0.160	**0.810**	0.569	**0.850**	0.234	**0.910**	0.461	0.671	**0.733**	**0.680**	0.393	0.296
PNMF	0.404	0.177	0.8	0.557	0.735	0.246	**0.910**	0.500	0.662	0.506	0.513	0.413	0.268
ONMF	0.383	0.1383	0.798	0.550	0.669	0.246	**0.910**	**0.500**	0.665	0.504	0.575	0.417	0.258
RMNMF	0.520	0.212	0.693	0.585	0.554	0.337	0.809	0.429	0.328	0.491	0.433	0.443	0.215
ConvexNMF	0.407	0.160	0.638	0.543	0.573	0	0.808	0.432	0.551	0	0.422	0.450	0.099
KMeans	0.499	0.1061	0.650	0.517	0.561	0.333	0.707	0.4478	0.492	0.224	0.380	0.437	0.095

with support vector machines [76,51]. Finally, our comparisons to several state-of-the-art algorithms such as robust manifold, random walk NMF, and NMF, found no significant difference between these algorithms and Robust Kernel CNMF that uses Tukey kernel. Something worth noting is that the two algorithms RNMF and RKNMF use the framework of augmented Lagrangian optimization, which could be used to improve the results of KCNMF, and could also kernelize the two algorithms RNMF and RMNMF.

Tukey KCNMF performs quite well on non-image data such as Fault and Wineq. State-of-the-art algorithms such as NSC, NCUT and NMFR do not have as good a performance as Robust Kernel KCNMF algorithms on non-image data. The poor performance of Robust Kernels for image data may be due to the CNMF Algorithm's sub-optimal optimization process. A future testing with LSD and Robust Kernels such as Tukey kernel could show better performance, because the linear approach, LSD, shows very good results on Image data. Andrews robust kernel exhibited a strange behavior because on some data such as Abalone and Umist, it outperformed the other robust kernels, even though for the other datasets, it showed a very poor performance. This behavior may be due to the mapping generated by the cosine function used in the Andrews kernel. On the contrary, by using a robust kernel such as the RBF kernel, the kernel K-Means was able to obtain a robust cluster estimation in spite of the noise contamination. Huber and Cauchy kernels had a poor performance with Kernel CNMF. These behaviors due to the Kernel CNMF does not converge to any value when all the values of the Gram Matrix are negative. Further modifications to kernel CNMF are necessary to work with these two kernels.

Additionally, a sensibility study of bandwidth parameters (c or σ) in Movement Libras and Fault datasets was performed. In Movement Libras, there is a clear behavior of the dependency of a good selected bandwidth parameter on the clustering accuracy. There is a gap of almost forty percent between a good parameter selection and a bad one. In WineQ, there is not a clear behavior of the dependency

Table 4
The rank of each clustering method on the different datasets. The last column corresponds to the average rank, a lower value is better. 3.

	Abalone	AR	AT&T	Scale Balance	Coil	Fault	Jaffe	Libras Movement	ORL	Segment	Umist	Wineq	Yale	**Average**	**Rank**
Tukey KCNMF	9	3	7	13	7	4	7	4	8	6	9	2	5	6.46	4
RBF KCNMF	3	1	8	11	9	6	9	6	7	3	14	7	8	7.08	6
Andrews KCNMF	1	5	13	2	12	11	8	5	9	9	8	5	19	8.23	8
Huber KCNMF	11	14	18	12	18	3	18	18	11	17	19	14	18	14.69	19
Cauchy KCNMF	12	12	19	15	19	2	19	19	18	17	18	9	10	14.54	18
RBF KKMeans	4	11	10	14	11	18	13	15	12	10	9	3	11	10.85	11
Tukey KKMeans	7	15	15	18	8	4	10	14	14	13	7	6	16	11.31	12
Andrews KKMeans	8	19	16	1	17	17	14	11	13	8	12	4	15	11.92	14
Huber KKMeans	4	17	17	17	13	9	16	12	15	7	14	18	17	13.54	16
Cauchy KKMeans	4	18	14	19	10	1	15	12	17	16	11	1	14	11.69	13
NSC	15	4	5	5	2	10	1	7	6	4	2	11	1	5.62	1
NCUT	17	5	3	6	3	12	1	8	2	5	4	8	2	5.85	2
LSD	15	8	1	9	4	13	1	3	1	2	3	17	3	6.15	3
NMFR	19	9	1	4	1	16	1	9	2	1	1	19	4	6.69	5
PNMF	14	7	4	7	5	14	1	1	5	11	6	16	6	7.46	7
ONMF	18	13	6	8	6	14	1	1	4	12	5	15	7	8.46	9
RMNMF	2	2	9	3	16	7	11	17	19	14	13	12	9	10.31	10
ConvexNMF	13	9	12	10	14	19	12	16	10	17	16	10	12	13.08	15
KMeans	10	16	11	16	15	8	17	10	16	15	17	13	13	13.62	17

