Learning low-dimensional manifolds under the L0-norm constraint for unsupervised outlier detection

Yoshinao Ishii¹ · Satoshi Koide¹ · Keiichiro Hayakawa¹

Received: 9 July 2020 / Accepted: 1 June 2021 / Published online: 10 July 2021
© The Author(s) 2021

Abstract

Unsupervised outlier detection without the need for clean data has attracted great attention because it is suitable for real-world problems as a result of its low data collection costs. Reconstruction-based methods are popular approaches for unsupervised outlier detection. These methods decompose a data matrix into low-dimensional manifolds and an error matrix. Then, samples with a large error are detected as outliers. To achieve high outlier detection accuracy, when data are corrupted by large noise, the detection method should have the following two properties: (1) it should be able to decompose the data under the L0-norm constraint on the error matrix and (2) it should be able to reflect the nonlinear features of the data in the manifolds. Despite significant efforts, no method with both of these properties exists. To address this issue, we propose a novel reconstruction-based method: “L0-norm constrained autoencoders (L0-AE).” L0-AE uses autoencoders to learn low-dimensional manifolds that capture the nonlinear features of the data and uses a novel optimization algorithm that can decompose the data under the L0-norm constraints on the error matrix. This novel L0-AE algorithm provably guarantees the convergence of the optimization if the autoencoder is trained appropriately. The experimental results show that L0-AE is more robust, accurate and stable than other unsupervised outlier detection methods not only for artificial datasets with corrupted samples but also artificial datasets with well-known outlier distributions and real datasets. Additionally, the results show that the accuracy of L0-AE is moderately stable to changes in the parameter of the constrained term, and for real datasets, L0-AE achieves higher accuracy than the baseline non-robustified method for most parameter values.

Keywords Outlier detection · Unsupervised learning · Autoencoder · Reconstruction-based method · Manifold learning · Alternative optimization

1 Introduction

Outlier detection, which is the identification of samples that are different from the majority (inliers) in the data, is a fundamental problem in data mining and has been studied for a long time [1,16,26]. The demand for outlier detection has further

increased as data collection for various targets has become possible because of recent improvements in sensor and communication technologies [7,14,20]. However, as more data can be obtained, the cost of labeling data has become enormous. Therefore, unsupervised outlier detection without the need for clean data is highly practical and has been important for real problems in recent years [12,32].

Reconstruction-based methods, such as principal component analysis (PCA) [21], are popular methods used for unsupervised outlier detection. These methods assume that outliers are incompressible and thus cannot be effectively reconstructed from low-dimensional projections. PCA linearly decomposes a data matrix X into a low-rank matrix L and an error matrix E that indicates the outlieriness of elements of X such that the l_2 -norm of E is minimized. Because the l_2 -norm is sensitive to outliers, the detection accuracy of PCA for data corrupted by large noise tends to be low. Data corrupted by large noise are hereinafter referred to simply

This paper is an extended version of the article that appeared in PAKDD’2020-“L0-norm constrained Autoencoders for Unsupervised Outlier Detection” [18].

✉ Yoshinao Ishii
e1597@mosk.tytlabs.co.jp
Satoshi Koide
koide@mosk.tytlabs.co.jp
Keiichiro Hayakawa
kei-hayakawa@mosk.tytlabs.co.jp

¹ Toyota Central R&D Lab., Inc., Nagakute, Aichi 480–1192, Japan

as “corrupted data.” Robust PCA (RPCA) [6], which is an improved method of PCA, linearly decomposes X into L and a sparse error matrix S . As the decomposition of RPCA is performed under an l_0 -norm constraint on S , RPCA is robust against corrupted data. However, because both methods decompose X linearly, there is the drawback of poor accuracy for data that have nonlinear features.

In recent years, the robust deep autoencoder (RDA) [34] has been proposed as an unsupervised outlier detection method that can accurately detect outliers, even for corrupted and nonlinear data. In the RDA, a low-dimensional nonlinear manifold that flexibly captures nonlinear features of input data is trained to reconstruct the data using autoencoders (AEs) [15] instead of the linear manifold used in RPCA. The RDA aims to decompose data as $X = L_D + S$ using an alternating optimization algorithm, where L_D is a matrix whose elements are close on the low-dimensional nonlinear manifold and S is a sparse error matrix. To learn a low-dimensional nonlinear manifold that is not affected by corrupted elements, the RDA aims to decompose data under the l_0 -norm regularization on S ; however, the RDA actually sparsifies S by solving the l_1 -norm relaxed regularization problem because of the computational intractability of l_0 -norm regularization. For corrupted data, l_0 -norm regularization, which can avoid the effect of the values of the corrupted elements, is more robust for outliers than l_1 -norm regularization, for which it is difficult to avoid the effects of the values of the corrupted elements (see Sect 5.3.1). Additionally, in the alternating optimization method of the RDA, no theoretical analysis of convergence is made. In practice, it has been confirmed that the progress of training the RDA model may be unstable (see Sect. 5.5).

