Robust One-Class Kernel Spectral Regression
Shervin Rahimzadeh Arashloo and Josef Kittler, Life Member, IEEE

Abstract—The kernel null-space technique and its regression-based formulation (called one-class kernel spectral regression, a.k.a. OC-KSR) is known to be an effective and computationally attractive one-class classification framework. Despite its outstanding performance, the applicability of kernel null-space method is limited due to its susceptibility to possible training data corruptions and inability to rank training observations according to their conformity with the model. This work addresses these shortcomings by studying the effect of regularising the solution of the null-space kernel Fisher methodology in the context of its regression-based formulation (OC-KSR). In this respect, first, the effect of a Tikhonov regularisation in the Hilbert space is analysed where the one-class learning problem in the presence of contaminations in the training set is posed as a sensitivity analysis problem. Next, driven by the success of the sparse representation methodology, the effect of a sparsity regularisation on the solution is studied. For both alternative regularisation schemes, iterative algorithms are proposed which recursively update label confidences, facilitating ranking of the training observations based on their fit to the model. Through extensive experiments conducted on different data sets, the proposed methodology is found to enhance robustness against contamination in the training set, as compared with the baseline kernel null-space technique as well as other existing approaches in the one-class classification paradigm while providing the functionality to rank training samples effectively.

Index Terms—Kernel null-space technique, contamination, regression, regularisation.

I. INTRODUCTION

While for a wide variety of pattern classification problems there is an abundance of data for the system design, there are applications where the quantity of the training data is severely limited, or the data is in some respect unrepresentative or imbalanced. This is manifest in applications, where the cost associated with collecting samples from a specific class is prohibitively high, or samples take unexpected form at the test time. There also exist other scenarios where the recognition task is inherently open, leading to inaccurately defined class(es). In these situations, a particular class may not be very well represented by the available training samples or be totally void of samples, causing a degradation in the performance of conventional multi/binary-class classifiers.

An alternative and more effective approach in such cases is offered by the one-class classification (OCC) paradigm [1]. One-class classification aims to identify patterns from the target class, conforming to a specific behaviour, and distinguish them from non-target objects. It differs from the conventional multi/binary-class formulation in that it primarily uses samples from a single class for training. More concretely, assume \( X = \{x_1, \ldots, x_n\} \) to be a set of observations where \( x_i \in \mathbb{R}^d \) is a realisation of a multivariate random variable \( x \) characterised by the target probability density function \( p(x) \). One-class learning tries to specify the support domain of \( p(x) \) using a one-class classifier \( h(z) \) as

\[
h(z) = [f(z) \geq \tau] = \begin{cases} 
1 & z \in \mathcal{T} \\
0 & \text{otherwise}
\end{cases}
\]

where \( \mathcal{T} \) denotes the target class and function \( f(.) \) encodes the similarity of an observation \( z \) to \( \mathcal{T} \), while \([.]\) stands for the Iverson brackets. Parameter \( \tau \) controls the fraction of observations which lie within the support domain of the target distribution.

One-class classification is typically believed to be a more difficult problem than its multi/binary-class counterpart due to the lack of non-target training samples, impeding the estimation of a decision boundary between the target and non-target observations. Yet, it forms the basis of a variety of different applications including intrusion detection [2], novelty detection [3], fault detection in safety-critical systems [4], fraud detection [5], insurance [6], health care [7], surveillance [8], etc.

In order to identify a pattern as normal or novel, a one-class classifier is trained on a training set consisting of (typically) normal samples and then used to gauge the similarity of a test sample to those previously observed in the training set. In this case, the generalisation capability of the one-class learner on new data plays a central role. In a real-life one-class learning problem, however, not all observations in the training set would conform to the model equally. In practice, the training set may be corrupted and incorporate noisy or non-target observations which may degrade performance, emphasising the requirement for mechanisms to identify such samples. The identification of the degree of normality of samples in a given set is also useful in its own right for specific applications such as ranking and retrieving items from a database for a given query, pruning contaminated data sets as well as in decision threshold setting. It also opens the possibility of using the detected non-target samples to refine the decision boundary. Consequently, a ranking of the set of training observations according to their degree of fit with the model is highly desirable in a one-class learning paradigm.

While there exist a variety of different OCC methods and fine-grained categorisation of such methods exists [9], [10], [11], they can be roughly identified as either generative or non-generative [12]. Generative approaches include a model for data generation whereas non-generative methods lack a transparent link to the data. Among the well-known instances of the generative approaches are the parametric and non-

S.R. Arashloo is with the Department of Computer Engineering, faculty of engineering, Bilkent university, Ankara, Turkey, 06800, e-mail: S.rahimzadeh@cs.bilkent.edu.tr.
J. Kittler is with CVSSP, department of electronic engineering, faculty of engineering and physical sciences, university of Surrey, Guildford, Surrey, UK, GU2 7XH, e-mail: J.kittler@surrey.ac.uk.
parametric density estimation methods [13], [14], [15], neural-
network based approaches [16], [17], sparse representation
methods [18], [19], etc. Notable representatives of non-
generative methods include support vector-based approaches
[20], [21], convex hull methods [22], [23], cluster approaches
[24] and subspace techniques [25], [26], [27], [28], [29]. As
a result of the emphasis put on classification, rather than on
generative process modelling, the non-generative methods are
typically believed to yield better classification performance.

Among other non-generative methods, the kernel null-space
approach [30], [27], [31] is known to be a very successful
methodology for one-class classification. In particular, it
has been found to perform better than many state-of-the-art
one-class classification techniques by exploiting non-linear
projection function corresponding to the Fisher criterion for
classification. More specifically, in this methodology, a
discriminative feature subspace is inferred such that observations
from a specific class are mapped onto a single point, resulting
in a zero within-class scatter. On the other hand, a positive
between-class scatter is obtained since any other sample is
mapped onto another distinct point in the feature subspace.

