Discriminative Features for Identifying and Interpreting Outliers

Xuan Hong Dang, Ira Assent
Department of Computer Science
Aarhus University
Aarhus, Denmark
{dang,ira}@cs.au.dk

Raymond T. Ng
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada
rng@cs.ubc.ca

Arthur Zimek, Erich Schubert
Institut für Informatik
Ludwig-Maximilians-Universität München
Munich, Germany
{zimek,schube}@dbs.ifi.lmu.de

***Abstract*—We consider the problem of outlier detection and interpretation. While most existing studies focus on the first problem, we simultaneously address the equally important challenge of outlier interpretation. We propose an algorithm that uncovers outliers in subspaces of reduced dimensionality in which they are well discriminated from regular objects while at the same time retaining the natural local structure of the original data to ensure the quality of outlier explanation. Our algorithm takes a mathematically appealing approach from the spectral graph embedding theory and we show that it achieves the globally optimal solution for the objective of subspace learning. By using a number of real-world datasets, we demonstrate its appealing performance not only w.r.t. the outlier detection rate but also w.r.t. the discriminative human-interpretable features. This is the first approach to exploit discriminative features for both outlier detection and interpretation, leading to better understanding of how and why the hidden outliers are exceptional.**

I. Introduction

Outlier identification is a key problem for many practical applications. Unlike other data mining tasks such as clustering, classification, or frequent pattern analysis which aim to find popular patterns, outlier detection is to capture a small set of objects that deviate significantly from the larger number of common objects in a data set. Mining and interpreting that kind of inconsistent patterns poses particular challenges and issues. The difficulty often lies in the fact that the population of outliers is small, compared to the number of regular objects, limiting the learning capability of most algorithms. It is also very hard to precisely define, quantify and interpret the notion of "significant deviation" of a data object, especially in high dimensional spaces.

The problem of outlier detection in various application domains attracted significant research effort [11]. Unfortunately, most existing studies focus on the problem of outlier detection only but often ignore the equally important problem of outlier explanation. In many application scenarios, the user is not only interested in detecting outliers but would like to gain deeper insights of *why* these are exceptional w.r.t. the other, regular, objects (or inliers). Generally, an anomaly degree of an object can be considered as a first step toward outlier explanation since this piece of information shows how likely the object was generated by the same mechanism as the majority of the data. Yet, from the viewpoint of a practical user, such a numerical score provides limited information, especially for high dimensional data, as it is lacking the information in which data view the object is most exceptional. In terms of interpretation, we thus claim that an identified outlier should be explained clearly in a compact view, as a succinct subset of original features, that shows its exceptionality. This type of knowledge, obviously, not only helps the domain experts, who often have little or no expertise in data mining, to validate the practical existence of the discovered outliers but also further improves their understanding of the data. An important question is thus how to extract a small number of relevant features that can be used to explain the exceptional properties of an outlier, without falling trap to the "data snooping bias" [37]. This means, by observing the data from these selective features, an outlier must be well *discriminated* from common inliers. It is worth noting that, although several algorithms for outlier detection in subspace projections have been developed recently, they all attempt to explore descriptive feature subspaces of regular inliers to facilitate the process of computing deviation (i.e., anomaly) degrees for outliers. They are hence less successful in uncovering the most discriminative features to distinguish outliers from regular patterns.

In order to illustrate the importance of uncovering the most discriminative features, let us consider a motivating real-world example using a set of images from the AR Face database [28], shown in Figure 1. Intuitively, two images at the outer left and outer right are anomalous. By analyzing the descriptive features of these images, a method based on principle component analysis (PCA), for example, can show that the images all share the same descriptive features as characterizing for the same person and two outliers are the ones deviating most from these features. However, by looking for the discriminative features, it is possible to show that the outer left image is an outlier based on the features positioned at the mouth and central face areas whereas the outer right image is an anomaly due to the appearance of the glasses around the eye features. Notice that these two outliers share similar nearby inliers; yet the features identifying them as exceptional are clearly different in both cases. Therefore, solely relying on the descriptive

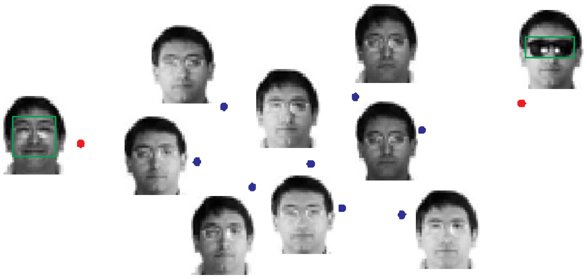

Figure 1. The outer left image is an outlier based on the discriminative features around the central face area whereas the outer right one is an outlier relied on the discriminative features around the eye area. Images from the AR face database [28].

features of majority patterns may not help the user to fully understand why the objects are anomalous. Instead, exploring the discriminative features would yield the best features justifying the distinctiveness of outliers.

In this paper, we present a novel algorithm for exploring and characterizing local outliers in high dimensional numerical datasets. Our algorithm is not only able to provide a ranking list over the outlier degrees of data objects but also discovers a succinct subset of discriminative features to explain why an outlier is exceptional compared to the majority patterns. The proposed technique takes a graph embedding approach in which a neighborhood graph is constructed to model the geometrical structure of the underlying data distribution. For learning a subspace in which an outlier candidate can be discriminated from regular objects, our algorithm exploits: (i) the local graph connections between the outlier candidate and its nearby inliers; and (ii) the graph connections within the set of its neighboring inliers. Using the notion of graph Laplacian combined with the L_2-norm shrinkage, we show that the algorithm can learn an optimal projection that transforms both the outlier candidate and its nearby inliers to a reduced dimensional subspace of which the outlier is optimally discriminated from the neighbors (by maximizing (i)) while at the same time retaining the natural structure of its neighboring inliers (by minimizing (ii)) to ensure the quality of outlier explanation. The induced subspace thus provides all essential information for the anomalous properties of the outlier. The proposed technique possesses many appealing properties: (1) it makes no specific assumptions on the statistical distribution regarding the original data structure; (2) it has a solid mathematical background; (3) it guarantees the solution for the induced subspace to be globally optimal. Through experimental analysis over a number of real world datasets, we demonstrate the appealing performance of our proposed algorithm not only over the outlier ranking quality against most well-known algorithms, but also over the set of discriminative human-interpretable features. To our knowledge, the proposed algorithm is the first that exploits discriminative features for both outlier detection and interpretation, providing more insights into outliers and thus leading to a better understanding of why the outliers are exceptional patterns.

II. Related work

According to the variety of application domains for outlier detection, such as bioinformatics, direct marketing, or various types of fraud detection, many algorithms have been developed to deal with this problem. Depending on whether the labels for outliers (or inliers) are available or not, the algorithms can be classified as supervised [35], semi-supervised [15], or unsupervised [9] techniques. Alternatively, w.r.t. the analyzed data types, they can be categorized as dealing with spatial temporal data [12], structured graph data [4], transactional/categorical data [33], or numerical data [26]. Here, we focus on the *unsupervised* setting on *numerical* high dimensional datasets [37].

The seminal database-oriented method DB-outlier [21] requires two parameters, distance d and data fraction p. An object is considered an outlier if at least a fraction p of all instances have a distance $> d$. In other studies [30], a similar definition is used where an object is viewed as an outlier if its distance to the k-th nearest neighbor is sufficiently large, usually greater than a given threshold d. It can be seen that both definitions are related and that the identified anomalous objects are called *global* outliers since their properties w.r.t. the given thresholds (e.g., d, p) are compared with all other objects of the data set. In contrast, so-called “local” techniques seek *local* outliers, whose outlier degrees are defined w.r.t. their neighborhoods rather than w.r.t. the entire data set [32]. In the LOF model [9], the (relative) density around an object is loosely estimated as the inverse of the average distances from the object to its nearest neighbors and its local outlier factor is calculated based on the ratio between its density and the average density computed from all its k neighbors. There are several studies attempting to find outliers in spaces with reduced dimensionality [37]. Some of them consider every single dimension [19] or every combination of two dimensions [16] as the reduced subspaces, others [20,29] go further in refining the number of relevant subspaces, assuming either [29] that outliers exist in subspaces with non-uniform distributions, or [20] that outliers appear in subspaces exhibiting high dependencies among their related dimensions. These studies, either exploring subspace projections [20,29] or subspace samples [16,19,27], appear to be appropriate for the purpose of outlier identification. Nonetheless, as the outlier score of an object is aggregated from multiple subspaces, it remains unclear which subspace should be selected to interpret the outlier. In addition, the number of explored subspaces for every object should be large in order to obtain good outlier ranking results. These techniques are hence closer to outlier ensembles [31,36] than to outlier interpretation. The SOD method [23] pursues a different approach, seeking an axis-parallel hyperplane (w.r.t. an object’s neighbors), spanned by the attributes with the highest data variances. The anomaly degree of the object is computed in the space orthogonal to this hyperplane. The COP method [25] generalizes this idea by looking for the arbitrarily oriented subspaces of highest variance and further provides an error vector for each identified outlier as a form of explanation. The ABOD model [26] exploits the variance of angles among objects to compute outlying degrees. Intuitively, an outlier lying outside of clusters tends to exhibit a relatively low variance of angles between pairs of vectors pointing to other objects. ABOD also provides an error vector as an outlier interpretation.