In this paper, we propose L_0 -norm constrained AE (L0-AE), which is a novel reconstruction-based unsupervised outlier detection method that can learn low-dimensional manifolds under an l_0 -norm constraint for the error term using AE. In Table 1, we briefly summarize the comparison of the features of each reconstruction-based method. Compared with other reconstruction-based methods, L0-AE can provably guarantee the convergence of optimization under the l_0 -norm constraint in addition to considering nonlinear features. To the best of our knowledge, there is no reconstruction-based method that enforces neither relaxation of the l_0 -norm nor linearity. The key contributions of this work are as follows:

1. We propose a new alternating optimization algorithm that decomposes data nonlinearly under an l_0 -norm constraint for the error term (Sect. 3.1).
2. We prove that our alternating optimization algorithm always converges under the mild condition that AE is trained appropriately using gradient-based optimization instead of the RDA (Sect. 3.3).

Table 1 Comparison of features of reconstruction-based methods

	PCA	RPCA	AE	RDA	L0-AE
Decomposition	$X = L + E$	$X = L + S$	$X = \tilde{L} + E$	$X = L_D + S$	$X = \tilde{L} + S + E$
Minimization	$\ X - L\ _2^2$	$\ L\ _* + \lambda\ S\ _1$	$\ X - f_{AE}(X; \theta)\ _2^2$	$\ L_D - f_{AE}(L_D; \theta)\ _2 + \lambda\ S\ _1$	$\ X - f_{AE}(X; \theta) - S\ _2^2$
Constraints	$\text{rank}(L) \leq k'$	$\ X - L - S\ _2^2 = 0$	—	$X - L_D - S = 0$	$\ S\ _0 \leq k$
Convexity	Yes	Yes	No	No	No
Nonlinear model?	No	No	Yes	Yes	Yes
Considering l_0 -norm	No	Yes	No	No	Yes
Convergence of Alternating optim.	—	—	—	Not proved	Guaranteed if AE Is trained appropriately

3. Through extensive experiments, we show that L0-AE is more robust and accurate than other reconstruction-based methods not only for corrupted data but also datasets that contain well-known outlier distributions. Additionally, for real datasets, we show that L0-AE achieves higher accuracy than the baseline non-robustified method for most parameter values and can learn more stably than the RDA (Sect. 5).

2 Preliminaries

In this section, we describe related reconstruction-based methods. Throughout the paper, we denote a given data matrix by $X \in \mathbb{R}^{N \times D}$, where N and D denote the number of samples and feature dimensions of X , respectively.

2.1 Robust principal component analysis

RPCA [6] is a modification of PCA [21].

First, we describe PCA. PCA decomposes a data matrix X into a low-rank matrix L and a small error matrix E such that

$$X = L + E. \quad (1)$$

This decomposition can be formulated as the following optimization problem

$$\begin{aligned} \min_L \|X - L\|_2^2 \\ \text{s.t. } \text{rank}(L) \leq k', \end{aligned} \quad (2)$$

where $\|\cdot\|_2$ is the l_2 -norm and k' is a parameter that determines the maximum value allowed for $\text{rank}(L)$; that is, PCA seeks the best low-rank representation L of the given data matrix X with respect to (w.r.t.) the l_2 -norm. PCA is used for outlier detection because $E \circ E$ indicates element-wise outlierness, where $\circ$ is the element-wise product. Generally, the outlierness of each sample is obtained by adding $E \circ E$ along the feature dimension. PCA is sensitive to outliers because of the use of the l_2 -norm, which makes the use of PCA for outlier detection difficult.

To avoid this drawback of PCA, RPCA decomposes X into a low-rank matrix L and a sparse error matrix S such that

$$X = L + S \quad (3)$$

based on the assumption that there are not many corrupted elements in X . This decomposition problem can be represented as follows: minimize the rank of the matrix L satisfying the sparsity constraint $\|S\|_0 \leq k$, where $\|\cdot\|_0$ is the l_0 -norm that represents the number of nonzero elements and k is a parameter that determines the maximum value

of $\|S\|_0$. The Lagrangian reformulation of this optimization problem is

$$\begin{aligned} \min_{L, S} \text{rank}(L) + \lambda \|S\|_0 \\ \text{s.t. } \|X - L - S\|_2^2 = 0, \end{aligned} \quad (4)$$

where λ is a parameter that adjusts the sparsity of S instead of k . However, this non-convex optimization problem (4) is NP-hard [2]. A convex relaxation whose solution matches the solution of Eq. (4) under broad conditions has been proposed as follows:

$$\begin{aligned} \min_{L, S} \|L\|_* + \lambda \|S\|_1 \\ \text{s.t. } \|X - L - S\|_2^2 = 0, \end{aligned} \quad (5)$$

where $\|\cdot\|_*$ is the nuclear norm and $\|\cdot\|_1$ is the l_1 -norm (i.e., the sum of the absolute values of the elements). RPCA is also used for outlier detection because $S \circ S$ indicates the outlierness of each element.

RPCA is more robust against outliers than PCA because of the use of the l_0 -norm. However, RPCA cannot represent data distributed in a nonlinear manifold because the mapping from X to L is restricted to be linear.

2.2 Robust deep autoencoders

The RDA [34] is a method that relaxes the linearity assumption of RPCA. The RDA uses an AE for nonlinear mapping from X to L instead of the linear mapping of RPCA.