Although the superiority of this approach over some other
alternatives has been confirmed in different studies [30], [27],
[31], nevertheless, in practice, it suffers from a number of
limitations. In the one-class learning paradigm it is natural
to expect that not all training samples equally adhere to
the model inferred. In extreme cases, a fraction of training
observations may be corrupted or even counter-examples. As
such, it is desirable to rank training observations according
to their conformity with the inferred model. A ranking of the
training samples may then allow one to enhance the perfor-
mance by discarding noisy observations from the training set
or using incompatible samples as counter-examples to refine
decision boundaries and thus enhance the robustness against
contamination in the training set. In addition, the ranking of
training data also facilitates setting an appropriate decision
threshold for the system to achieve (approximately) a desired
error rate on the test set. Furthermore, there exist applications
where one is only interested in ranking observations in a
given data set in a completely unsupervised fashion. However,
as in the kernel null-space methodology, all training samples
corresponding to a particular class are projected onto the
same point in the optimal feature subspace, there exists no
straightforward mechanism to gauge the compatibility of
(subsequently rank) training samples with the underlying
model. The sensitivity of the kernel null-space technique to
noisy observations in the training set, and their propensity to
seriously degrade its performance in practical settings, will be
demonstrated in the experimental evaluation section.

A. Overview of the proposed approach

This work addresses the aforementioned limitations of the
kernel null-space methodology by: 1-allowing the labels of
observations to vary and thus, in contrast to fixed hard labels,
to provide soft confidences so as to facilitate ranking of the
training data; 2-studying the effect of regularisation on the
regression framework and its ability to deal with training set
contamination; 3-refining the decision boundary of a one-class
classifier by automatically detecting outliers in the training set
and utilising them as counter-examples. Regarding the first
contribution, an alternating direction minimisation approach
is proposed where the discriminant function in the Hilbert
space and label confidences are optimised concurrently. The
second contribution is achieved by introducing two alternative
regularisation techniques, either based on Tikhonov regularisa-
tion or on sparsity, which are shown to improve the one-class
kernel system robustness against contaminations in the training
set. Finally, in terms of the third contribution, a variant of the
proposed approach is presented which makes use of any
detected non-target samples in the training set to refine the
decision boundary and enhance the system performance.

B. Outline of the paper

The rest of the paper is organised as follows: In Section II, a
review of related work is provided. A brief overview of the ker-
nel null-space technique and its regression-based formulation
is presented in Section III. The proposed robust kernel spectral
regression approach is introduced in Section IV. The results
of an experimental evaluation of the proposed methodology
are presented and discussed in Section V. Finally, conclusions
are drawn in Section VI.

II. RELATED WORK

Similar to the kernel null-space methodology, there exist
other approaches operating on a Rayleigh quotient criterion.
As an instance of the unsupervised methods in this category,
the work in [15] utilises kernel PCA for novelty detection
where the reconstruction residual of a test sample with respect
to the inferred subspace is considered as a novelty measure.
Other work [32] proposes a strategy to improve the conver-
gence of the kernel algorithm based on an iterative kernel
PCA. A robustified PCA which deals with outliers in the
training set is proposed in [33].

A one-class kernel Fisher discriminant classifier is presented
in [25], [34] which is based on the idea of separating the
data from their negatively replicated counterparts and involves
an eigenvalue decomposition of the kernel matrix. In this
approach, once the data are mapped onto the feature space,
a Mahalanobis distance to the mean of the fitted Gaussian
is used as a test statistic. As indicated in [34], for kernel
maps projecting input data into a higher-dimensional space,
the Gaussianity assumption for the target data may not hold in
general. An increasing deviation from normality may render
the results of the method presented in [25], [34] unreliable.
Another work in [27] proposed a Fisher-based null space
method where all training samples of one class are projected
onto a single point. The proposed method treats multiple
known classes jointly and detects novel instances with respect
to the set of classes using a single model operating on a joint
subspace where the training samples of all known classes are
presumed to have zero variance. In a follow-up work [35], it is
proposed to incorporate locality in the null space approach of
[27] by considering only the most similar patterns to a query
sample. In [31], an incremental version of the method in [27] is proposed to enhance the computational efficiency.

A generalised Rayleigh quotient specifically designed for outlier detection is presented in [28], [36] where the method tries to find an optimal hyperplane which is closest to the target data and farthest from the outliers utilising two scatter matrices corresponding to the outliers and target data. In [36], the generalised eigenvalue problem is replaced by an approximate conjugate gradient solution to moderate the computational cost of the method in [28]. A later study [37] tries to address the limitations of the method in [28], [36] in terms of the availability of outlier samples and a difference in the densities of target and non-target observations via the null-space approach.

While the majority of existing work on one-class classification based on the Rayleigh quotient formulation require computationally intensive eigen-decomposition of large matrices, the work in [30], presents a one-class approach which replaces costly eigen-analysis computations by a regression-based formulation [38]. Among other unsupervised novelty detection techniques, the DPCP approach [39] learns a linear subspace from data corrupted by outliers using non-convex $l_1$ optimisation. It is shown that DPCP can tolerate as many outliers as the square of the number of inliers, thus improving upon other robust PCA methods. Another method, known as Outlier Pursuite (OP), is an efficient convex optimisation-based algorithm [40] to perform a robust principal component analysis that under mild assumptions on the uncorrupted points recovers the exact optimal low-dimensional subspace and identifies the corrupted points. The FMS method is a non-convex robust subspace recovery approach [41], designed to be least affected by corruptions in the training set. It has been demonstrated to converge to a close vicinity of the correct subspace in few iterations. A leading unsupervised technique, among others, is that of SRO [42]. It obtains a weighted directed graph, defines a Markov Chain via self-representation, and identifies outliers via random walks.

III. BACKGROUND

A brief overview of the kernel Fisher null-space approach [30], [27], [31] and its regression-based formulation [30] is provided next.