III. Local Outliers with Graph Projection

A. General concepts

In many practical applications, data are collected and described in high dimensional spaces. However, many dimensions or features can be irrelevant for outlier exploration and in many cases the intrinsic structure behind the observed data can be captured by only a small number of latent regularities. The success of an outlier mining algorithm therefore depends strongly on how these structures are captured and represented. For our specific problem of identifying and interpreting *local* outliers, capturing the local structure of the data is more important than capturing the global structure (e.g., the global data variance) since the outlierness of each object is depending on the deviation of the object's characteristic from the characteristic of its local neighborhood. We therefore adopt in this work a graph based model to capture the local geometry.

Following this approach, given a collection of data objects $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ where each $\mathbf{x}_i$ is represented as a vector in $\mathbb{R}^{\mathcal{D}}$ (a data space with $\mathcal{D}$ dimensions), we construct an undirected graph $G(V, E)$ to model the local neighborhood relationships among all data instances. Each vertex $v_i \in V$ corresponds to a data object $\mathbf{x}_i \in X$. With k as a supplied parameter, we place an edge $E(i,j) \in E$ between two vertices v_i and v_j if the corresponding object $\mathbf{x}_i$ is among the k nearest neighbors of object $\mathbf{x}_j$ or inversely, $\mathbf{x}_j$ is among the k nearest neighbors of object $\mathbf{x}_i$. In addition, for each edge $E(i,j)$ connecting v_i and v_j, we compute a non-negative weight $K(i,j)$ to reflect how strong the connection between v_i and v_j is, or equivalently, how similar $\mathbf{x}_i$ and $\mathbf{x}_j$ are. This weight represents the neighborhood relationship between two objects $\mathbf{x}_i$ and $\mathbf{x}_j$, and we adopt the widely used radial symmetric Gaussian kernel function[1] for this task given by: $K(i,j) = \frac{1}{(2\pi\sigma)^{\mathcal{D}/2}} \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2 \cdot \sigma^2}\right)$ (with σ being the width of the Gaussian adapted from the data [14]) if v_i, v_j are connected; and $K(i,j) = 0$ if v_i, v_j are not connected (i.e., $\mathbf{x}_i$ and $\mathbf{x}_j$ are not among their k nearest neighbors either way). This forms an affinity matrix K with $K(i,j) \geq 0$. This matrix K is not only symmetric but also sparse. The graph is fully connected and captures the local neighborhood property of every observed data object.

B. Objective function

A similar graph structure has been used in manifold unsupervised learning [6,34] or extended to some supervised settings [10,13] where the objective is to capture the data structures by using a small number of projected dimensions. These approaches, however, cannot be directly applied to the outlier exploration problem since the structures behind the observed data can be multiple clusters rather than a single manifold whereas the labels over regular and outlying objects are also not available to enable a supervised learning. For uncovering local outliers along with discriminative (relevant) features to explain each one as an anomalous object, a local projection of the data is more important than a global projection [6,13] since different outliers could be discriminated by different subsets of original features.

We develop an objective function that can learn an optimal subspace to discriminate an outlier well from nearby inliers while retaining the important structure of the data. Toward this goal, for each data instance $\mathbf{x}_i$, we extract from the above global graph a neighboring subgraph $G^{(i)}$ which comprises the vertices corresponding to the k nearest neighbors of $\mathbf{x}_i$ together with the weights and edge connections among them. This subgraph captures the local geometrical data structure of $\mathbf{x}_i$'s vicinity. We denote the weights and edges of the subgraph by $K^{(i)}$ and $E^{(i)}$, respectively. By space transformation, we would like to find a projection that maps the objects of a local substructure into a lower dimensional subspace such that the geometrical data structure is retained as much as possible. More specifically, let $\mathbf{y}_p$, $\mathbf{y}_q$ in the lower dimensional space $\mathbb{R}^d$ be the mapping points of $\mathbf{x}_p$, $\mathbf{x}_q$, the neighboring objects of $\mathbf{x}_i$, in the original space $\mathbb{R}^{\mathcal{D}}$ (with $d \ll \mathcal{D}$), we form our first objective function:

$$\text{minimize} \sum_p \sum_q \|\mathbf{y}_p - \mathbf{y}_q\|^2 K^{(i)}_{(p,q)} \tag{1}$$

If $\mathbf{x}_p$ and $\mathbf{x}_q$ are close neighboring objects in the original space, the weight $K^{(i)}_{(p,q)}$ between them is high and this objective function will penalize a large value if their respective mapping instances $\mathbf{y}_p$ and $\mathbf{y}_q$ are mapped far apart in the new transformed subspace. Therefore, minimizing this function is equivalent to optimally preserving the local structure of nearby instances around $\mathbf{x}_i$ and the d transformed dimensions are their most descriptive features in our reduced dimensional subspace. On the other hand, by considering $\mathbf{x}_i$ as an outlier candidate, its mapping point in the new transformed subspace should be as far as possible from its nearby mapping neighbors: if $\mathbf{x}_i$ is truly an outlier, it should be well discriminated from its neighboring objects. Our second objective function formulates this:

$$\text{maximize} \sum_p \|\mathbf{y}_i - \mathbf{y}_p\|^2 K^{(i)}_{(i,p)} \tag{2}$$

[1] Other kernel functions can also be used to compute $K(i,j)$, e.g., the dot product if $\mathbf{x}_i$, $\mathbf{x}_j$ are the term vectors in document data.

Again, $\mathbf{y}_i$ is the mapping point of $\mathbf{x}_i$ in our reduced dimensional subspace and this function will incur a high penalty if $\mathbf{x}_i$ and $\mathbf{x}_p$ are far part in $\mathbb{R}^{\mathcal{D}}$ but being mapped close in this $\mathbb{R}^d$ subspace.

To gain more insights into our two objective functions, let $D^{(i)}$ be a diagonal matrix whose entry $D^{(i)}_{pp} = \sum_q K^{(i)}_{(p,q)}$ and let W be the matrix having size of $\mathcal{D} \times d$ that maps $\mathbf{x}_p$'s into $\mathbf{y}_p$'s.[2] The squared vector norm in our first objective function (Eq. (1)) can be written as a function of W:

$$\begin{aligned}
J_m(W) &= \sum_p \sum_q \|W^T\mathbf{x}_p - W^T\mathbf{x}_q\|^2 K^{(i)}_{(p,q)} \\
&= \sum_p \sum_q tr\left(W^T(\mathbf{x}_p - \mathbf{x}_q)(\mathbf{x}_p - \mathbf{x}_q)^T W\right) K^{(i)}_{(p,q)} \\
&= tr\left(\sum_p \sum_q \left(W^T(\mathbf{x}_p - \mathbf{x}_q) K^{(i)}_{(p,q)} (\mathbf{x}_p - \mathbf{x}_q)^T\right) W\right) \\
&= 2 \cdot tr\left(W^T X^{(i)} D^{(i)} X^{(i)^T} W\right) \\
&\qquad - 2 \cdot tr\left(W^T X^{(i)} K^{(i)} X^{(i)^T} W\right)
\end{aligned} \tag{3}$$

where $tr(.)$ is the trace of a matrix and $X^{(i)}$ is the matrix having the k nearest neighbors of $\mathbf{x}_i$ as its column vectors. By defining $L^{(i)} = D^{(i)} - K^{(i)}$ as the Laplacian matrix of our equation and ignoring the constant, we are able to re-write our first objective function:

$$J_m(W) = tr\left(W^T X^{(i)} L^{(i)} X^{(i)^T} W\right) \tag{4}$$

For the second objective function, note that only the relationship between $\mathbf{x}_i$ and its k nearest neighbors is of concern. The weight entries thus form a vector $K^{(i)}_{(i,.)}$, rather than a full matrix. However, in order to be consistent with the matrix form in our first objective function (Eq. (4)), we represent it as a sparse matrix by:

$$K^{(i')} = \begin{pmatrix} 0 & K^{(i)}_{(i,2)} & \cdots & K^{(i)}_{(i,k)} \\ K^{(i)}_{(2,i)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ K^{(i)}_{(k,i)} & 0 & \cdots & 0 \end{pmatrix}$$

and consequently, $D^{(i')}$ as the diagonal matrix:

$$D^{(i')} = \begin{pmatrix} \sum_p K^{(i)}_{(i,p)} & 0 & \cdots & 0 \\ 0 & K^{(i)}_{(i,2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K^{(i)}_{(i,k)} \end{pmatrix}$$