First, we describe the AE. The AE [15] is a special type of multilayer neural network in which the number of nodes in the output layer is the same as that in the input layer. Generally, the model parameters θ of AEs are trained to minimize the reconstruction error, which is defined by the following optimization problem:

$$\min_{\theta} \|X - f_{\text{AE}}(X; \theta)\|_2^2, \quad (6)$$

where $f_{\text{AE}}(X; \theta)$ is an output of the AE with an input X and parameters θ . If an AE with a bottleneck layer can be trained so that its reconstruction error becomes small, $f_{\text{AE}}(X; \theta)$ forms a low-dimensional nonlinear manifold; that is, the AE is a generalization of PCA and decomposes data such that

$$X = \bar{L} + E, \quad (7)$$

where $\bar{L} = f_{\text{AE}}(X; \theta)$ denotes a matrix that forms a low-dimensional nonlinear manifold. The AE can be used as an outlier detection method that is a nonlinear version of PCA. Similar to PCA, the AE is sensitive to outliers because of the l_2 -error minimization.

Next, we describe the RDA. Similar to RPCA, the RDA aims to decompose X into $X = L_D + S$, where S is a sparse error matrix whose nonzero elements indicate the reconstruction difficulty and L_D is easily reconstructable data for the AE. This decomposition of the RDA is defined as the following optimization problem:

$$\begin{aligned} \min_{\theta, S} & \|L_D - f_{AE}(L_D; \theta)\|_2 + \lambda \|S\|_0 \\ \text{s.t.} & X - L_D - S = 0. \end{aligned} \quad (8)$$

This decomposition is more robust than Eq. (6) and can capture nonlinear features of X , unlike Eq. (4).

As Eq. (8) is difficult to optimize, an l_1 -relaxation of the following form has been also proposed [34]:

$$\begin{aligned} \min_{\theta, S} & \|L_D - f_{AE}(L_D; \theta)\|_2 + \lambda \|S\|_1 \\ \text{s.t.} & X - L_D - S = 0. \end{aligned} \quad (9)$$

2.2.1 RDA for structured outliers

The outlier detection methods mentioned above assume that outliers are unstructured. However, in real applications, outliers are often structured [34]; that is, outliers are concentrated on a specific sample. To improve the detection accuracy for data that have structured outliers along the sample dimension, a method with grouped norm regularization has been proposed:

$$\begin{aligned} \min_{\theta, S} & \|L_D - f_{AE}(L_D; \theta)\|_2 + \lambda \|S^T\|_{2,1} \\ \text{s.t.} & X - L_D - S = 0, \end{aligned} \quad (10)$$

where the $l_{2,1}$ -norm is defined as

$$\|X\|_{2,1} = \sum_{j=1}^D \|x_j\|_2 = \sum_{j=1}^D \left(\sum_{i=1}^N |x_{ij}|^2 \right)^{1/2}. \quad (11)$$

The RDA essentially uses Eq. (10) as an objective function for outlier detection.

2.2.2 Alternating optimization algorithm of the RDA

To optimize Eq. (10), the RDA applies an alternating optimization method to θ and S . In particular, the RDA uses a back-propagation method to train AE to minimize $\|L_D - f_{AE}(L_D; \theta)\|_2$ with S fixed and a proximal-gradient-based method with a fixed step size to minimize the penalty term $\|S^T\|_{2,1}$ with θ fixed.

3 L0-norm constrained autoencoders

Although the RDA can detect outliers even for nonlinear data, there are several concerns with the RDA. First, because of the NP-hardness of the problem, the RDA uses the l_1 or $l_{2,1}$ -norm instead of the l_0 -norm, which makes the result highly sensitive to outliers. Second, the alternating optimization method of the RDA does not guarantee convergence. In practice, it has been experimentally confirmed that the progress of training the RDA model may be unstable (Sect. 5.5).

To address these issues, we propose an unsupervised outlier detection method that can decompose data nonlinearly using AE under an l_0 -norm constraint for the sparse matrix S . We prove that our algorithm always converges under a certain condition. For clarity, we first describe L0-AE for unstructured outliers and then extend it for structured outliers.

3.1 Formulation of the objective function

We consider decomposing a data matrix X into three terms: a low-dimensional nonlinear manifold $\tilde{L} = f_{AE}(X; \theta)$, sparse error matrix S and small error matrix E similar to stable principal component pursuit [35], which overcomes the drawbacks of RPCA for noisy data, such that

$$X = f_{AE}(X; \theta) + S + E. \quad (12)$$

Note that we can calculate element-wise outlierness using $(X - f_{AE}(X; \theta)) \circ (X - f_{AE}(X; \theta))$ because $X - f_{AE}(X; \theta)$ equals $S + E$, which denotes the total error matrix.

Equation (12) can be satisfied freely depending on E if E is not constrained. To train an AE that captures the features of X successfully, $\|E\|_2^2$ needs to be as small as possible. Therefore, we aim to obtain an AE model that successfully reconstructs the data by excluding the influence of outliers and a sparse matrix S by solving the following optimization problem that minimizes $\|E\|_2^2 = \|X - f_{AE}(X; \theta) - S\|_2^2$ with l_0 -norm regularization of S :

$$\min_{\theta, S} \|X - f_{AE}(X; \theta) - S\|_2^2 + \lambda \|S\|_0. \quad (13)$$

Specifically, this optimization problem (13) means “minimizing the reconstruction error of the AE without considering the sparse noisy elements S .” Because of the l_0 -norm regularization of S , it is robust against outliers contained in X . However, Eq. (13) is difficult to optimize, as is Eq. (8) in the RDA.