A. Null-space Fisher analysis

In statistical pattern classification, a widely used design objective is the Fisher criterion where one tries to infer a projection function from the input space into a feature space such that the within-class scatter of the data associated with each class is minimised while maximising the between-class scatter. More specifically, in a Fisher classifier, the aim is to maximise the criterion function, $\mathbf{J}(\varphi)$,

$$\arg\max_{\varphi} \mathbf{J}(\varphi) = \arg\max_{\varphi} \frac{\varphi^\top \mathbf{S}_b \varphi}{\varphi^\top \mathbf{S}_w \varphi} \quad (2)$$

where $\mathbf{S}_b$ denotes the between-class scatter matrix, $\mathbf{S}_w$ denotes the within-class scatter matrix and $\varphi$ is a basis defining one axis of the subspace.

In a one-class classification problem and in the absence of non-target training observations, the origin may be used as a counter-example. With regards to the Fisher criterion, a theoretically optimal projection is the one yielding a zero within-class scatter while providing a positive between-class scatter, referred to as a null projection [30], [27], [31]. Thus, in a null-space Fisher classifier:

$$\varphi^\top \mathbf{S}_w \varphi = 0$$
$$\varphi^\top \mathbf{S}_b \varphi > 0 \quad (3)$$

A null projection function attains the optimum of $\mathbf{J}(\varphi)$ in Eq. 2 and thus provides the best separability with respect to the Fisher criterion. It can be shown that one may compute at most $C - 1$ null projection directions, with $C$ being the number of classes. In a one-class formulation, since target observations are assumed to form a single class while (hypothetical) non-target samples correspond to a second class, only a single optimiser for Eq. 2, given as the eigenvector corresponding to the largest eigenvalue of the following eigen-problem, exists:

$$\mathbf{S}_b \varphi = \lambda \mathbf{S}_w \varphi \quad (4)$$

Once the null projection direction is determined, the projection of a sample $x$ onto the null-space (hereafter referred to as response) is found as

$$y = \varphi^\top x \quad (5)$$

In order to handle data with inherently non-linear structure, non-linear (kernel) extensions of this methodology are proposed [30], [27], [31]. In kernel methods, a kernel function is utilised to implicitly project the data into a high dimensional space, known as the reproducing kernel Hilbert space (RKHS) in an attempt to make the data more easily separable in this new space. These methods typically require eigen-decompositions of dense matrices.

B. One-Class Kernel Spectral Regression

As eigen-computations associated with the kernel null-space technique are computationally demanding, an alternative approach based on spectral regression (called one-class kernel spectral regression, a.k.a. OC-KSR) was proposed in [30]. The OC-KSR method is based on two principles: (1) in order for the within-class scatter to be zero while achieving a positive between-class scatter, the observations corresponding to a specific class must be mapped onto the same point in the optimal feature subspace, distinct from all other classes, i.e. the responses (elements of $y$) corresponding to samples of a particular class must be equal, while being different from those corresponding to other class(es); and (2) in a RKHS, the function realising the above projection from the input space onto a feature subspace can be represented in terms of real numbers $\alpha_i$’s and a positive semi-definite kernel function $\kappa(.,.)$ as [30]

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i \kappa(\cdot, x_i) \quad (6)$$
where $x_i$’s denote training observations. In the OC-KSR approach [30], finding the optimal coefficients $\alpha_i$’s is then posed as a regression problem:

$$f(.)^{\text{opt}} = \arg\min_{f(.)} \left\{ \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \delta \|f\|_K^2 \right\}$$

(7)

or equivalently as

$$\alpha^{\text{opt}} = \arg\min_{\alpha} \left\{ \|K\alpha - y\|^2 + \delta \alpha^\top K\alpha \right\}$$

(8)

where $\delta$ is a regularisation parameter and $\|\cdot\|^2_K$ denotes the norm in the RKHS, while $K$ stands for the kernel matrix. The optimal solution $\alpha$ to the problem above satisfies

$$(K + \delta I_n)\alpha = y$$

(9)

where $I_n$ denotes an identity matrix of size $n$. In the OC-KSR method of [30], the kernel function is an RBF and thus the kernel matrix is positive definite. As the kernel matrix is invertible, no regularisation is imposed (i.e. the regularisation parameter $\delta$ is set to zero) on $\alpha$. Based on the availability of training data, two cases are considered in the OC-KSR method: 1-if only positive training samples are available, the target response vector $y$ is shown to be $y = (1, \ldots, 1)^\top$. When both positive and negative training observations are available, the target response vector $y$ is given as $y = (\frac{1}{n_0}, \ldots, \frac{1}{n_0}, 0, \ldots, 0)^\top$ (up to a scale factor) where without loss of generality it is assumed that the non-target instances are collected as the last $n_0$ samples of the training set. Given $y$, solving Eq. 9 for $\alpha$ is performed efficiently using the Cholesky factorisation via the Sherman’s march algorithm [43]. Once $\alpha$ is determined, a test sample may be projected onto the null feature subspace using Eq. 6. In the decision making stage, the (Euclidean) distance between the projection of a test sample and that of the target class is used as a dissimilarity criterion.

IV. ROBUST ONE-CLASS KERNEL SPECTRAL REGRESSION

In this section, the proposed approach to build a robust one-class classifier is based on the kernel null-space methodology in the presence of contaminations in the training set is presented. In the proposed method, starting from the given labels of the training samples, label confidences are updated iteratively and soft labels are assigned to the training observations, reflecting their degree of normality. As a result, although the proposed method is designed to detect novelty with respect to a given set of training samples, nevertheless, it can be directly used for observation ranking purposes in an unsupervised setting as will be discussed in the subsequent sections.