Then, keeping in mind the symmetry of the matrix $K^{(i')}$ and all its entries being zero except for those in the first row and column, we can re-phrase our second objective function in terms of optimizing the projection matrix W:

$$\begin{aligned}
J_M(W) &= \sum_p \left\|W^T\mathbf{x}_i - W^T\mathbf{x}_p\right\|^2 K^{(i')}_{(i,p)} \\
&= \frac{1}{2} \times \sum_p \sum_q \left\|W^T\mathbf{x}_p - W^T\mathbf{x}_q\right\|^2 K^{(i')}_{(p,q)} \\
&= tr\left(W^T X^{(i')} \left(D^{(i')} - K^{(i')}\right) X^{(i)^T} W\right) \\
&= tr\left(W^T X^{(i')} L^{(i')} X^{(i')^T} W\right)
\end{aligned} \tag{5}$$

in which from the first row to the second row of the equation, we have included $\mathbf{x}_i$ as the first column of $X^{(i')}$ and thus the indices p, q (in the 2nd row) also go through it. It is crucial to mention here that while we want to learn a single transformation matrix W for both objective functions, the matrix $X^{(i)}$ used in Eq. (4) is different from $X^{(i')}$ in Eq. (5) by the single (first) column (i.e., $\mathbf{x}_i$). A simple cure for this matter therefore is to also add $\mathbf{x}_i$ as the first column of $X^{(i)}$, making two matrices $X^{(i)}$ and $X^{(i')}$ identical, and accordingly, a 0-vector also must be added to $L^{(i)}$ as its first row and column, yielding the consistency of matrix size matching in Eq. (4).[3] For simplicity, we use the notation $X^{(i)}$ and in combination with the results above, the two objective functions can be combined into a single one:

$$\begin{aligned}
\text{maximize}\, J(W) &= J_M(W) - J_m(W) \\
&= tr\left(W^T X^{(i)} L^{(i'')} X^{(i)^T} W\right) \\
\text{subject to}\quad & W^T X^{(i)} D^{(i)} X^{(i)^T} W = I \\
\text{and}\quad & \mathbf{w}_p^T \mathbf{w}_q = 0 \quad \text{for } p \neq q
\end{aligned} \tag{6}$$

where we have used $L^{(i'')}$ to denote $(L^{(i')} - L^{(i)})$ and used the linear map property of $tr(.)$; and vectors $\mathbf{w}_p$, $\mathbf{w}_q$ are columns of W. The first constraint is added as we need directions of columns in W rather than their magnitude (since $X^{(i)}$ is projected onto them) whereas the second constraint is incorporated to impose the independence among projected dimensions.

As observed, for each data instance $\mathbf{x}_i$, our objective is to learn a low dimensional mapping subspace in which $\mathbf{x}_i$, if it is an outlier, is well discriminated from its nearby objects while at the same time the geometrical structure of the data is preserved. This approach can be seen as closely related to large-margin nearest-neighbor (LMNN [8]), in which a similar projection is learned to separate classes in a supervised setting; but in contrast to LMNN we try to learn such a projection for every single object to separate it from its neighbors. The appropriate original features selected in the new transformed subspace are defined by the coefficients learnt from the columns of the projection matrix W. A naïve

[2] Please note the notation difference between $D^{(i)}$ and dimension $\mathcal{D}$.

[3] This is equivalent to adding a vector with 1st entry being equal to $K_{(i,i)}$ (corresponding to $\mathbf{x}_i$) and zero elsewhere as the first row and column of the matrices $K^{(i)}$ and $D^{(i)}$.

approach is to restrain W's column entries to be either 1 or 0 (i.e., corresponding original features are selected or not selected) and choose W which leads to the largest value in Eq. (6). Nonetheless, this search process is discrete as features are either retained or discarded, and the number of subspaces to be explored is also exponential in the number of features, making this approach computationally expensive. We thus deal with this challenge in a more tractable way by imposing the penalty on the norm of each column of W, leading to the final optimization function:

$$\begin{aligned} W^* =& \arg\max_{W} \left\{ tr\left(W^T X^{(i)} L^{(i'')} X^{(i)^T} W \right) - \alpha W^T W \right\} \\ & \text{subject to } \; W^T X^{(i)} D^{(i)} X^{(i)^T} W = I \\ & \text{and } \; \mathbf{w}_p^T \mathbf{w}_q = 0 \quad \text{for } p \neq q \end{aligned} \tag{7}$$

Incorporating the penalty retrained on the vector norm is often called the regularization and is well studied in the statistics community [18]. The parameter $\alpha \geq 0$ is used to apply the amount of shrinkage imposed on W's columns. Though other forms, like the L_1 norm penalty, can also be used here, we employ the quadratic L_2 norm for the ease of optimization. The results between L_1 and L_2 penalties are not much different as long as the number of selective features is small [18], yet using the L_1 penalty makes the solutions nonlinear and thus requires more complex techniques (e.g., quadratic programming) to optimize. We show in the following section a closed form solution for our optimization setting in Eq. (7).

C. Subspace learning

In solving the trace optimization associated with constraints, we can use the Lagrange multipliers method. Eq. (7) can be recast as the Lagrangian of the following problem:

$$\begin{aligned} \mathcal{L}(W, \Lambda) =& W^T \left(X^{(i)} L^{(i'')} X^{(i)^T} - \alpha I \right) W \\ & - \Lambda \left(W^T X^{(i)} D^{(i)} X^{(i)^T} W - I \right) \end{aligned} \tag{8}$$

in which Λ is a diagonal matrix, its entries being the Lagrange multipliers. Solving this objective function for W will satisfy our added constraints over W whereas the columns of W are also naturally orthogonal to one another (as shortly presented). Taking the derivative of $\mathcal{L}$ with respect to W and equating it to zero gives us:

$$\left(X^{(i)} L^{(i'')} X^{(i)^T} - \alpha I \right) W = \Lambda X^{(i)} D^{(i)} X^{(i)^T} W \tag{9}$$

This equation has the form of a generalized eigenvalue problem. Nonetheless, notice that the right hand side matrix $X^{(i)} D^{(i)} X^{(i)^T}$ is not full rank since, in general, we have the number of data dimensions exceeding the number of nearest neighbors (actually, plus 1 as $\mathbf{x}_i$ has been included in $X^{(i)}$). Hence, this matrix is not directly invertible. Dealing with this issue, it is better to decompose $X^{(i)}$ into three matrices $X^{(i)} = U^{(i)} \Sigma^{(i)} V^{(i)^T}$, of which columns in $U^{(i)}$ and $V^{(i)}$ are the left and right singular vectors and the diagonal elements in $\Sigma^{(i)}$ are the singular values, and $X^{(i)}$ is approximated by keeping the most significant singular values in the diagonal matrix $\Sigma^{(i)}$.[4] By denoting $\widetilde{W} = U^{(i)^T} W$ and $\widetilde{X}^{(i)} = \Sigma^{(i)} V^{(i)^T}$, it is straightforward to see that:

$$\begin{aligned} & W^T \left(X^{(i)} L^{(i'')} X^{(i)^T} - \alpha I \right) W \\ & \quad = \widetilde{W}^T \left(\Sigma^{(i)} V^{(i)^T} L^{(i'')} V^{(i)} \Sigma^{(i)} - \alpha I \right) \widetilde{W} \\ & \quad = \widetilde{W}^T \left(\widetilde{X}^{(i)} L^{(i'')} \widetilde{X}^{(i)^T} - \alpha I \right) \widetilde{W} \end{aligned} \tag{10}$$

and

$$\begin{aligned} X^{(i)} D^{(i)} X^{(i)^T} W &= \widetilde{W}^T \Sigma^{(i)} V^{(i)^T} D^{(i)} V^{(i)} \Sigma^{(i)} \widetilde{W} \\ &= \widetilde{W}^T \widetilde{X}^{(i)} D^{(i)} \widetilde{X}^{(i)^T} \widetilde{W} \end{aligned} \tag{11}$$