To facilitate the optimization described later, we consider the following l_0 -norm constrained optimization problem

instead of Eq. (13):

$$\begin{aligned} \min_{\theta, S} \|X - f_{\text{AE}}(X; \theta) - S\|_2^2 \\ \text{s.t. } \|S\|_0 \leq k. \end{aligned} \quad (14)$$

Equation (13) can be considered as the Lagrangian reformulation of this optimization problem (14). In unsupervised settings, it is not possible to know the optimal sparsity of S for a problem in advance; therefore, λ and k that adjust the sparsity of S are both hyperparameters. In practical situations, domain experts can often estimate the rate of outliers in the data. In Section 5.3.1, we experimentally show how to set the appropriate $C_p = k/N$, which is the normalized value of k , for the data when the estimated rate of outliers is available. If we can solve Eq. (14), we can obtain $\tilde{L}$, which captures nonlinear features of X and can avoid the influence of outliers completely.

3.2 Alternating optimization algorithm

In the following, we propose an alternating optimization algorithm for θ and S for the l_0 -norm constrained optimization problem (14). We denote the reconstruction error terms $X - f_{\text{AE}}(X; \theta)$ by $Z(\theta)$. Then the objective function can be expressed as $\|Z(\theta) - S\|_2^2$. In the optimization phase of θ with S fixed, we use a gradient-based method. In the optimization phase of S with θ fixed, the optimum S of $\|Z(\theta) - S\|_2^2$ can be obtained using the following method.

When $\|S\|_0 \leq 1$ (S has at most one nonzero element), S that minimizes $\|Z(\theta) - S\|_2^2$ is the matrix that zeroes out the element with the largest absolute value of $Z(\theta)$. Similarly, when $\|S\|_0 \leq k$, S that minimizes $\|Z(\theta) - S\|_2^2$ is the matrix that zeroes out the elements with the top- k largest absolute values in $Z(\theta)$; that is, S that minimizes Eq. (14) with θ fixed can be determined in a closed form as follows:

$$s_{ij} = \begin{cases} z_{ij} & (|z_{ij}| \geq c) \\ 0 & (\text{otherwise}), \end{cases} \quad (15)$$

where z_{ij} is an (i, j) -element of $Z(\theta)$ and c is the k th largest value in $\{|z_{ij}| \mid 1 \leq i \leq N, 1 \leq j \leq D\}$.

Theorem 1 S obtained by Eq. (15) always minimizes Eq. (14) with θ fixed.

Proof To optimize Eq. (14) w.r.t. S under $\|S\|_0 \leq k$, we can choose at most k nonzero elements. We denote a set of subscripts of nonzero elements in S by $H = \{(i_1, j_1), \dots, (i_{k'}, j_{k'})\}$ where $k' \leq k$. The optimality of Eq. (15) is proved by showing the following two facts.

(i) With H fixed, the optimal S is described by $s_{ij} = z_{ij}$ if $(i, j) \in H$ and $s_{ij} = 0$ otherwise.

(ii) The optimal H consists of subscripts that have the top- k largest absolute values in $\{|z_{ij}| \mid 1 \leq i \leq N, 1 \leq j \leq D\}$.

(i) With H fixed, the objective function Eq. (14) is

$$\sum_{(i,j) \in H} (s_{ij} - z_{ij})^2 + \text{const}. \quad (16)$$

Note that, as $|H| \leq k$, we do not need to consider the constraint. Hence, the optimal S is clearly given by $s_{ij} = z_{ij}$ if $(i, j) \in H$ and $s_{ij} = 0$ otherwise.

(ii) We denote the optimal S given H by S_H^* . First, it is clear that, if $H \subset H'$, $S_{H'}^*$ achieves an equal to or better objective value than S_H^* . This is because H' can make more terms zero than H . This implies that if $|H|$ is less than k , we can determine an equivalent or better H' with k elements.¹ Hence, without loss of generality, we can assume $|H| = k$.

Consider an optimal H^* . If we assume that there exists a subscript in H^* that does not correspond to any of the top- k absolute values, then by replacing that subscript with a subscript corresponding to any of the top- k absolute values that is not already included in H^* , we obtain an H' that necessarily yields a lower value for the objective function. Therefore, H^* is not optimal, which is a contradiction. Hence, the optimal H consists of subscripts that have the top- k largest absolute values in $\{|z_{ij}| \mid 1 \leq i \leq N, 1 \leq j \leq D\}$.