As noted earlier, the proposed approach is based on a regularised regression formulation, i.e.

$$P(\alpha, y) = \|K\alpha - y\|^2 + \mathcal{R}(\alpha)$$

(10)

where $y$ is the expected responses for observations and $\mathcal{R}(\alpha)$ encodes a desired regularisation of the solution $\alpha$. The regularisation, in general, may serve different purposes. First, when the number of variables exceeds the number of observations, the least-squares problem is ill-posed and it is therefore impossible to solve the associated optimisation problem as infinitely many solutions exist. Regularisation in this case allows the introduction of additional constraints that help to uniquely determine the solution. The second scenario, where regularisation may be deployed, corresponds to the case where the number of variables does not exceed the number of samples, but the model learned suffers from poor generalisation capability. In such cases, regularisation is used to improve the generalisation performance of the model by constraining it during the training phase. The regularisation term thus imposes a limitation on the function space available by introducing a penalty to avoid certain regions of the function space. In the current work, given sparse and noisy samples of a function $f(.)$ from a corrupted data set, regularisation constraints the function and maintains a trade-off between data fidelity and some constraint on the solution function. The imposed constraint may, for example, enforce the solution to be sparse or to reflect other prior knowledge about the solution such as constraining its norm in the corresponding space. In such cases, regularisation methods typically correspond to priors on the solution to a least squares problem.

Following a regularised regression formulation, the proposed approach to handle contaminations in the training set is to optimise the objective function $P$ not only with respect to $\alpha$ but also with respect to $y$. While optimisation with respect to $\alpha$ is the standard approach to specify the parameters characterising the projection function given by Eq. 6, optimisation with respect to $y$ reflects (and compensates for) the absence of prior knowledge regarding the conformity of individual observations to the model. This is fundamentally different from the ordinary OC-KSR method [30] as well as other flavours of the kernel null-space technique [27], [31] where $y$ is fixed and $\alpha$ optimisers of $P$ are sought. Moreover, the method in [30] does not impose a regularisation on the model when deriving the final solution (i.e. $\delta$ is set to zero). In essence, given an initial guess regarding object labels (i.e. the responses $y$), the proposed approach updates label confidences and derives soft responses for all training samples, while at the same time optimising the objective function with respect to $\alpha$.

This is realised via a block coordinate descent minimisation approach, alternating between minimising $P$ with respect to $\alpha$ and minimising it with respect to $y$.

Optimising $P$ with respect to $y$ can be realised by setting its partial derivative with respect to $y$ equal to zero, i.e.

$$\frac{\partial P}{\partial y} = -2(K\alpha - y) = 0$$

(11)

which gives $y = K\alpha$. As such, the generic scheme of the proposed approach can be summarised as Algorithm 1. It can be shown that the alternating minimisation scheme of Algorithm 1 is convergent. This can be readily confirmed by denoting $\alpha_{t+1} = \arg\min_{\|\alpha\|_1} P(\alpha,y_t)$ and $y_{t+1} = \arg\min_y P(\alpha_{t+1},y)$ for iteration $t+1$. Consequently, one has

$$P(\alpha_t, y_t) \geq P(\alpha_{t+1}, y_t) \geq P(\alpha_{t+1}, y_{t+1})$$

(12)

The regularisation types considered in this work are those of Tikhonov and sparsity. As a result, the objective function
Algorithm 1 The generic scheme of the proposed approach
1: repeat
2: \( \alpha = \arg\min_{\|\alpha\| = 1} P(\alpha, y) \)
3: \( y = K\alpha \)
4: until convergence

Algorithm 2 The proposed approach based on Tikhonov regularisation
1: repeat
2: \( \alpha = (K + \delta I_n)^{-1}y \)
3: \( \alpha = \alpha / \|\alpha\| \)
4: \( y = K\alpha \)
5: until convergence

\( P(\alpha, y) \) is bounded from below and thus the non-increasing sequence \( \lim_{t \to \infty} P(\alpha_t, y_t) \) is convergent.

Solving the minimisation problem \( \alpha^{opt} = \arg\min_{\|\alpha\| = 1} P(\alpha, y) \) is dependent upon the specific regularisation imposed on the solution. Although other possibilities exist, in this work, two commonly used regularisation schemes of Tikhonov and sparsity are considered for \( R(\alpha) \), discussed next.

A. Tikhonov regularisation

In the case of a Tikhonov regularisation (also known as ridge regression), \( R(\alpha) = \delta \alpha^{\top}K\alpha \) and the objective function \( P(\alpha, y) \) would be

\[
P(\alpha, y) = \|K\alpha - y\|^2 + \delta \alpha^{\top}K\alpha \tag{13}\]

In this case, the problem \( \alpha^{opt} = \arg\min_{\|\alpha\| = 1} P(\alpha, y) \) can be solved by setting the partial derivative of \( P(\alpha, y) \) with respect to \( \alpha \) to zero, i.e.

\[
\frac{\partial P}{\partial \alpha} = 2K(K\alpha - y) + 2\delta K\alpha = 0 \tag{14}
\]

which gives

\[
\alpha^{opt} = (K + \delta I_n)^{-1}y \tag{15}
\]

Combining the two equations for \( \alpha \) and \( y \), the proposed approach based on a Tikhonov regularisation is given as Algorithm 2. In a way, Tikhonov regularisation favours models that provide predictions that are as smooth functions of the data as possible. In other words, such a regularisation scheme penalises coefficients of larger magnitude, thereby producing a more parsimonious solution, incorporating a set of coefficients with smaller variance, which is particularly advantageous when making inference in a noisy data set.

1) Optimal regularisation parameter: Algorithm 2 provides a procedure to find the optimal parameters \( \alpha \) and \( y \). Yet, it does not specify how to choose the Tikhonov regularisation parameter \( \delta \). In order to infer the optimal Tikhonov regularisation parameter, the one-class learning problem in the presence of outliers is posed as a sensitivity analysis problem in this work. In this respect, the optimal Tikhonov regularisation parameter is derived so as to minimise the sensitivity of the solution \( \alpha \) with respect to contaminations in the training set. For this purpose, let us consider a set of contaminated observations \( X = \{x_1, \ldots, x_n\} \) for which the true labels are recorded in vector \( y \). It is assumed that we are not informed of the true labels in the contaminated data set. Instead, an initial guess for object labels may be assumed. In the absence of any prior knowledge, the initial assumption for all observations (recorded in vector \( y' \)) may be that of being target samples, i.e. \( y' = \{1, \ldots, 1\} \). \( y' \) corresponds to the noisy assumption of labels which deviates from the true labels by \( \Delta y \), i.e. \( y' = y + \Delta y \). Using the assumed noisy labels \( y' \), the weight vector \( \alpha' \) for the Tikhonov regularised problem is given as \( \alpha' = (K + \delta I)^{-1}y' \). In practice, it is desirable to have \( \alpha' \) as close as possible to the true \( \alpha \) (derived based on \( y \)) so that the responses obtained as \( K\alpha' \) are close to the true responses \( y = K\alpha \). This is essentially a sensitivity analysis problem where one is given a perturbed response vector \( y' = y + \Delta y \) and the goal is to find a solution \( \alpha' \) with a minimal difference from the ideal solution \( \alpha \). Assuming \( ||\alpha'|| > 0 \), the sensitivity of the solution with respect to perturbations in object labels is defined as

\[
S = \frac{||\alpha' - \alpha||}{||\alpha'||} \tag{16}\]