Rather than directly seeking W from Eq. (8), we find it via the equation $W = U^{(i)} \widetilde{W}$ whereas $\widetilde{W}$ is found from:

$$\left(\widetilde{X}^{(i)} L^{(i'')} \widetilde{X}^{(i)^T} - \alpha I \right) \widetilde{W} = \Lambda \widetilde{X}^{(i)} D^{(i)} \widetilde{X}^{(i)^T} \widetilde{W} \tag{12}$$

In order to show the matrix $\widetilde{X}^{(i)} D^{(i)} \widetilde{X}^{(i)^T}$ on the right hand side being nonsingular, we need the following proposition:

Proposition 1: Let $V^{(i)}$ be our matrix with orthogonal column vectors, then its row vectors are also orthogonal, i.e., $V^{(i)} V^{(i)^T} = I$

Proof: Let $\mathbf{a}$ be an arbitrary vector, we need to show $V^{(i)} V^{(i)^T} \mathbf{a} = \mathbf{a}$. It is true that $V^{(i)^T} V^{(i)} = I$ as $V^{(i)}$'s column vectors are orthogonal. Therefore, the inversion of $V^{(i)}$ is equal to $V^{(i)^T}$ and given $\mathbf{a}$, there is a uniquely determined vector $\mathbf{b}$ such that $V^{(i)} \mathbf{b} = \mathbf{a}$. Consequently,

$$V^{(i)} V^{(i)^T} \mathbf{a} = V^{(i)} V^{(i)^T} V^{(i)} \mathbf{b} = V^{(i)} \mathbf{b} = \mathbf{a}$$

It follows that $V^{(i)} V^{(i)^T} = I$ since $\mathbf{a}$ is an arbitrary vector. □

Theorem 1: Let B be the matrix $\widetilde{X}^{(i)} D^{(i)} \widetilde{X}^{(i)^T}$ in which $\widetilde{X}^{(i)} = \Sigma^{(i)} V^{(i)^T}$, then B is non-singularity.

Proof: The proof of this theorem is straightforward given Proposition 1 and keeping in mind that both $D^{(i)}$ and $\Sigma^{(i)}$ are diagonal positive semi-definite matrices. □

By defining $A = (\widetilde{X}^{(i)} L^{(i'')} \widetilde{X}^{(i)^T} - \alpha I)$, it follows that:

$$B^{-1} A \widetilde{W} = \Lambda \widetilde{W} \tag{13}$$

This is a generalized eigenvalue problem where we can find the solution by first looking for the largest eigenvalue λ_1 and the corresponding eigenvector $\widetilde{\mathbf{w}}_1$. Then, the second largest eigenvalue/eigenvector λ_2 and $\widetilde{\mathbf{w}}_2$ are found by taking the constraint $\widetilde{\mathbf{w}}_1^T \widetilde{\mathbf{w}}_2 = 0$ and so on. Overall, the diagonal values in Λ and the corresponding column vectors in $\widetilde{W}$ are the eigenvalues (in descending order) and the eigenvectors, respectively, of $B^{-1} A$. It is further observed that if Λ and $\widetilde{W}$ are the optimal solutions for Eq. (12), then Λ and

[4] In this work, we consider singulars whose values are smaller than 10^{-5} as 0 and remove them from $\Sigma^{(i)}$.

$W = U^{(i)}\widetilde{W}$ are also the optimal solutions for Eq. (9). Additionally, since $U^{(i)}{U^{(i)}}^T = I$, column vectors in W are also pairwise orthogonal and they are finally our globally optimum solution.

D. Outlier score computation

Generally, our solution developed above can find up to d, equal to the rank of matrix $X^{(i)}$, as the number of dimensions to which the outlier candidate $\mathbf{x}_i$ can be discriminated from its neighboring objects. Nonetheless, it is noted that the importance degrees of all induced dimensions (i.e., eigenvectors) are not the same and essentially can be assessed by their corresponding eigenvalues. Since the eigenvalues and eigenvectors are going in pairs, by ordering the eigenvalues in the descending order, the corresponding top eigenvectors are thus the most significant dimensions. Let d be the number of top eigenvectors selected from W and let $X^{(i)}$ comprise $\mathbf{x}_i$'s neighbors as its columns, then the outlier score (OS) of $\mathbf{x}_i$ can be computed as the statistical distance from $\mathbf{x}_i$ to its neighboring objects in the transformed space as follows:

$$OS(\mathbf{x}_i) = \frac{1}{d}\sum_{p=1}^{d}\sqrt{\frac{\max\left\{\left(\mathbf{w}_p^T\mathbf{x}_i - \frac{1}{k}\sum(\mathbf{w}_p^T X^{(i)})\right)^2, \sigma_p^2\right\}}{\sigma_p^2}}$$

in which σ_p is the standard deviation of the neighboring objects projected onto the eigenvector $\mathbf{w}_p$. The variance in the second term of the max operation is used as the lower bound to constrain the projection of $\mathbf{x}_i$ not too close to the projected center. This can happen when $\mathbf{x}_i$ is a regular object surrounded by similar other regular objects in the same data cluster. Thus, the smallest value of the defined outlier score is limited by 1. The higher the value of $OS(\mathbf{x}_i)$, the more $\mathbf{x}_i$ deviates from its neighbors.

Let us note that, according to the categorization of local outlier detection methods [32], the score OS is an outlier model of first order locality.

E. Final discriminative features

Using the leading eigenvectors helps to visualize the deviation of an outlier candidate $\mathbf{x}_i$, along with the geometrical structure of its nearby objects, in the lower d-dimensional subspace. This visualization, however, does not directly show the user which features in the original space are the most crucial in discriminating $\mathbf{x}_i$ as an exceptional object. Nevertheless, recall that since our approach in this work is the space transformation, the coefficients of eigenvectors therefore unveil how the original features have been combined to induce the optimal subspace. In other words, they contain essential information regarding the features contributing most to the formation of the new transformed space. Additionally, we assume that $\mathbf{x}_i$, as a local outlier, can be linearly separated from its neighbors, the dimensionality of such a transformed subspace might be as small as one. In this case, the first eigenvector of Eq. (9), corresponding to the largest eigenvalue of the $B^{-1}A$ matrix, will be the optimal direction. Let $\mathbf{w}$ denote this leading eigenvector and let the absolute values of its entries be ordered decreasingly. It would be expected that there exists an appreciable difference in the absolute coefficients of relevant and irrelevant features due to the regularization over the $\mathbf{w}$'s L_2 norm. As such, in explaining the exceptional properties of the object $\mathbf{x}_i$, we select top q original features corresponding to the top q largest absolute coefficients in $\mathbf{w}$ such that $|w_q| - |w_{q+1}| \geq 2 \times \frac{1}{q-1}\sum_{i=1}^{q-1}(|w_i| - |w_{i+1}|)$. This means the difference in coefficients between the relevant and irrelevant features is larger at least by a factor 2 than most of the differences between two ordered relevant features. Nonetheless, for cases where such a gap of difference is not very prominent, we also provide the user a parameter γ, whose value is within the range $(0, 1)$, to generate the set of q features such that $\sum_{i=1}^{q}|w_i| \geq \gamma \times \sum_{j=1}^{\mathcal{D}}|w_j|$ (in this work, we typically choose $\gamma = 0.8$). It is worth mentioning that, since each outlier may have different features discriminating it as an anomalous object, the value q might also be different across different outliers.