From the facts (i) and (ii), S that zeroes out the elements with the top- k largest absolute values in $Z(\theta)$ with θ fixed is the optimal matrix that minimizes the objective function value; that is, S obtained by Eq. (15) always minimizes Eq. (14) with θ fixed. $\square$

Using the fact that the optimal S is described by $s_{ij} = z_{ij}$ if $(i, j) \in H^*$ and $s_{ij} = 0$ otherwise, we rewrite our proposed formulation (14) and alternating optimization method to make it algorithmically concise as follows:

$$\min_{A, \theta} \|A \circ Z(\theta)\|_2^2 \quad \text{s.t. } \|A\|_0 \geq ND - k, \quad (17)$$

where $A \in \{0, 1\}^{N \times D}$ is a binary-valued matrix. In the alternating optimization of Eq. (17), θ is optimized using gradient-based optimization and A is optimized using

$$a_{ij} = \begin{cases} 1 & (|z_{ij}| < c) \\ 0 & (\text{otherwise}). \end{cases} \quad (18)$$

If multiple elements in $Z(\theta)$ match c , ties are broken arbitrarily so that $\|A\|_0 \geq ND - k$.

¹ Note that this holds even if the number of nonzero elements in Z is less than k .

The procedure of our proposed optimization algorithm is as follows:

Input: $X \in \mathbb{R}^{N \times D}$, $k \in [0, N \times D]$ and $Epoch_{\max} \in \mathbb{N}$

Initialize $A \in \mathbb{R}^{N \times D}$ as a zero matrix, epoch counter $Epoch = 0$ and an AE $f_{AE}(\cdot; \theta)$ with randomly initialized parameters.

Repeat the following $Epoch_{\max}$ times:

1. Obtain reconstruction error matrix $Z(\theta)$: $Z(\theta) = X - f_{AE}(X; \theta)$

2. Optimize A with θ fixed:

Get threshold $c = k$ th largest absolute value in $Z(\theta)$ and update A using Eq. (18)

3. Update θ with A fixed:

Minimize $\|A \circ Z(\theta)\|_2^2$ using gradient-based optimization

Return the element-wise outlieriness $R \in \mathbb{R}^{N \times D}$, which is computed as follows:

$$R = (X - f_{AE}(X; \theta)) \circ (X - f_{AE}(X; \theta)). \quad (19)$$

In step 3, the number of iterations in each gradient-based optimization process affects the performance of L0-AE. In practice, the values of a performance indicator and convergence of L0-AE are sufficient if the update occurs once based on a gradient-based optimization method (e.g., Adam [22]) (see Sect. 5). In this case, the total computational cost of L0-AE is the sum of the cost of a normal AE and sorting to obtain the top- k error value.

3.3 Convergence property

In this section, we prove that our alternating optimization algorithm always converges under the assumption that AE is trained appropriately using gradient-based optimization. Here, we denote the objective function $\|A \circ Z(\theta)\|_2^2$ by $K(A, \theta)$ and the variables A and θ at the t th step of each alternating optimization phase by A^t and θ^t , respectively. Under this assumption, the convergence of the proposed alternating optimization method can be shown as follows:

Theorem 2 Suppose $K(A^t, \theta^t)$ is updated to $K(A^{t+1}, \theta^t)$ using Eq. (18), and assume that $K(A^{t+1}, \theta^t) \geq K(A^{t+1}, \theta^{t+1})$ with gradient-based optimization. Then there exists a value $a^\infty \geq 0$ such that $\lim_{t \rightarrow \infty} K(A^t, \theta^t) = a^\infty$.

Proof By updating A using Eq. (18), the obtained A^* minimizes Eq. (17) for any $Z(\theta)$. Hence, for any θ^t , we have $K(A^t, \theta^t) \geq K(A^{t+1}, \theta^t)$. Furthermore, $K(A^{t+1}, \theta^t) \geq K(A^{t+1}, \theta^{t+1})$ holds by assumption, which indicates that $K(A^t, \theta^t) \geq K(A^{t+1}, \theta^{t+1})$. This implies that a sequence $\{K(A^t, \theta^t)\}$ is a monotonically non-increasing and non-negative sequence. Therefore, by applying the monotone convergence theorem [4], there exists a value $a^\infty = \lim_{t \rightarrow \infty} K(A^t, \theta^t) \geq 0$. $\square$

Remark The assumption $K(A^{t+1}, \theta^t) \geq K(A^{t+1}, \theta^{t+1})$ holds when the learning rate of the AE model is sufficiently small. Although this assumption might not hold for a fixed learning rate in practice, L0-AE results in better convergence than the RDA (see Sect. 5.5). Note that, by Theorem 2, our training algorithm is stable, although it does not guarantee that the parameters converge to a stationary point.

3.4 Algorithm for structured outliers

In the following, we describe an alternating optimization algorithm for data with structured outliers along the sample dimension. To detect structured outliers, Eqs. (17) and (18) are, respectively, reformulated as follows:

$$\min_{A, \theta} \|A \circ (X - f_{AE}(X; \theta))\|_2^2 \quad \text{s.t. } \|A\|_{2,0} \geq N - k, \quad (20)$$

$$a_{i \cdot} = \begin{cases} 1 & (\sum_{j=1}^D (z_{ij})^2 < c') \\ 0 & \text{(otherwise)}, \end{cases} \quad (21)$$

where the $l_{2,0}$ -norm is defined as

$$\|A\|_{2,0} = \sum_{i=1}^N \left\| \sum_{j=1}^D a_{ij}^2 \right\|_0, \quad (22)$$

and c' is the k th largest value of the vector $\sum_{j=1}^D (z_{\cdot j})^2$. The sample-wise outlieriness r' is calculated using the $R = [r_{ij}]$ defined by Eq. (19) as follows:

$$r'_i = \sum_{j=1}^D r_{ij}. \quad (23)$$

L0-AE uses this version of the formulation and the alternating optimization method for outlier detection.