Apparently, minimising the sensitivity of the solution would maximise the similarity between the true labels \( y \) and the inferred responses \( y' \).

The sensitivity of the regression solution in the Hilbert space using a Tikhonov regularisation in a general setting is studied in [44] and summarised in the Theorem below and the following Corollary.

**Theorem** Let \( \kappa(\ldots) \) be a kernel function and let the kernel matrix \( K \) corresponding to a set of observations \( X \) be positive definite. Then the optimal value of parameter \( \delta > 0 \) with respect to the sensitivity of the regularised solution of the problem \( E_Z + \delta ||\cdot||_K^2 \) is

\[
\delta_{opt} = \frac{\lambda_{min}(c(K)) - c(K) + 1}{2\sqrt{c(K)}} \tag{17}
\]

where \( E_Z = \frac{1}{n}||K\alpha - y||^2 \)

\( \lambda_{min} \) denotes the smallest eigenvalue of the kernel matrix \( K \) and \( c(.) \) stands for the condition number of a matrix, defined as the ratio of the largest eigenvalue to the smallest one.

**Corollary** In addition to the requirements of the Theorem above, if the kernel matrix is normalised, then the optimal value of parameter \( \delta > 0 \) with respect to the sensitivity of the regularised solution of the problem \( E_Z + \delta ||\cdot||_K^2 \) is given as

\[
\delta_{opt} = \frac{1}{1 + \lambda_{min}} - \frac{\lambda_{min}(2 - \sqrt{\lambda_{min}})}{2} \tag{18}
\]

For a proof of the theorem and the corresponding corollary cf. [44].

As the kernel function used in this work is that of a radial basis function, the kernel matrix would be positive definite. Consequently, for a minimum sensitivity solution in the proposed approach, one may follow Algorithm 2, setting the regularisation parameter based on Eq. 18 for a normalised kernel matrix.
B. Sparse regularisation

In addition to the widely used Tikhonov regularisation, other regularisation approaches encouraging sparseness of the solution have a rich history as a guideline for inference. The underlying motivation for seeking a sparse characterisation is the desire to provide the simplest possible explanation of an observation as a linear combination of as few as possible atoms from a given dictionary. One of the most celebrated instantiations of sparseness, the principle of minimum description length in model selection, requires that within a pool of models, the model that yields the most compact representation should be preferred for decision making. Such sparse methods select a small subset of relevant atoms to characterise the solution.

A sensible sparsity constraint to impose on the solution of the kernel spectral regression approach is the $l_0$-norm, defined as the number of non-zero elements in the solution. However, solving an $l_0$-regularised learning problem has been demonstrated to be NP-hard. The $l_1$-norm is shown to induce sparsity and can be used to approximate the optimal $l_0$-norm via convex relaxation. A least squares problem in the presence of an $l_1$ regularisation term is known as Lasso in statistics, and a basis pursuit in signal processing. Sparse $l_1$-norm models allow for scalable algorithms that can handle problems with a large number of parameters. Encouraging the solution of the kernel null-space approach to be sparse can conveniently be performed by enforcing an $l_1$-regulariser, i.e. $R = \delta \sum_{i=1}^{n} |\alpha_i|$. Consequently, the objective function $P(\alpha, y)$ in this case would be

$$P(\alpha, y) = \|K\alpha - y\|^2 + \delta \sum_{i=1}^{n} |\alpha_i|$$ (19)

The degree of sparseness of a solution is controlled via parameter $\delta$. In a sparse formulation of the regression problem, each response $y_i$ is generated using only a few observations from the training set. In particular, if the solution $\alpha$ is very sparse, a large number of observations would have no contribution to the final solution, the immediate implication of which is a reduction in the computational complexity of the algorithm in the test phase. A further and more important consequence of forming a sparse model, as also suggested by other studies [45], is that of classification performance where a more compact model could improve performance compared with its non-sparse counterpart, especially in the presence of corrupted data.

Solving lasso problems efficiently is a subject of ongoing and fast developing research. While efficiency of solving the lasso problem is desirable, as this stage of the method is presumed to be performed offline, it has no impact on the efficiency of the proposed approach in the deployment phase. The complexity of the method in the test phase is governed by the degree of sparsity of the solution, i.e. $\alpha$ and not by the procedure deriving such a solution. In this work, the Least Angle Regression (LARS) algorithm [46] is used to find the optimal solution corresponding to $\alpha = \arg \min_{\alpha} \|K\alpha - y\|^2 + \delta \sum_{i=1}^{n} |\alpha_i|$. Using the LARS algorithm, solutions with all possible cardinalities on $\alpha$ can be computed. For the lasso formulation, the proposed approach is given in Algorithm 3.