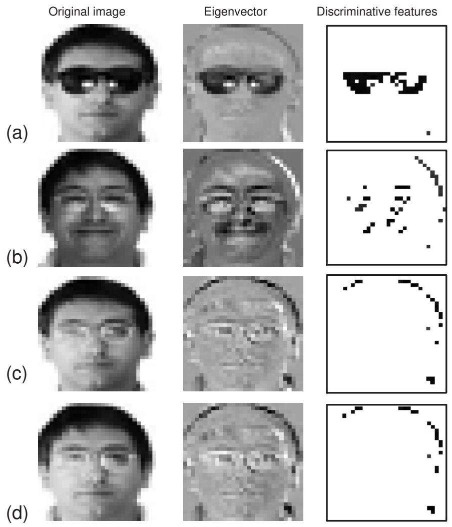

Figure 2. Discriminative features for outliers from the AR face images. Graphs in the first two rows (a-b) relate to the true two outlying images whereas graphs in the last two rows (a-b) relate to the two regular images ($k = 5$). See text for explanation.

To illustrate the selection of discriminative features for outlier explanation, we continue with our example on the AF face data set presented in Section 1 (images having size of 32×30 pixels are viewed as vectors $\mathbf{x}_i$'s of 960 features). In Figures 2(a-b), we show two anomalous images having the highest local outlier scores of 1.838 and 1.643. The leading eigenvector (reshaped into 32×30 images), onto which the outlying image and their neighbors are projected, is shown in the second column and the final set of relevant features are shown in the third column of each figure. As observed from Figure 2(a), the most significant coefficients of the leading eigenvector $\mathbf{w}$ are the ones around the eye area and by our proposed method based on the significant difference between features' coefficients, the top 68 features (i.e., original pixels) have been selected as the relevant features explaining the image as an outlier. Likewise, for the second anomalous image shown in Figure 2(b), out of 960 original features, 43 pixels have been selected to interpret it as an exceptional object. Though this set of discriminative features are not concentrated compared to that of the first outlier and consist of some less relevant pixels (around the head shape), they are clearly intuitive and consistent to the "anomalous" expression of the human face. Due to smiling, the features around the mouth, cheek-bone and eyes have been learnt with the most significant coefficients and consequently our algorithm has selected them as the discriminative features to characterize for its anomalous properties. To provide more insights, we further plot the features selected for the next two images in the outlier ranking list, having OS respectively of 1.182 and 1.159, in the Figures 2(c) and (d). As observed, no special features within the face area have been selected since these images are truly regular and they are very much similar to other images in this example.

F. Algorithm Complexity

Our algorithm developed above is named LOGP which stands for Local Outliers with Graph Projection and its computation complexity is analyzed as follows. LOGP first needs to build the global graph which requires the calculation of the nearest neighbors and the K matrix, which takes $O(\mathcal{D}N^2)$ in the worst case or can reduce to $O(\mathcal{D}N \log N)$ if the implementation of the k-$\mathcal{D}$ tree data structure is used [7]. The size of the matrix $X^{(i)}$ is $\mathcal{D} \times k$ and its singular value decomposition takes $O(\mathcal{D}k \log(k))$ with the Lanczos technique [17]. Likewise, the eigen-decomposition of the matrix $B^{-1}A$ takes $O(\mathcal{D}^2 \log \mathcal{D})$ since its size is $\mathcal{D} \times \mathcal{D}$. As we compute these steps for all instances to render the outlier ranking list, the computation amounts to $O(\mathcal{D}Nk \log(k) + \mathcal{D} \log \mathcal{D}))$. The overall complexity is thus at most $O(\mathcal{D}N(\log N + k \log(k) + \mathcal{D} \log \mathcal{D}))$.

IV. Evaluation

A. Methodology

To evaluate the performance of the proposed method, we compare it on multiple real world datasets to a representative selection of established methods: (1) LOF [9] which is one of the most well known algorithms in seeking local outliers from varying density data; (2) SOD [23] which seeks outliers in axis-parallel subspaces; (3) COP [25] which finds outliers in arbitrarily oriented subspaces; (4) ABOD [26] which discovers outliers based on angles between vector triples in high dimensional spaces; and (5) HiCS [20] which seeks outliers in multiple subspaces. All reference implementations are available in ELKI [2].

To evaluate the algorithm performance, we use the well-established receiver operating characteristic (ROC) curve, which visually shows the relationship between the true positive rate (y-axis) and the false positive rate (x-axis). The optimal curve goes straight up the y-axis, only then right. A random result will be close to the diagonal, while values below the diagonal indicate a reverse ordering. This curve can be summarized into a single value when desired, known as the area-under-curve, ROC AUC.

An inherent problem in evaluating algorithm performance of unsupervised methods is parametrization. Reporting results for optimal parameters only is an unrealistic scenario, as it is not trivially possible to find these parameters for a real problem. Instead, we try a best-effort-approach to choose a realistic set of parameters. Where applicable, we prefer a simple maximum ensemble approach as discussed for LOF in the original publication [9] and pointed out as being an early, model-centered, ensemble approach recently [3]. This ensemble approach will not be able to combine different algorithms [24,31] but is reasonable to use with slightly different parametrizations of the same method. For LOF, SOD, and LOGP we construct such a object-wise maximum ensemble [3] for $k = 5 \ldots 25$, which is a reasonable parameter range for these algorithms. The SNN neighborhood size ℓ for SOD was set to the same value as k and $\alpha = 0.8$. With COP, choosing small values of k is not sensible, as k must be larger than the dimensionality for local PCA [1]. Instead, we chose $k = 3 \cdot \mathcal{D}$. For the CMU faces data set (which has many more dimensions than instances), we first reduced the dimensionality to 8 using PCA, then chose $k = 24$. To achieve more stable results, COP was configured to use robust weighted PCA [22] locally and the χ normalization. For ABOD, we use the exact version (since the data sets are small enough) with polynomial kernel of degree 2. For HiCS, we set the various parameters to the values suggested by the authors [20]; inside HiCS we used LOF with $k = 10$. We do not use an ensemble approach here, as HiCS itself already is an ensemble method [3,37] and it did not further improve results in preliminary experiments. Furthermore, HiCS already suffers from a high runtime. For the proposed method LOGP, we usually set $\alpha = 0.1$ in Eq. (8), and use an object-wise minimum ensemble to combine the results of $k = 5 \ldots 25$.

B. CMU Face data set

The first real world data set we use to demonstrate the performance of all algorithms is the high dimensional CMU Face data set [5]. This data set contains grayscale images captured from 20 people. There are up to 32 images for each person covering every combination of the 4 facial expressions (neutral, happy, sad, angry), 4 head positions (straight, left, right, up), and 2 eye states (open and sunglasses). Each image has a size of 32 × 30 pixels which can be interpreted as vectors in a space of 960 dimensions.

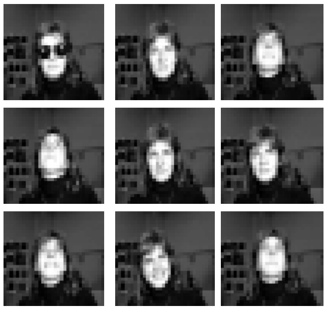

Figure 3. Nine images selected from a person in the CMU Face data set, where the first image is labelled as an outlier due to the appearance of the sunglasses.

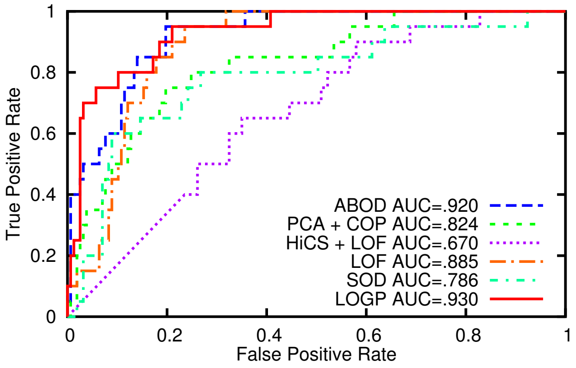

Figure 4. The ROC curve performance of all algorithms on the outlier detection rate over the CMU face database.