As with the update of A using Eq. (18), the update of A using Eq. (21) always minimizes the objective function (20) with θ fixed. The convergence of this algorithm using Eq. (21) is easily proved in a similar manner to Theorem 2.

Remark The concept of our optimization methodology for structured outliers can be regarded as least trimmed squares (LTS) [27], in which the sum of all squared residuals except the largest k squared residuals is minimized.

4 Related work

Recently, highly accurate neural network-based anomaly detection methods, such as AE, variational AE (VAE) and generative adversarial network-based methods [3,28,36],

have been proposed. However, they assume a different problem setting from ours; that is, the training data do not include anomalous data, and finding anomalies in test datasets is the target task. Therefore, these methods do not have a mechanism that excludes outliers during training. In [10], the equivalence of the global optimum of the VAE and RPCA is shown under the condition that a decoder has some type of affinity; however, connections between VAE and RPCA are not shown for general nonlinear activation functions.

The discriminative reconstruction AE (DRAE) [30] has been proposed for unsupervised outlier removal. The DRAE labels samples for which the reconstruction errors exceed a threshold as “outliers” and omits such samples for learning. To appropriately determine the threshold, the loss function of the DRAE has an additional term to separate the reconstruction errors of inliers and outliers. Because of this additional term, the DRAE does not solve an l_0 -norm constrained optimization problem; that is, the learned manifold is affected by outlier values, which degrades the detection performance (see Section 5).

RandNet [9] has been proposed as a method to increase robustness through an ensemble scheme. Although this ensemble may improve robustness, it does not completely avoid the adverse effects of outliers because each AE uses a non-regularized objective function. The deep structured energy-based model [31], which is a robust AE-based method that combines an energy-based model and non-regularized AE, has the same drawback. In [19], a method that simply combines AE and LTS was proposed; however, no theoretical analysis for the combined effects of AE and LTS was presented.

5 Experimental results

5.1 Experimental settings

5.1.1 Datasets

We used both artificial and real datasets. Four types of artificial datasets were used to evaluate the detection performance against three types of well-known outliers [13,17]: global outliers, clustered outliers and local outliers, in addition to outliers contained in image data. Figure 1 illustrates the artificial datasets. In particular, the data in Fig. 1a have large noise, which can be regarded as an example of corrupted data. Figure 1a shows the dataset containing global outliers: We sampled 9,000 inlier samples $(x, 2x^2 - 1) \in \mathbb{R}^2$ where $x \in [-1, 1]$ was sampled uniformly. Furthermore, we sampled 1000 outliers uniformly from $[-1, 1] \times [-1, 1]$ that were regarded as corrupted samples. Figure 1b shows the two-dimensional dataset containing clustered outliers:

Fig. 1 Artificial datasets: black/red points are inliers/outliers in (a)–(c)

We sampled 9000 inlier samples in the same manner as in Fig. 1a. Furthermore, we generated 1000 outliers with a normal distribution at the mean of $(-0.1087, 0.6826)$ (randomly generated central coordinates) and the standard deviation of 0.04. Figure 1c shows the two-dimensional dataset containing the local outliers that were generated by referring to [17]. We sampled a total of 9750 inlier samples. Half of them were uniformly sampled from the center $(0.5, 0.5)$, distance $r \in [0, 0.5]$ and angle $rad \in [0, 2\pi]$ from the center. The other half were uniformly sampled from the center $(-0.5, -0.5)$, $r \in [0.4, 0.5]$ and $rad \in [0, 2\pi]$ from the center. Furthermore, we generated 250 outliers uniformly within a radius of 0.35 from the center $(-0.5, -0.5)$. Figure 1d shows a part of the image dataset that was generated by referring to [34]. This dataset was generated using “MNIST”: The inliers were 4750 randomly selected “4” images, and the outliers were 250 randomly selected images that were not “4.”

As real datasets, we used 11 datasets from Outlier Detection Datasets (ODDS) [25], which are commonly used as benchmarks for outlier detection methods. Additionally, we also used the “kdd99_rev” dataset introduced in [36]. Table 2 summarizes the 12 datasets. Before the experiments, we normalized the values of the datasets by dimension into the range of -1 to 1.

Table 2 Summary of the real datasets

Dataset	Dims.	Samples	Outlier rate [%]
Cardio	21	1831	9.61
Cover	10	286,048	0.96
Kdd99_rev	118	121,597	20.00
Mnist	100	7603	9.21
Musk	166	3062	3.17
Satellite	36	6435	31.64
Satimage-2	36	5803	1.22
Seismic	28	2584	6.58
Shuttle	9	49,097	7.15
Smtp	3	95,156	0.03
Thyroid	6	3772	2.47
Vowels	12	1456	3.43

5.1.2 Evaluation method

Following the evaluation methods in [9,23,24], we compared the area under the receiver operating characteristic curves (AUCs) of the outlier detection accuracy. The evaluation was performed as follows: (1) all samples (whether inlier or outlier) were used for training; (2) the outlierness of each sample was calculated after training; and (3) the AUCs were calculated using outlierness and inlier/outlier labels. Note that we did not need to specify the detection threshold in this evaluation scheme.