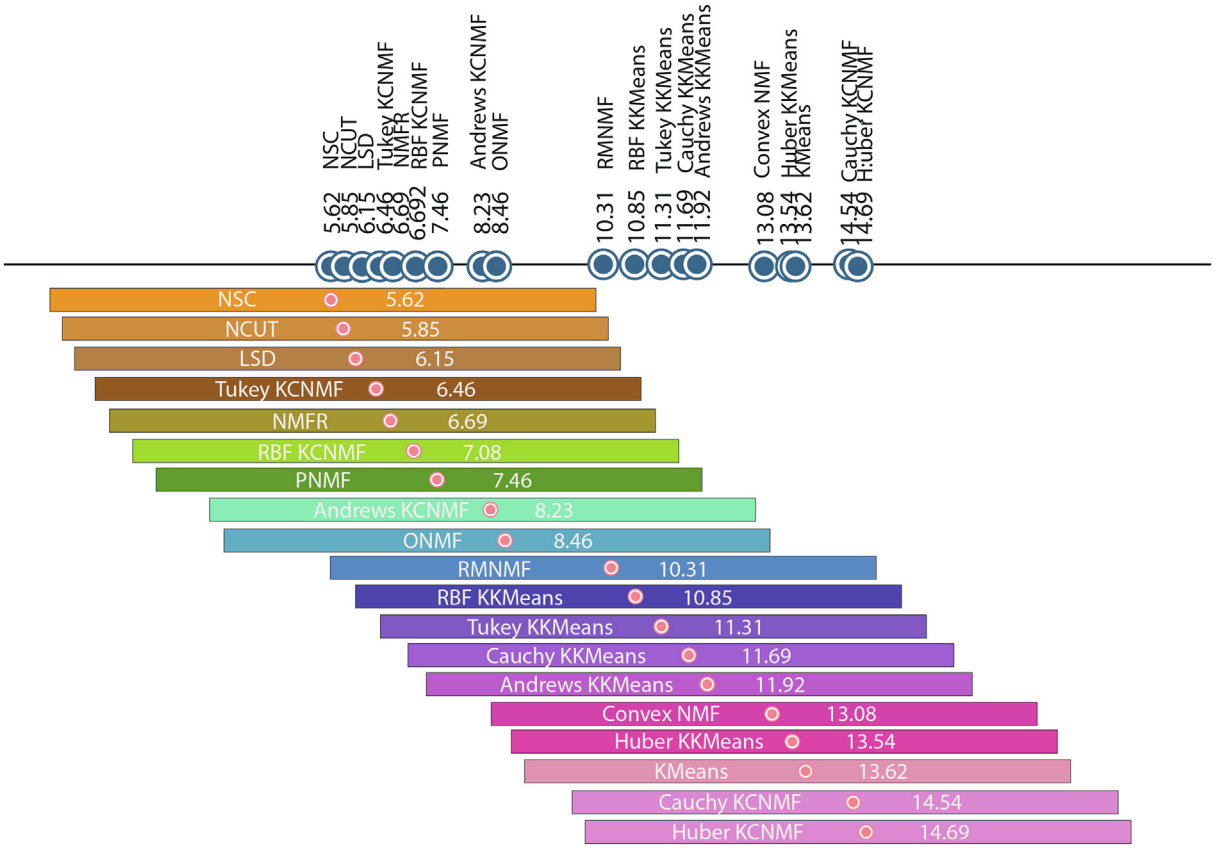

Fig. 5. Pairwise Nemenyi test with critical difference of 4.72 for results of clustering accuracy in Table 3. NSC is the algorithm with less average rank (5.62) followed by NCUT with 5.85 of average rank. The worst method according to average rank is Huber KCNMF with average rank of 14.69.

of a good selected bandwidth parameter. In robust statistics, the bandwidth parameter controls the robustness of the M-estimator. There seems to be a relationship between the robustness of the method and the selected bandwidth parameter.