Table I
ROC AUC RESULTS ON CMU FACE DATABASE.

algorithm	key parameters	AUC
ABOD	polynomial kernel of degree 2	.920
PCA + COP	PCA to 8 dimensions, COP $k = 24$	.825
HiCS + LOF	1000 candidates per iteration, LOF $k = 10$	.670
LOF	max score for $k = 5 \ldots 25$	.885
SOD	$\alpha = 0.8$, max score for $\ell = k = 5 \ldots 25$	.786
LOGP	$\alpha = 0$, min score for $k = 5 \ldots 25$	.930

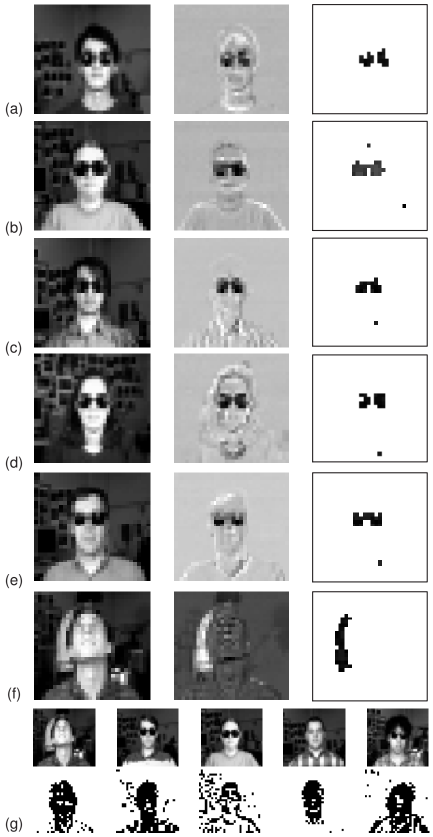

Figure 5. Five (a-e) outlying images (1st column) along with their leading eigenvectors (2nd column) and the discriminative features (3rd column) identified by LOGP. Images in (f) relate to a false positive, a normal object with a high outlier score. The last two rows in (g) show SOD's 5 top ranked outlying images and their corresponding subspace features.

Instead of randomly generating artificial outliers or manually modifying some images to be outliers, we adopt a more natural way by downsampling images with sunglasses from each person to 1 and keeping all remaining frontal images (without sunglasses, and not looking left or right) as regular objects, resulting in a data set with 20 outliers and 157 inliers (some combinations are missing from the original data). For illustration, several randomly selected images from the same person are shown in Figure 3 where the first image is labelled as a ground-truth outlier.

Outlier identification: In Figure 4, we plot the ROC curve performance of all algorithms. For each curve, we further report the area under the curve (AUC) which provides a way to numerically compare the algorithms' performance in Table I. For SOD, choosing $\ell = k$ yields better performance than setting ℓ to the estimated class size of 8 images per person. For COP, we reduced the data set dimensionality to 8 dimensions using PCA. For HiCS, we had to increase the candidate cutoff parameter to 1000 because of the high dimensionality to yield meaningful results. For LOGP, we disabled regularization by using $\alpha = 0$ to decrease runtime, and we had to estimate the σ^2 kernel bandwidth manually.

As seen from this figure, LOGP performed best with respect to the ROC AUC metric, and in particular on the first few items only (i.e. with a false positive rate of less than 10%). ABOD, designed for high-dimensional data, also worked very well. Note that a simple changing variable in solving Eq. (7) shows more stable results for this experiment and since the bias-tradeoff technique [14] is less suitable for very high dimensional data, we have chosen $\sigma = 3$. LOF, while not being a subspace method, performs very well in this parameterization (taking the maximum score of each object for $k = 5 \ldots 20$, which is maybe best explained with the gains from an ensemble approach. While HiCS is designed to find informative subspaces, it does not scale well to this very high dimensionality. It tries to find subspaces bottom up, but the discriminative subspaces are of medium dimensionality here (i.e. multiple pixels) and cannot be easily identified in 2 dimensions. However, in order to avoid exponential complexity, HiCS only retains a limited set of subspace candidates at each iteration: out of the $\binom{960}{2} = 460320$ 2-dimensional subspace candidates it only retains a fixed (parameterizable) number of $100 - 1000$. If this subset is not chosen well, it becomes hard to find the most relevant subspaces in higher dimensionality. SOD suffers from a different problem of data set size: both to find a reasonable neighbor set and to compute the relevant subspace, it needs much more neighbors than this small but high-dimensional data set can provide.

Outlier interpretation: We further explore the set of discriminative features generated by our LOGP algorithm for each of its 20 top uncovered outlying images. We show the 5 top outliers, along with their uncovered discriminative features, in Figures 5(a-e). As expected, LOGP performs quite well with this very high dimensional data set where most features chosen to explain the anomalous images are around the eyes area, due to the appearance of the sunglasses. Furthermore and surprisingly, among these top discovered outliers, LOGP also ranks high an image which is shown in Figure 5(f). Note that, according to our ground truth that labels pictures with sunglasses as outliers, this image is a regular object rather than an outlier; yet by inspecting the set of discriminative features discovered by LOGP, this image is clearly outstanding as an exception compared to other images due to a person unintentionally appearing at the background. As visualized from the last graph of Figure 5(f), his white T-shirt is well captured as the discriminative features for this outlying image by our algorithm.

Recall that SOD, COP, and HiCS are all subspace-based techniques, of which COP uses error vectors to characterize outliers whereas HiCS produces multiple subspaces for a *single* outlier. We therefore select SOD for visualizing its subspace for each uncovered outlier. In Figure 5(g), we show its top five outlying images along with the corresponding subspaces in which these outliers have been found. As observed, SOD performance also seems to be effective with 3 true outlying images ranked top. It even further gives the image with a person at the background the highest outlier score. Nonetheless, by looking at the feature subspaces generated by the SOD, we found that, unlike the LOGP method, it is hard to explain why these images are exceptional. The reason, as justified above, is the fundamental difference in the subspace learning objectives of two approaches. We therefore do not attempt to compare their uncovered subspaces in the subsequent experiments.

C. Pen digit data set

We next provide experiments on the pen digit data set [5], consisting of 1602 data samples, where each sample corresponds to a hand written digit. As a digit is being written on a pen-based tablet, 8 (x, y) positions of the pen are recorded and they consequently form the 16 attributes of the digit. Similar to the CMU face data set, we keep instances from two randomly selected digits as regular objects while downsampling to 2 instances from each of 8 remaining digits as outliers, yielding the data set with 16 anomalous digits and 334 inliers from digits 1 and 5.

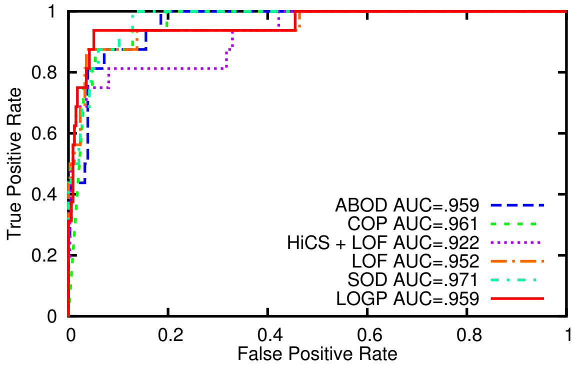

Figure 6. ROC curve performance of all algorithms on the outlier detection over the pen digit data set.

Table II
ROC AUC RESULTS ON PEN DIGITS DATABASE.

algorithm	key parameters	AUC
ABOD	polynomial kernel of degree 2	.959
COP	COP $k = 3 \cdot \mathcal{D} = 48$	.961
HiCS + LOF	100 candidates per iteration, LOF $k = 10$	.922
LOF	max score for $k = 5 \dots 25$	.952
SOD	$\alpha = 0.8$, max score for $\ell = k = 5 \dots 25$	.972
LOGP	$\alpha = 0.1$, min score for $k = 5 \dots 25$	.959

Outlier identification: In Figure 6, the performance in terms of ROC curve have been plotted for all algorithms. On this data set, all algorithms fared very well (indicating that we chose reasonable parameterization). LOGP came in at a close third after SOD and COP, and tied with ABOD. HiCS performance was much better (likely due to this data set only having 16 dimensions), but not competitive to the LOF ensemble. The surprising winner on this data set was SOD. Note that, however, the score is dominated by when the methods find the last outlier. LOGP is the first method to have found 15 out of 16 outliers, but late at finding all outliers. Such differences in algorithmic performance are not adequately reflected by the ROC AUC measure, and should not be considered significant.