C. Known fraction of contaminations

In the discussions thus far, it is assumed that the fraction of the data contaminated in the training set is not known in advance. In case the number of contaminations in the training set (denoted as $n_0$) is known, this information can be incorporated into the learning algorithm. This is realised by ranking training observations according to their compatibility with the model and identifying $n_0$ least compatible observations as non-target samples which results in the splitting of the training set into a target set and a non-target set. For this purpose, at each iteration, the responses for all observations are sorted and the smallest $n_0$ elements of $y$ are set to zero, while others are set to one. This procedure essentially corresponds to updating the initial normality assumption regarding observations, which is expected to improve the performance. In this respect, the samples corresponding to the $n_0$ samples which are less likely to be positive observations would form a non-target set, which is then used for the refinement of the solution. The generic learning scheme thus obtained is given as Algorithm 4 where the $SR(.)$ routine corresponds to a sorting and updating of the elements of an argument vector as described above. The minimisation step $\alpha = \arg \min_{|\alpha| = 1} P(\alpha, y)$ in Algorithm 4 is performed as discussed previously, depending on the specific type of regularisation employed.

D. Decision strategy

Once the optimal projection parameter $\alpha$ is inferred, the projection of a test sample onto the feature subspace is realised as per Eq. 6. The projections of target samples in the proposed formulation are expected to lie at points closer to point 1 in the feature subspace while those corresponding to non-target samples are expected to lie at points closer to the origin. Consequently, the decision rule for a test sample $z$ is defined as

$$f(z) = \sum_{i=1}^{n} \alpha_i \kappa(z, x_i) \geq \tau \quad z \text{ is a target object}$$

$$f(z) = \sum_{i=1}^{n} \alpha_i \kappa(z, x_i) < \tau \quad z \text{ is an outlier}$$ (20)

where $\tau$ is a threshold for deciding normality.
V. Experimental Evaluation

In this section, an experimental evaluation of the proposed approach along with a comparison to the state-of-the-art methods is presented. On each data set, the training data includes both positive and negative instances to simulate a contaminated training set.

A. Data sets

1) Face: This data set is created to perform a toy experiment in face recognition. The data set contains face images of different individuals and the task is to recognise a subject among others. For each subject, a one-class classifier is built using the training data associated with the subject under consideration while all other subjects are assumed as outliers with respect to the built model. The experiment is repeated in turn for each subject in the dataset. The features used for face image representation are obtained via the frontal-pose PAM deep CNN model [47] applied on the face bounding boxes. The data set is created out of the real-access videos of the Replay-Mobile dataset [48] which provides face bounding boxes. In this work, ten subjects are used to form the data set where each subject is represented using 30 positive training instances. The number of negative training observations for each subject, i.e. contaminations, is also 30 images selected randomly from subjects other than the subject under consideration. The number of positive and negative test samples are similarly set to 30 images each.

2) MNIST: MNIST is a collection of $28 \times 28$ pixel images of handwritten digits 0-9 [49]. In our experiments, a single digit is considered as the target digit while all others correspond to non-target observations. In the experiments on this data set, the number of positive and negative training instances are set to 50 images each. Similarly, 50 positive and 50 negative images are included in the test set. The target class is assumed to be digit ‘3’ while all other digits represent anomalies with respect to the target class.

3) Coil-100: The Coil-100 data set [50] contains 7,200 images of 100 different objects. Each object has 72 images taken at pose intervals of 5 degrees, with the images being of size $32 \times 32$ pixels. In the experiments conducted on this data set, a single object is randomly chosen to be the target class while all others are considered as novelties. Raw pixel intensities are used as feature representations in this data set. The number of positive train and test instances for the target class is 36 each. Similarly, 36 negative train and 36 negative test observations are included in the experiments on this data set.

B. Convergence behaviour

In this section, the convergence behaviour of the proposed block coordinate descent optimisation method is analysed. For this purpose, the proposed models are trained on the face, MNIST and Coil-100 data sets. The error, defined as the norm of the deviation in $\alpha$ between two consecutive iterations, vs. number of iterations of the corresponding algorithms is plotted. The results are averaged over ten random splits of the data into training and test sets. A zero norm deviation would be indicative of the convergence of the optimisation algorithm. The mean and std. (shaded regions in the figure) of the error vs. number of iterations are given in figures 1, 2 and 3 for the face, MNIST and Coil-100 data sets, respectively. As can be observed from the figures, on all three data sets, the algorithm, on average, converges faster with the Tikhonov regularisation as compared with the sparse model. Nevertheless, regardless of the regularisation type imposed on the solution, convergence is quite fast. The methods converge in just a few iterations (on average in as few as five iterations).

C. The effect of regularisation parameter

In this section, the effect of regularisation parameters on the performance of the proposed methodology is evaluated. The contamination is increased from 5% to 50% in steps of 5%. Regarding the ridge regression formulation, the optimal Tikhonov regularisation parameter along with three other parameters and the case where no regularisation is imposed, are considered. For the sparse model, the degree of sparseness is varied from 50% to 90% in steps of 10% where $n\%$ sparsity corresponds to the case where $n\%$ of variables in $\alpha$ are zero. The results corresponding to this experiment for the Tikhonov regularised formulation and the sparse model on the three data sets are presented in Fig. 4 and 5, respectively. The plots
present the average AUC’s over ten random splits of data into the training and test sets.

Regrading the Tikhonov regularised model, the worst performance corresponds to the case where no regularisation is applied. Increasing the regularisation parameter from zero towards the optimal value gradually enhances system robustness on all three data sets. Regarding the proposed sparse model, the sparser the solution towards the 90% sparsity level, the better the robustness of the method is. This behaviour is confirmed on all three data sets where a maximum sparseness of 90% achieves the best average overall performance over the entire range of contaminations. A common behaviour for both types of regularisation is that increasing the regularisation effect improves robustness against contaminations towards higher levels of corruption at the cost of a small decrease in the performance for the lower percentages of contaminations. Nevertheless, the best average performance over the entire range of contaminations is obtained for the Tikhonov regularised model with the optimal regularisation parameter and the sparse model set to a 90% sparsity level, indicating the effectiveness of regularisation in improving the robustness of the kernel null-space approach.