6. Conclusions and future work

We showed that performing location estimation in a feature space obtained from special kernels such as the Gaussian kernel, is equivalent to performing location estimation in the original data space using a robust M-estimator. M-estimators have been used recently in different areas of machine learning, pattern recognition, and data mining. A connection between M-estimators and kernels opens the interesting possibility of working with contaminated data sets in a non-linear feature space. Building new kernels that match the notion of high dimensionality with robust statistics opens new possibilities in kernel mean embedding, support vector machines, scattered data approximation, kernel principal component analysis among others. Tukey Robust kernel is not positive definite PD^n for all n. Andrews Robust kernel is not positive definite even not conditional

positive definite. Huber and Cauchy Robust kernel are conditionals positive definite of order 1. We could find - in the practical realm - that Tukey kernels and Andrews kernels also work fairly well, and that furthermore, with the aid of Convex Nonnegative Matrix Factorization, they are as good as Nonnegative Spectral Cut and Normalized Cuts, since it could not be statistically proven that there is a significant difference between these methods.

Tukey and Andrews kernels showed good performance in the unsupervised machine learning task of clustering. Further testing in other areas such as dimensionality reduction with principal component analysis, or classification with support vector machines, is needed to show the potential advantage of using robust kernels. Future research calls to build new algorithms with new robust and sparse kernels. Tukey and Andrews kernels have the particularity of being sparse in accordance with the regularization parameter. Besides, new kernels can be defined by combining other kernels, this means that a mixture of robust kernels may be robust. Exploring the robustness of these mixtures is also part of our future work.

CRediT authorship contribution statement

Joseph A. Gallego: Methodology, Writing - original draft, Software. **Fabio A. González:** Conceptualization, Methodology, Validation, Writing - original draft. **Olfa Nasraoui:** Conceptualization, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by a visiting scholar grant awarded to F. González by the Fulbright foundation. We acknowledge partial support from National science foundation CAREER Award NSF IIS 0533317 to Olfa Nasraoui.