Outlier interpretation: To save space, we report in Figures 7(a-e) the five top ranking anomalous digits together with the features selected to discriminate them as outliers. In these figures, we show the true label above each plotted digit (top graph) whereas its corresponding discriminative features are marked with labels (bottom graph). Compared to the CMU face data set, it is slightly harder to observe the set of discriminative features; however, combining both the visualization over the digits and their set of discriminative features gives us the explanation as follows. First, due to

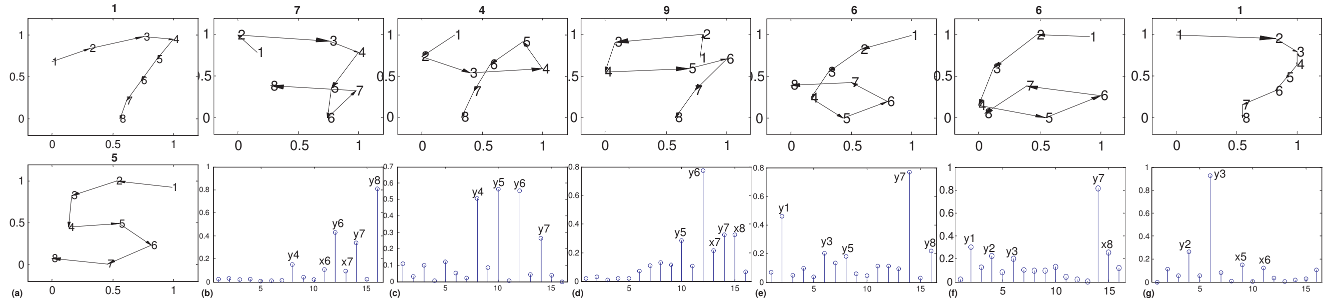

Figure 7. Outlying hand-written digits uncovered by LOGP. Two samples from distributions of regular digits 1 and 5 are shown in (a). Top five outlying digits are shown in the (b-f)'s top row, whereas their discriminative features (with labels) are shown in the bottom row. A regular digit ranked as one of LOGP's top outliers is shown in (g). For graphs plotting the digits, the horizontal and vertical axes correspond to the x- and y- coordinates of the pen-based tablet. For graphs plotting the features, the horizontal and vertical axes correspond the feature index and feature coefficient respectively.

the randomness when the data set is created, all instances from two digits 1 and 5 have been labelled as regular objects whilst each 2 of the remaining digits have been selected as hidden outliers. Then, by further inspecting the data, we found that the digits 7 and 4, out of the top five outlying digits shown in Figures 7(b-f), are closest to the data distribution of digit 1 and thus they can be viewed as the local outliers w.r.t. this distribution. Likewise, the digits 6 and 9 are closest to the distribution of digit 5 and hence are the anomalous objects w.r.t. this distribution. For the sake of discussion, we also plot in Figure 7(a) the popular writing of the digit 1 (top graph) and 5 (bottom graph). Now, selecting the digit 7 shown in Figure 7(b) as an example, its most discriminative features are the y-coordinates of the last 3 strokes compared to that of digit 1. More specifically, while these strokes in digit 7 go down to the bottom, up right then left, making $y6 = 0$, $y7 = 0.35$ and $y8 = 0.4$, these stokes go down constantly in digit 1, resulting in $y6 = 0.5, y7 = 0.2$ and $y8 = 0$. This writing style clearly distinguishes the digit 7 from that of the popular digit 1. It is further observed that this digit resembles the writing style with the digit 1 on their beginning strokes (both starting from top left then going right). Correspondingly, LOGP has learnt the leading features (i.e., from $x1, y1$ to $x3, y3$) with coefficients close to 0 and consequently, they are well excluded from the set of discriminative features for the digit 7. A similar interpretation can also be applied to the outlying digit 4 w.r.t. digit 1, where its y-coordinates of 3rd, 4th and 5th strokes go right, up left then down left (i.e., $y4, y5, y6, y7$), obviously contrasting to those of digits 1's which go down gradually.

For the outlying digit 9 (Figure 7(d)) which is closest to the distribution of digit 5, the most discriminative feature is $y6$. It is obviously seen that while the value of $y6$ in digit 9 positions in the upper half, that value in digit 5 locates in the lower half. The next discriminative feature is $x8$ which locates close to $x = 0.5$ in digit 9 but close to $x = 0$ in digit 5. Likewise, we can observe the clear difference in the other selected features $y5, x7$ and $y7$ that all helps to distinguish digit 7 from digit 1.

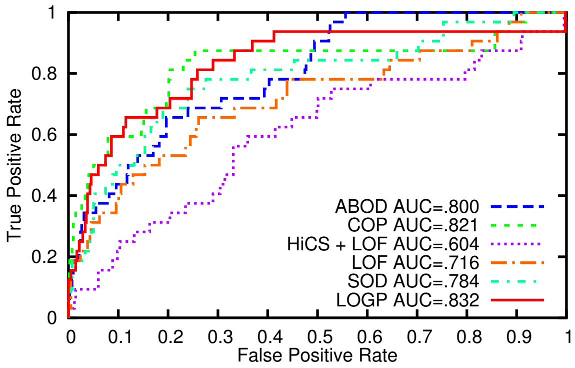

Figure 8. ROC curve performance of all algorithms on the outlier detection over the English letter data set.

For the features discriminating the two outlying examples of digit 6 in Figures 7(e-f) from the 5's distribution, $y7$ tends to be the top selected feature. As observed, its value in the both outlying digits 6 locates close to $y = 0.5$ but close to $y = 0$ in the digit 5. The subsequent discriminative features can be the $y1$ and $y3$. However, since the writing styles of the two examples of digit 6 shown in Figures 7(e-f) are slightly different at the ending stroke, $y8$ is selected as another discriminative feature in the first digit 6 but not in the second one. It is also worth mentioning here that in the 16 top ranking outliers found by our algorithm, there are also several digits coming from the two regular distributions of 1's and 5's. Nonetheless, since digits are written by different people, we found that these outliers, though wrongly labelled from the ground truth, are still quite anomalous compared to the popular writing of 1's and 5's. In the last Figure 7(g), we show an example of such cases and as seen, it is rather hard to say the written number is the digit 1.

D. Other UCI datasets

We further test our algorithm and its competing techniques on two more benchmark UCI datasets [5]: (i) the image segmentation and (ii) the English letter recognition. The image segmentation data set consists of 2,310 instances of outdoor images being categorized into 7 classes {brickface, sky, foliage, cement, window, path, grass}. Each instance is a

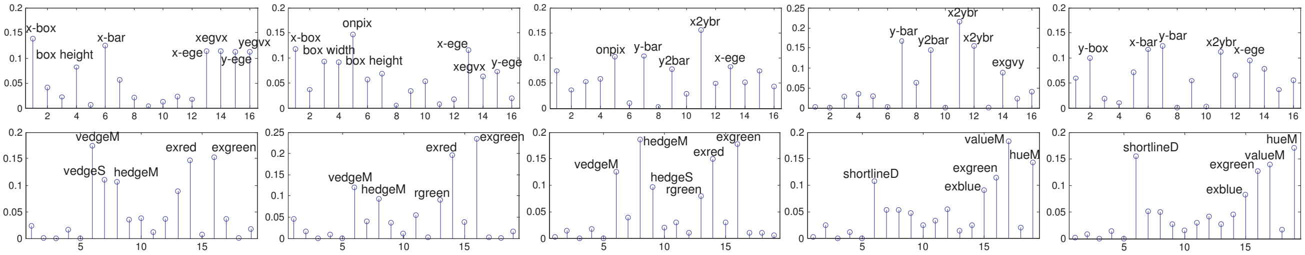

Figure 10. Discriminative features (those with labels) from the top five outliers found from the English letter data set (top row) and the image segmentation data set (bottom row). In each graph, the x-axis is the feature index whereas the y-axis is the feature coefficient.