D. One-class classification in the presence of contaminations in the training set

In this experiment, the proposed regularised null-space kernel spectral regression methods are evaluated on a one-class classification problem where the model is trained on a set of observations whose contamination is varied from 10% to 50% and then evaluated on a separate set of test samples. Such an evaluation scheme is commonly referred to as a supervised one-class classification [1] where it is assumed that the training data incorporates labelled instances only for the normal class. In this experiment, the proposed methodology is compared against several other methods including:

- **Tikh** is the proposed robust spectral regression approach using a Tikhonov regularisation term.
- **Spar** is the proposed robust spectral regression approach using a sparsity inducing regularisation term.
- **Org** corresponds to the one-class kernel null Foley-Sammon transform. The methods presented in [30], [27], [31] theoretically obtain the same result and only differ in terms of computational complexity. As the proposed approach is based on a kernel null-space formulation, the methods in [30], [27], [31] serve as a baseline to gauge improvements obtained using the proposed approach.
- **SVDD** is the Support Vector Data Description approach to solve the one class classification problem [20]. As a widely used method, it serves as a second baseline for comparison.
- **GP** is derived based on the Gaussian process regression and approximate Gaussian process classification [51] where in this work the predictive mean is used as the one class score.
- **K-m** is the k-means clustering-based approach where k centres are assumed for the target observation. The novelty score of a sample is defined as the minimum distance of a query to cluster centres. The optimum value for parameter k is experimentally set to 5 to achieve the best average performance over all data sets.
- **FB** corresponds to the feature bagging algorithm [52] which detects outliers by bagging anomaly scores where the scores are generated by different individual outlier detection models operating on random subsets of input features. For each outlier detection model, the abnormal score is derived using a small subset from the set of
available features. The base outlier detection method is that of the Local Outlier Factor [53].

- **Pa.** (Parzen) corresponds to the non-parametric density estimation approach based on Parzen-window estimators with Gaussian kernels [54] and works by sliding a window centred at each sample, and computing a probability density function for each label. Samples for which the maximum probability is less than a threshold are identified as outliers.

1) Implementation details: For the Org method, the implementation corresponding to the work in [30] is employed as it is the most efficient among other variants of the kernel null-space technique. The SVDD method is based on the implementation provided in the data description toolbox [55]. The GP method is based on the implementation available publicly 1. Regarding the FB and Parzen approaches the implementations provided in a supplement website 2 are utilised. The K-means clustering approach is based on its implementation in Matlab 2017b. No pre-processing is applied on the features other than normalising them to have a unit $l_2$-norm.

In the experiments on each data set, the data is randomly divided into the train and test sets and the percentage of contaminations is increased from 10% to 50% of the total training data composed of positive and negative samples. Note that a 50% contamination in the training set represents a quite high degree of data corruption. The performance is then reported as the average AUC’s over 10 repetitions of random splitting of the data into the train and test sets.

2) Results on the face data set: The results in terms of average AUC’s over ten random splitting of the data into the train and test sets on the face data set are plotted in Fig. 6 over all subjects. The average AUC’s (in percentage) for different subjects over the whole range of contaminations (from 10% to 50%) are reported in Table I. The corresponding AUC plots for each individual subject are provided as supplementary material.

From the figures and the table, it can be observed that the
best performing method, on average, over the full range of contaminations is the proposed approach based on Tikhonov regularisation. Moreover, in terms of the average performance over all subjects, the second best performing method is that of the proposed sparse approach. The improvement obtained using the proposed approach over the baseline method [30], [27], [31] over the entire range of contamination is more than 3% on average using a Tikhonov regularisation and over 2% on average using a sparse regularisation. However, as the percentage of contaminations in the training set increases, the effectiveness of the proposed regularisation-based approach becomes more evident reaching more than 10% of improvement in terms of AUC for a 50% contamination in the training set.

3) Results on the MNIST data set: The results corresponding to different methods on the MNIST data set are provided in Fig. 7. From the figure it can be observed that the most robust method against contaminations in the training set is the proposed approach based on Tikhonov regularisation. The superiority of this formulation over other alternatives becomes more evident as more contaminations are included in the training set. The second best performing method, on average, similar to the experiments on the face data set is again the proposed approach using a sparse regularisation term.

4) Results on the Coil-100 data set: The results corresponding to different methods on the Coil-100 data set are presented in Fig. 8 in terms of average AUC’s. From the figure it can be observed that, interestingly, the two best performing methods are those of Parzen window and the proposed approach based on Tikhonov regularisation. Nevertheless, the proposed sparse model is not much worse. The results corresponding to the MNIST and Coil-100 data set are also summarised in Table II. As can be observed from the table, the two best performing methods on the MNIST data set are the proposed methods based on Tikhonov and Sparse regularisation. In particular, the proposed approach based on the Tikhonov regularisation, on average, is nearly 5% better than the original kernel null-space method [30], [27], [31]. On the Coil-100 data set, interestingly, the Parzen window method achieves the best average performance followed by the Tikhonov regularisation-based approach.
versions Tikh and Spar on all the three data sets examined. Second, the improvements obtained by Tikh+ and Spar+ are monotonically increasing with the increasing number of non-target observations in the training set. This shows that, in the proposed methodology, non-target samples can play an effective role. Third, it can also be observed that, interestingly, on the MNIST data set the additional information in terms of number of negative samples in the training set may even boost the performance of the Spar+ method. This is reflected in the average AUC of Spar+ method corresponding to a 50% corruption being higher than that of a 10% corruption in the data set.

2) Unknown fraction of outliers: In this section, we examine the possibility of automatically estimating the fraction of contamination from the data. The approach is based on the expectation that if the training data is contaminated, but we assume no contamination, a good fit would not be obtained. Similarly, if one assumes too much contamination, normal data will start mixing with negative samples resulting in a sub-optimal fit. In this respect and in order to estimate the contamination percentage, we optimise the objective function $P$ for different hypothetical values of contamination $N_0$ and measure the normalised error as $\|K\alpha - y\|^2/N_0$ where $y$ is also normalised to have a unit norm. The value of $N_0$ which gives the minimum error is then considered as an estimate of the true $n_0$.