References

[1] R. Wilcox (Ed.), Copyright, third edition Edition, Statistical Modeling and Decision Science, Academic Press, Boston, 2012. doi:https://doi.org/10.1016/B978-0-12-386983-8.00016-0.
[2] P.J. Rousseeuw, M. Hubert, Robust statistics for outlier detection, Wiley Interdisciplinary Reviews, Data Min. Knowl. Disc. 1 (1) (2011) 73–79.
[3] G. Shevlyakov, N. Vilchevski, Robustness in Data Analysis: criteria and methods, 2001.
[4] F.R. Hampel, Robust statistics, Wiley Sieres in Probability and Statistics (1985).
[5] G. Li, Z. Chen, Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo, J. Am. Stat. Assoc. 80 (391) (1985) 759–766.
[6] C. Croux, A. Ruiz-Gazen, High breakdown estimators for principal components: the projection-pursuit approach revisited, J. Multivariate Anal. 95 (1) (2005) 206–226.
[7] M. Hubert, P.J. Rousseeuw, S. Verboven, A fast method for robust principal components with applications to chemometrics, Chemometrics and Intelligent Laboratory Systems 60 (1–2) (2002) 101–111.
[8] C. Croux, P. Filzmoser, M.R. Oliveira, Algorithms for projection–pursuit robust principal component analysis, Chemometrics Intell. Lab. Syst. 87 (2) (2007) 218–225.
[9] S.-A. Berrani, C. Garcia, Robust detection of outliers for projection-based face recognition methods, Multimedia Tools Appl. 38 (2) (2008) 271–291.
[10] A. Ben-Tal, S. Bhadra, Efficient methods for robust classification under uncertainty in kernel matrices, J. Mach. Learn. Res. 13 (2012) 2923–2954.
[11] L.A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Iscar, A review of robust clustering methods, Adv. Data Anal. Classif. 4 (2–3) (2010) 89–109.
[12] K.V. Branden, M. Hubert, Robust classification in high dimensions based on the simca method, Chemometrics Intell. Lab. Syst. 79 (1–2) (2005) 10–21.
[13] M. Amer, M. Goldstein, S. Abdennadher, Enhancing one-class support vector machines for unsupervised anomaly detection, in: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description - ODD '13 (2013) 8–15 doi:10.1145/2500853.2500857.
[14] A. L. B. Barros, G. A. Barreto, Building a robust extreme learning machine for classification in the presence of outliers, in: International Conference on Hybrid Artificial Intelligence Systems, Springer, 2013, pp. 588–597.
[15] R.N. Davé, R. Krishnapuram, Robust clustering methods: a unified view, Fuzzy Systems, IEEE Trans. 5 (2) (1997) 270–293.
[16] U. Boryczka, Finding groups in data: Cluster analysis with ants, Appl. Soft Comput. 9 (1) (2009) 61–70.
[17] Z. Chen, X. Shixiong, L. Bing, A robust fuzzy kernel clustering algorithm, Appl. Math. 7 (3) (2013) 1005–1012.
[18] J. Cuesta-Albertos, A. Gordaliza, C. Matrán, et al., Trimmed k-means: An attempt to robustify quantizers, An. Stat. 25 (2) (1997) 553–576.
[19] P. A. Forero, V. Kekatos, G. B. Giannakis, Outlier-aware robust clustering, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 2244–2247.
[20] J. Hardin, D.M. Rocke, Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator, Comput. Stat. Data Anal. 44 (4) (2004) 625–638.
[21] J. Kim, C. Scott, Robust kernel density estimation, The, J. Mach. Learn. Res. (2012) 3381–3384.
[22] J.-H. Chen, M-estimator based robust kernels for support vector machines, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. 1 (1) (2004) 168–171 Vol. 1. doi:10.1109/ICPR.2004.1334039.
[23] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust statistics: the approach based on influence functions, Vol. 114, John Wiley & Sons, 2011.
[24] P. Huber, Robust statistics, Wiley Sieres in Probability and Statistics (2011).
[25] H. Rieder, P. Huber, Robust statistics, data analysis and computer intensive methods, Springer-Verlag, New York, 1996.
[26] D.E. Tyler, Robust statistics: Theory and methods, J. Am. Stat. Assoc. 103 (482) (2008) 888–889.
[27] C.-T. Liao, S.-H. Lai, Robust kernel-based learning for image-related problems, IET Image Process. 6 (6) (2012) 795–803.
[28] S. A. Shah, V. Koltun, Robust continuous clustering, in: Proceedings of the National Academy of Sciences of the United States of America 114 (37) (2017) 9814–9819. arXiv:1803.01449, doi:10.1073/pnas.1700770114.
[29] F.H. Ruymgaart, A robust principal component analysis, J. Multivariate Anal. 11 (4) (1981) 485–497.
[30] J. Liu, J. Li, W. Xu, Y. Shi, A weighted lq adaptive least squares support vector machine classifiers–robust and sparse approximation, Expert Syst. Appl. 38 (3) (2011) 2253–2259.
[31] F. De la Torre, M. J. Black, Robust principal component analysis for computer vision, in: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on 1 (2001) 362–369.
[32] H.H. Huang, Y.R. Yeh, An iterative algorithm for robust kernel principal component analysis, Neurocomputing 74 (18) (2011) 3921–3930, https://doi.org/10.1016/j.neucom.2011.08.008.
[33] M. Svensén, C. M. Bishop, Robust Bayesian mixture modelling, Neurocomputing 64 (1-4 SPEC. ISS.) (2005) 235–252. doi:10.1016/j.neucom.2004.11.018.
[34] G. Gan, M.K.-P. Ng, K-means clustering with outlier removal, Pattern Recogn. Lett. 90 (2017) 8–14.
[35] J.C. Bezdek, R. Ehrlich, W. Full, Fcm: The fuzzy c-means clustering algorithm, Computers Geosci. 10 (2–3) (1984) 191–203.
[36] R. Krishnapuram, J.M. Keller, The possibilistic c-means algorithm: insights and recommendations, IEEE Trans. Fuzzy Syst. 4 (3) (1996) 385–393.
[37] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, A. Mayo-Íscar, Robust clustering tools based on optimal transportation, Statistics and Computing 29 (1) (2019) 139–160. arXiv:1607.01179, doi:10.1007/s11222-018-9800-z. doi: 10.1007/s11222-018-9800-z.
[38] Z. Zhou, G. Si, Y. Zhang, K. Zheng, Robust clustering by identifying the veins of clusters based on kernel density estimation, Knowl.-Based Syst. 159 (2018) 309–320, https://doi.org/10.1016/j.knosys.2018.06.021.
[39] L. Zhang, Z. Chen, M. Zheng, X. He, Robust non-negative matrix factorization, Front. Electr. Electron. Eng. China 6 (2) (2011) 192–200.
[40] Z. Yang, T. Hao, O. Dikmen, X. Chen, E. Oja, Clustering by nonnegative matrix factorization using graph random walk, in: Advances in Neural Information Processing Systems, 2012, pp. 1079–1087.
[41] M. Debruyne, M. Hubert, J. Van Horebeek, Detecting influential observations in kernel pca, Comput. Stat. Data Anal. 54 (12) (2010) 3007–3019.
[42] C. Lu, T. Zhang, R. Zhang, C. Zhang, Adaptive robust kernel pca algorithm, in: Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on, Vol. 6, IEEE, 2003, pp. VI–621.
[43] Z. Liang, D. Zhang, P. Shi, Robust kernel discriminant analysis and its application to feature extraction and recognition, Neurocomputing 69 (7–9 SPEC. ISS.) (2006) 928–933. doi:10.1016/j.neucom.2005.09.001.
[44] Z. Kang, C. Peng, Q. Cheng, Kernel-driven similarity learning, Neurocomputing 267 (2017) 210–219, https://doi.org/10.1016/j.neucom.2017.06.005.
[45] Z. Kang, H. Xu, B. Wang, H. Zhu, Z. Xu, Clustering with similarity preserving, Neurocomputing 365 (2019) 211–218, https://doi.org/10.1016/j.neucom.2019.07.086.
[46] F. A. González, D. Bermeo, L. Ramos, O. Nasraoui, On the robustness of kernel-based clustering, in: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer, 2012, pp. 122–129.