Table III
ROC AUC RESULTS ON ENGLISH LETTER DATA SET.

algorithm	key parameters	AUC
ABOD	polynomial kernel of degree 2	.800
COP	COP $k = 3 \cdot \mathcal{D} = 48$	.822
HiCS + LOF	100 candidates per iteration, LOF $k = 10$	.604
LOF	max score for $k = 5 \ldots 25$	.716
SOD	$\alpha = 0.8$, max score for $\ell = k = 5 \ldots 25$	.784
LOGP	$\alpha = 0.1$, min score for $k = 5 \ldots 25$	.831

Table IV
ROC AUC RESULTS ON IMAGE SEGMENTATION DATA SET.

algorithm	key parameters	AUC
ABOD	polynomial kernel of degree 2	.887
COP	COP $k = 3 \cdot \mathcal{D} = 57$	.957
HiCS + LOF	100 candidates per iteration, LOF $k = 10$	.891
LOF	max score for $k = 5 \ldots 25$	.952
SOD	$\alpha = 0.8$, max score for $\ell = k = 5 \ldots 25$	.932
LOGP	$\alpha = 0.1$, min score for $k = 5 \ldots 25$	.968

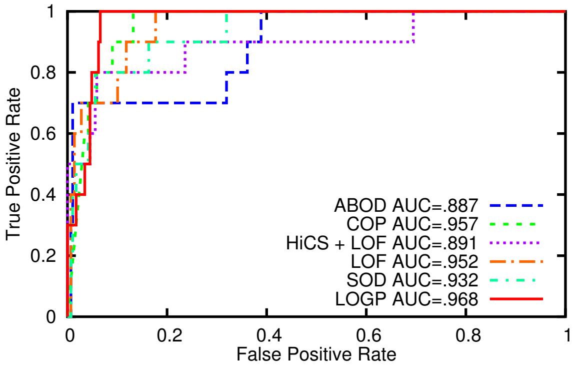

Figure 9. ROC curve performance of all algorithms on the outlier detection over the image segmentation data set.

3×3 region described by 19 attributes. To adapt it into a data set having natural outliers, we keep all instances from two randomly selected classes while sampling 2 instances from each of the 5 remaining classes. This yields 10 outliers and 282 inliers. The English letter recognition data set is much larger which contains 20,000 instances (unique stimuli) from 26 capital letters in the English alphabet. Each instance was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 to 15. Again, to create a data set with natural outliers, we keep all instances from 10 randomly selected letters as regular objects and sample 2 instances from each of the 16 remaining letters as hidden outliers, yielding a data set with 32 outliers and 417 inliers.

Outlier identification: We show in Figures 8 and 9 the ROC curve performance of all algorithms on these two datasets. The results still indicate that LOGP yields the best outlier detection quality for these moderately high dimensional datasets; but both COP and SOD also work remarkably well. It remains an open question if the performance can further be improved by constructing a mixed ensemble of all the methods discussed, as they may likely find different outliers.

Outlier interpretation: In an attempt to explain for the discriminative features, we plot in Figure 10 the top five outliers uncovered by LOGP, selecting from each of the two datasets. Similar to those of the Pen digit data, the x- and y-axes show the feature index and coefficient. As observed from the outliers uncovered from the English letter data, out of 16 original features, a few have been selected as the discriminative features. The first two outliers tend to share the same feature set spanned by $\{$*x-box, box height, x-ege, xegvx, y-ege*$\}$ whereas the third and fifth ones have $\{$*y-bar, x2ybr, x-ege*$\}$ as their discriminative features. These two pairs of instances are the letters H and Y while the 4th outlier is L, which are mostly different from the set of letters (B,C,E,I,J,L,O,S,U,X) that have been selected as the regular objects in the data. It is also interesting to observe the discriminative features discovered for the five outliers in the image segmentation data set, shown in the bottom row of Figure 10. One may see that the space spanned by $\{$*vedgeM, hedgeM, exred, exgreen*$\}$ would be appropriate to explain the exceptional property of the first three outliers whilst the one spanned by $\{$*shortlineD, exblue, exgreen, valueM, hueM*$\}$ is appropriate to discriminate the last two outliers. Taking a closer look to the data, these two types of outliers were indeed exceptional to the two main distributions, the “brickface” and “cement” classes, which have been randomly chosen as the common patterns of the image segmentation data.

V. Conclusion

In this paper, we have developed a novel algorithm that addresses two equally important problems, outlier detection and outlier interpretation. This contrasts with the majority of existing algorithms that often provide solutions only for the problem of outlier detection. Our proposed algorithm takes a mathematical approach from spectral graph theory to learn an optimal subspace in which an outlier is well discriminated from regular objects whereas the intrinsic geometrical structure of the data is retained to assure the outlier explanation quality. We showed that the sets of discriminative features uncovered for hidden outliers are intuitive, meaningful, and human-interpretable which are important properties to enhance the understanding of the hidden outliers. Through experimental analysis on a number of real-world benchmark datasets, we also demonstrated its appealing performance in the outlier detection rate, compared against state-of-the-art algorithms. Along the way, we give further evidence that application of ensemble techniques for outlier detection [3,36] is quite sensible and promising.

References

[1] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek, "Robust, complete, and efficient correlation clustering," in *Proc. SDM*, 2007.

[2] E. Achtert, H.-P. Kriegel, E. Schubert, and A. Zimek, "Interactive data mining with 3D-Parallel-Coordinate-Trees," in *Proc. SIGMOD*, 2013, pp. 1009–1012.

[3] C. C. Aggarwal, "Outlier ensembles [position paper]," *SIGKDD Explor.*, vol. 14, no. 2, pp. 49–58, 2012.

[4] L. Akoglu, M. McGlohon, and C. Faloutsos, "oddball: Spotting anomalies in weighted graphs," in *Proc. PAKDD*, 2010, pp. 410–421.

[5] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[6] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *NIPS*, 2001.

[7] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[8] J. Blitzer, K. Q. Weinberger, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *NIPS*, 2005, pp. 1473–1480.

[9] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. SIGMOD*, 2000, pp. 93–104.

[10] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *IJCAI*, 2007, pp. 708–713.

[11] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM CSUR*, vol. 41, no. 3, pp. Article 15, 1–58, 2009.

[12] S. Chawla, Y. Zheng, and J. Hu, "Inferring the root cause in road traffic anomalies," in *ICDM*, 2012, pp. 141–150.

[13] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *CVPR*, 2005, pp. 846–853.

[14] X. H. Dang and J. Bailey, "A hierarchical information theoretic technique for the discovery of non linear alternative clusterings," in *Proc. KDD*, 2010, pp. 573–582.

[15] D. Dasgupta and N. S. Majumdar, "Anomaly detection in multidimensional data using negative selection algorithm," in *Proc. CEC*, 2002.

[16] A. Foss, O. R. Zaïane, and S. Zilles, "Unsupervised class separation of multivariate data through cumulative variance-based ranking," in *ICDM*, 2009.

[17] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, 2001.

[19] Z. He, S. Deng, and X. Xu, "A unified subspace outlier ensemble framework for outlier detection," 2005, pp. 632–637.

[20] F. Keller, E. Müller, and K. Böhm, "HiCS: high contrast subspaces for density-based outlier ranking," in *Proc. ICDE*, 2012.

[21] E. M. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proc. VLDB*, 1998, pp. 392–403.

[22] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "A general framework for increasing the robustness of PCA-based correlation clustering algorithms," in *Proc. SSDBM*, 2008, pp. 418–435.

[23] ——, "Outlier detection in axis-parallel subspaces of high dimensional data," in *Proc. PAKDD*, 2009, pp. 831–838.

[24] ——, "Interpreting and unifying outlier scores," in *Proc. SDM*, 2011, pp. 13–24.

[25] ——, "Outlier detection in arbitrarily oriented subspaces," in *Proc. ICDM*, 2012, pp. 379–388.

[26] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in *Proc. KDD*, 2008, pp. 444–452.

[27] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in *Proc. KDD*, 2005, pp. 157–166.

[28] A. Martinez and R. Benavente, "The AR face database," CVC, Tech. Rep. 24, 1998.

[29] E. Müller, M. Schiffer, and T. Seidl, "Statistical selection of relevant subspace projections for outlier ranking," in *Proc. ICDE*, 2011, pp. 434–445.

[30] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proc. SIGMOD*, 2000, pp. 427–438.

[31] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel, "On evaluation of outlier rankings and outlier scores," in *Proc. SDM*, 2012, pp. 1047–1058.

[32] E. Schubert, A. Zimek, and H.-P. Kriegel, "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data Min. Knowl. Disc.*, 2012.

[33] K. Smets and J. Vreeken, "The odd one out: Identifying and characterising anomalies," in *Proc. SDM*, 2011, pp. 804–815.

[34] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[35] R. Vilalta and S. Ma, "Predicting rare events in temporal domains," in *ICDM*, 2002, pp. 474–481.

[36] A. Zimek, M. Gaudet, R. J. G. B. Campello, and J. Sander, "Subsampling for efficient and effective unsupervised outlier detection ensembles," in *Proc. KDD*, 2013.

[37] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Stat. Anal. Data Min.*, vol. 5, no. 5, pp. 363–387, 2012.