Given an estimate of $n_0$, the rest of the procedure is similar to that of section V-E1. Applying the above idea to the Tikhonov-regularised approach (denoted as A-Tikh) a similar set of experiments to that of section V-E1 is performed including the A-Tikh approach. The results of the comparison of different variants of the Tikhonov and sparse regularised methods on the face data set are presented in Fig. 12. One may observe that the A-Tikh approach performs better than Tikh and Spar, where no automatic mechanism for detecting outliers is available. Moreover, the A-Tikh performs better than the Spar+ method which uses the knowledge of $n_0$. The effectiveness of the A-Tikh method is accentuated by its almost similar performance to Tikh+ for which the percentage of contaminations is provided as an input parameter. For the space constraints the results are shown only for the face data. However, the same behaviour was observed on the other data sets too.

F. Unsupervised observation ranking

In a final set of experiments, the proposed regularised kernel spectral regression methods are evaluated in an observation ranking paradigm. As noted earlier, the original null-space kernel one-class methods are unable to provide a ranking of training observations. Nevertheless, this limitation is removed in the proposed regularised methods by imposing an additional regularisation term while at the same time updating observation label confidences via an iterative alternating minimisation technique. The methods included in the comparison in this set of experiments are the state-of-the-art methods for unsupervised observation ranking and outlier detection. Consistent with the literature [1], unsupervised methods refer to those approaches which do not require training data and make the implicit assumption that normal instances are more frequently present than anomalies in a data set. Consequently, the methods included in this experiment are those which are specifically designed to operate on a given set of contaminated samples and provide a compatibility ranking. As such, different methods in this experiment are utilised to rank observations in the training set only. The methods compared in this experiment are:

- Tikh
- Spar

<table>
<thead>
<tr>
<th>Method</th>
<th>Tikh</th>
<th>Spar</th>
<th>Tikh+</th>
<th>Spar+</th>
<th>Org</th>
<th>SVDD</th>
<th>SVDD+</th>
</tr>
</thead>
<tbody>
<tr>
<td>face</td>
<td>96.80</td>
<td>95.78</td>
<td>97.17</td>
<td>96.40</td>
<td>93.67</td>
<td>93.84</td>
<td>93.51</td>
</tr>
<tr>
<td>MNIST</td>
<td>87.06</td>
<td>83.77</td>
<td>88.75</td>
<td>89.80</td>
<td>82.07</td>
<td>82.31</td>
<td>77.53</td>
</tr>
<tr>
<td>Coil-100</td>
<td>99.34</td>
<td>97.37</td>
<td>99.52</td>
<td>99.33</td>
<td>97.34</td>
<td>94.93</td>
<td>97.52</td>
</tr>
</tbody>
</table>
DPCP is a method for learning a linear subspace from data corrupted by outliers based on a non-convex $l_1$ optimisation [39]. It is shown that DPCP can tolerate as many outliers as the square of the number of inliers, thus improving upon other robust PCA methods.

OP is an efficient convex optimisation-based algorithm [40] to perform a robust principal component analysis that recovers the exact optimal low-dimensional subspace under mild assumptions and identifies the corrupted points.

FMS is a non-convex robust subspace recovery approach [41], designed to be least affected by corruptions in the training set and has been demonstrated to converge to a close vicinity of the correct subspace within few iterations.

SRO obtains a weighted directed graph, defines a Markov Chain via self-representation, and identifies outliers via random walks [42]. The SRO method can be considered as one of the leading unsupervised approaches for novelty detection.

The results, in terms of average AUC, corresponding to this experiment are provided in Fig. 13, 14 and 15 for the face, MNIST and Coil-100 data sets, respectively and also summarised in Table IV. From the table and the figures, it can be observed that the Tikhonov regularisation-based approach is the top performer among other competitors on all three data sets in an unsupervised observation ranking and novelty detection scenario. The proposed sparsity-based approach performs less well compared to the Tikhonov regression based formulation. Nevertheless, it still performs better than some other competitors including the recently proposed DPCP and FMS approaches.

### VI. Conclusion

The problem of one-class classification in a kernel Fisher null-space framework is studied. Two limitations of the null-space kernel Fisher analysis, namely its sensitivity to training set contamination and its inability to rank training samples, are addressed. We propose regularised variants of a regression-based formulation of the learning problem (Tikhonov and sparsity) where both projection parameters and object labels are inferred iteratively via an alternating direction minimisation approach. Through experiments on different data sets, it was demonstrated that: 1- the proposed regularisation approach combined with the alternating optimisation mechanism is effective in robustifying the baseline method; 2-the proposed iterative ridge regression-based formulation posing one-class learning as a sensitivity analysis problem is the top performer among other competitors; and 3-the performance of both alternative regularisation schemes is boosted by automatically detecting negative samples in the training set and forming a subset of counter-examples for enhanced training. While the Tikhonov-based regularisation provides superior performance as compared with its lasso counterpart, the sparse regularisation-based method provides computational advantages in the test phase where evaluation of a test sample incorporates only a small fraction of training samples (in the order of 10% of total training data).

The proposed methodology has been evaluated in a one-class classification paradigm by assessing its generalisation capability as well as in an observation ranking scheme to detect outliers in a given data. It has been found to perform better than the baseline method while providing very competitive performance compared to the state-of-the-art techniques.

### ACKNOWLEDGMENT

This work was supported in part by the EPSRC Programme Grant (FACER2VM) EP/N007743/1, EPSRC/dstl/MURI...


Shervin Rahimzadeh Arashloo received the Ph.D. degree from the centre for vision, speech and signal processing, university of Surrey, UK. He is an assistant professor with the Department of Computer Engineering, Bilkent University, Ankara, Turkey and also holds a visiting research fellow position with the centre for vision, speech and signal processing, university of Surrey, UK. His research interests includes secured biometrics, anomaly detection and graphical models with applications to image and video analysis.

Josef Kittler received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He published the textbook Pattern Recognition: A Statistical Approach (Englewood Cliffs, NJ, USA: Prentice-Hall, 1982) and over 600 scientific papers. He serves on the Editorial Board of several scientific journals in pattern recognition and computer vision.