[47] C. S. Ong, X. Mary, S. Canu, A. J. Smola, Learning with non-positive kernels, in: Proceedings of the twenty-first international conference on Machine learning, ACM, 2004, p. 81.
[48] J.T.-Y. Kwok, I.W.-H. Tsang, The pre-image problem in kernel methods, IEEE Trans. Neural Networks/Publ. IEEE Neural Networks Council 15 (6) (2004) 1517–1525, https://doi.org/10.1109/TNN.2004.837781.
[49] M.G. Genton, Classes of kernels for machine learning: a statistics perspective, J. Mach. Learn. Res. 2 (2) (2001) 299–312.
[50] R. Askey, Radial characteristics functions., Tech. rep., WISCONSIN UNIV MADISON MATHEMATICS RESEARCH CENTER (1973).
[51] N. Cristianini, J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge University Press, 2000.
[52] G.E. Fasshauer, Meshfree approximation methods with MATLAB, Vol. 6, World Scientific, 2007.
[53] R. Schaback, H. Wendland, Characterization and construction of radial basis functions, in: Multivariate Approximation and Applications, 2010, pp. 1–24. doi:10.1017/cbo9780511569616.002.
[54] Z. Wu, Compactly supported positive definite radial functions, Adv. Comput. Math. 4 (1) (1995) 283–292, https://doi.org/10.1007/BF03177517.
[55] C. Berg, J. P. R. Christensen, P. Ressel, Harmonic analysis on semigroups: theory of positive definite and related functions, Vol. 53, 1984. arXiv: arXiv:1011.1669v3.
[56] B. Schölkopf, The kernel trick for distances, Advances in Neural Information Processing Systems.
[57] K. Bache, M. Lichman, UCI machine learning repository (2013). http://archive.ics.uci.edu/ml.
[58] A. M. Martinez, The ar face database, CVC Technical Report 24.
[59] F. Samaria, The orl database of faces, AT&T Laboratories Cambridge 1.
[60] S. K. N. S. A. Nene, H. Murase., Columbia university image library (coil-20), Technical Report CUCS-005-96 1.
[61] M. J. Lyons, Coding facial expressions with gabor wavelets, in: 3rd IEEE International Conference on Automatic Face and Gesture Recognition 1.
[62] A. N. Graham Daniel, Face recognition: From theory to applications, Face Recognition: From Theory to Applications 163.
[63] O. Nasraoui, C. C. Uribe, C. R. Coronel, F. Gonzalez, Tecno-streams: Tracking evolving clusters in noisy data streams with a scalable immune system learning model, in: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, IEEE, 2003, pp. 235–242.
[64] O. Nasraoui, C. Rojas, Robust clustering for tracking noisy evolving data streams, in: Proceedings of the 2006 SIAM International Conference on Data Mining, SIAM, 2006, pp. 619–623.
[65] O. Nasraoui, R. Krishnapuram, A robust estimator based on density and scale optimization, and its application to clustering, in: IEEE International Conference on Fuzzy Systems, 1996, pp. 1031–1035.
[66] C. Ding, T. Li, M.I. Jordan, Convex and semi-nonnegative matrix factorizations, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 45–55.
[67] I. S. Dhillon, Y. Guan, B. Kulis, Kernel k-means: spectral clustering and normalized cuts (2004) 551–556.
[68] J.A. Hartigan, M.A. Wong, Algorithm as 136: A k-means clustering algorithm, J. R. Stat. Soc. Ser. C (Appl. Stat.) 28 (1) (1979) 100–108.
[69] P. Paatero, U. Tapper, Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values, Environmetrics 5 (2) (1994) 111–126, https://doi.org/10.1002/env.3170050203.
[70] C. Ding, T. Li, W. Peng, On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing, Comput. Stat. Data Anal. 52 (8) (2008) 3913–3927.
[71] M.-S. Yang, K.-L. Wu, J.-N. Hsieh, J. Yu, Alpha-cut implemented fuzzy clustering algorithms and switching regressions, IEEE Trans. Syst., Man, Cybern., Part B (Cybern.) 38 (3) (2008) 588–603.
[72] C. Ding, T. Li, W. Peng, H. Park, Orthogonal nonnegative matrix t-factorizations for clustering, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2006, pp. 126–135.
[73] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.
[74] R. Arora, M. Gupta, A. Kapila, M. Fazel, Clustering by left-stochastic matrix factorization, in, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 761–768.
[75] G. Aggarwal, S. Garg, N. Gupta, Combining clustering solutions with varying number of clusters, Int. J. Computer Sci. Issues (IJCSI) 11 (2) (2014) 240.
[76] W. Wen, Z. Hao, X. Yang, Robust least squares support vector machine based on recursive outlier elimination, Soft. Comput. 14 (11) (2009) 1241–1251, https://doi.org/10.1007/s00500-009-0535-9.