5.2 Methods and configurations

Robust PCA (RPCA)

We used the RPCA implemented in [11] as a baseline linear method.

Normal autoencoders (N-AE)

We implemented N-AE with the loss function $\|X - f_{\text{AE}}(X; \theta)\|_2^2$ as the baseline non-regularized detection method. For every AE-based method below, we used common network settings. We used three fully connected hidden layers (with a total of five layers), in which the number of neurons was $\text{ceil}([D, D^{\frac{1}{2}}, D^{\frac{1}{4}}, D^{\frac{1}{2}}, D])$ from the input to the output unless otherwise noted. These were connected as [input layer] $\rightarrow$ [hidden layer1: linear + relu] $\rightarrow$ [hidden layer2: linear] $\rightarrow$ [hidden layer3: linear + relu] $\rightarrow$ [output layer: linear], where “linear” means that the layer is a fully connected layer with no activation function and “linear + relu” means that the layer is a fully connected layer with the rectified linear unit function. We set the mini-batch size to $N/50$ and applied Adam [22] ($\alpha = 0.001$) for optimiza-

tion with $Epoch_{\text{max}} = 400$. We used Eq. (23) to calculate the sample-wise outlierness. To prevent undue advantages to our method (L0-AE) and the other AE-based methods, we searched this architecture by maximizing the average AUC of N-AE.

Robust deep autoencoders (RDA)

We implemented the RDA [34] with Eq. (10). To make the number of loops equal to those of the other AE-based methods, the parameter *inner_iteration*, which is the number of iterations required to optimize AE during one execution of l_1 -norm optimization, was set to 1. We varied the λ values in 10 ways, $\{1 \times 10^{-5}, 2.5 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 2.5 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 2.5 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}\}$, for each experiment with reference to [34] and adopted the result with the highest average AUC.

RDA-Stbl

This baseline was used to confirm the effect of the l_0 -norm constraint of L0-AE. There are three differences between the objective functions of L0-AE and the RDA: (1) the l_0 -norm constrained problem versus the l_1 -norm regularized problem; (2) for the input to AEs, X versus $X - S$; and (3) in the first term of the objective function, $\|\cdot\|_2^2$ versus $\|\cdot\|_2$. To bridge gaps (2) and (3), we introduce RDA-Stbl, which minimizes the objective function $\|L_D - f_{\text{AE}}(X; \theta)\|_2^2 + \lambda \|S^T\|_{2,1}$ such that $X - L_D - S = 0$, w.r.t. S and θ . By comparing the results of this model with those of L0-AE, we can confirm the effects that result from the difference between the l_0 -norm constraint and l_1 -norm regularization. We varied the λ values in 10 ways using the same values as those in the RDA and adopted the result with the highest average AUC.

L0-norm constrained autoencoders (L0-AE)

We used L0-AE for structured outliers (described in Sect. 3.4). The sample-wise outlierness of L0-AE was calculated using Eq. (23). We did not iterate the process to update the parameters of the AE at each gradient-based optimization step. Instead of k , we used $C_p = k/N$ ($0 \leq C_p \leq 1$), which was normalized by the number of samples. This type of normalized parameter is often used in other methods such as the one-class support vector machine (OC-SVM) [29] and isolation forest (IForest) [24]. Note that L0-AE is equivalent to N-AE when $C_p = 0$. We varied C_p from 0.05 to 0.5 in 0.05 increments for each experiment and adopted the result with the highest average AUC.

Variational autoencoder (VAE)

We adopted VAE for our problem setting. Outlierness was computed using the reconstruction probability [3]. Note that the number of output dimensions of hidden layer1 and layer3 is twice that of the other AE-based methods.

Discriminative reconstruction autoencoder (DRAE)

We chose λ , which determines the weight of the term in the objective function for separating the inlier and outlier reconstruction errors (see Sect. 4), as the parameter for the parameter search. We varied the λ values in 10 ways, $\{1 \times 10^{-3}, 2.5 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 2.5 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}, 2.5 \times 10^{-1}, 5 \times 10^{-1}, 1 \times 10^0\}$, for each experiment with reference to [30] and adopted the result with the highest average AUC.

We used Chainer (version 1.21.0) [8] for the implementation of the above AE-based methods. Additionally, we applied the following three conventional methods for a comparison of the AUCs for the real datasets.

One-class support vector machine (OC-SVM)

We used the OC-SVM implemented in scikit-learn and set $kernel = "rbf."$

Local outlier factor (LOF)

We used the LOF [5] implemented in scikit-learn and set " k " for the k -nearest neighbors to 100.

Isolation forest (IForest)

We used the IForest from a Python library *pyod* [33] with " n -estimators" (the number of trees) set to 100.

We tuned the above-mentioned parameters to achieve a high AUC on average over all the real datasets; for the other parameters, we used the recommended (default) values, unless otherwise noted.