Joseph A. Gallego is currently pursuing Ph.D. degree in Systems and Computing Engineering from the National University of Colombia, Colombia. He earned a Computing Systems Engineer degree and an Industrial Engineer degree from the National University of Colombia. He earned a MSc in Systems and Computer Engineering from the National University of Colombia. He is currently a student associated with the MindLab research group of the National University of Colombia, Bogot?, conducting research in Machine Learning, statistical learning, data mining, analysis of data among others.

Fabio A. González, PhD. is a full Professor at the Department of Computing Systems and Industrial Engineering at the National University of Colombia, where he leads the Machine Learning, Perception and Discovery Lab (MindLab). He earned a Computing Systems Engineer degree and a MSc in Mathematics degree from the National University of Colombia, and a MSc and PhD degrees in Computer Science from the University of Memphis. His research work revolves around machine learning and its applications in information retrieval, computer vision, natural language understanding and biomedical image analysis, with a particular focus on the representation, indexing and automatic analysis of multimodal data (data encompassing different types of information: textual, visual, signals, etc.).

Olfa Nasraoui received the Ph.D. degree in computer engineering and computer science from the University of Missouri, Columbia, in 1999. She is currently the endowed Chair of e-commerce and the Founding Director of the Knowledge Discovery and Web Mining Laboratory, University of Louisville, where she is also a Professor of computer engineering and computer science. She was a recipient of the National Science Foundation CAREER Award, two Best Paper Awards for theoretical contributions in computational intelligence at the ANNIE 2001 Conference, and a recent one at the Knowledge Discovery and Information Retrieval, KDIR 2018 Conference.