5.3 Evaluation results for artificial datasets

5.3.1 Results for the global outlier dataset

First, we evaluated the AUCs and robustness against the outliers of L0-AE and the baseline reconstruction-based methods using the global outlier dataset that has highly corrupted samples (Fig. 1a). We compared the average AUC, in addition to the average outlierness of inlier samples O_{avg}^i , average outlierness of outlier samples O_{avg}^o and ratio O_{avg}^o/O_{avg}^i (a higher value implies that fewer outliers are

close to a low-dimensional manifold). In this experiment, because $D = 2$, we could not set the number of neurons and parameters as mentioned in Sect. 5.2; instead, for N-AE to achieve a high AUC, we used [2, 100, 1, 100, 2], which were obtained empirically.

Table 3 presents the average of these measurements over 20 trials with different initial network weights, and Fig. 2 shows the distributions of outlierness for each method, except VAE, in the first run. Note that "params" in Table 3 indicates the parameters with which each method achieved the highest average AUC. As VAE uses the probability as outlierness, only the AUC was included in Table 3. These results show that L0-AE outperformed the other methods in terms of AUC and the distribution of outlierness between inliers and outliers (i.e., sparsity of the error matrix).

RPCA performed significantly poorer than the other methods because of its linearity. Among the AE-based methods, L0-AE performed best. In L0-AE, we observed that the learned manifold was almost entirely composed of inliers. Therefore, it can be confirmed that the l_0 -norm constraint of L0-AE functions as intended, and L0-AE can learn by almost completely eliminating the influence of corrupted samples while capturing nonlinear features. By contrast, the performances of the other AE-based methods were inferior to that of L0-AE because the other methods cannot completely exclude the influence of corrupted outliers. VAE was less accurate than the other AE-based methods; we consider that VAE was unable to demonstrate robustness because of the non-affinity of the decoder. For the DRAE, the reconstruction errors of inliers and outliers were relatively well separated, but the DRAE was more strongly affected by outliers than L0-AE because the DRAE objective function depends on how large outliers are, whereas the L0-AE objective function does not.

Next, we evaluated the parameter sensitivity of L0-AE using the artificial dataset with global outliers (Fig. 1a). We used 10 different datasets in which the outlier rate in the dataset was changed to $\{0.05, 0.1, 0.15, \dots, 0.45, 0.5\}$. The parameter sensitivity was evaluated based on the change in the average AUC over 10 trials with different random seeds when C_p was changed from 0.05 to 0.55 in 0.05 increments for each dataset.

Figure 3 shows the AUCs with different C_p values for each dataset. From the results, we confirmed that, even if C_p was changed significantly, the AUCs did not change significantly and remained above 80%. Therefore, we can say that the quality of the distribution of outlierness obtained by L0-AE is robust to changes in C_p . For all results, the maximum AUC values occurred at C_p values moderately greater than the true outlier rates. When C_p was moderately greater than the true outlier rate, we considered that outliers were safely detected as outliers; therefore, we believe that the highest AUCs could be achieved. When C_p was less than the true outlier rate, the AUCs were better than in the case of N-

Table 3 Average measurements from reconstruction-based methods for global outliers

Methods	RPCA	N-AE	RDA	RDA-Stbl	VAE	DRAE	L0-AE
AUC[%]	39.9	93.9	97.8	96.5	79.4	97.4	99.8
O_{avg}^i	1.30E-04	3.90E-03	7.12E-03	7.68E-03	—	6.57E-04	1.27E-05
O_{avg}^o	9.80E-05	1.60E-01	3.16E-01	1.70E-01	—	1.63E-01	4.56E-01
O_{avg}^o/O_{avg}^i	0.75	42.09	44.41	22.14	—	247.60	35965.54
params	—	—	1.0E-05	1.0E-05	—	1.0E-01	0.2

Bold indicates the best results of all methods

Fig. 2 Distributions of outlierness for each method in the first run for global outliers: sample IDs of inliers and outliers are 1~9000 and 9001~10,000, respectively **Fig. 3** Parameter sensitivity of L0-AE for the global outlier datasets with 10 outlier rates; each error bar represents a 95% confidence interval, gray vertical lines indicate the true outlier rates of each dataset, and gray horizontal lines indicate the AUCs of a normal AE ($C_p = 0$) for each dataset

AE ($C_p = 0$) because some outliers were not reconstructed. Therefore, in practical use, when the true outlier rate can be approximately estimated by a domain expert, it is appropriate to set C_p to a value moderately greater than the estimated outlier rate. On the basis of this experiment and experiments on various datasets presented below, we recommend setting C_p to a value approximately 0.1 greater than the estimated outlier rate. By contrast, for the parameters in other robustified reconstruction-based methods, it is difficult to determine which value should be set even if the true outlier rate is accurately estimated by a domain expert because they are abstract in that they determine the ratio between the reconstruction error and the regularization term.

5.3.2 Results for the clustered, local and image outlier datasets

In the following, we evaluate L0-AE for the artificial datasets other than the global outlier dataset. First, we compare the AUCs and the distributions of outlierness. Table 4 presents the average of the measurements over 20 trials with different random seeds and the parameters for achieving the maximum average AUC for the three datasets (Figs. 1b–d). For the clustered and local outlier datasets, we used the same number of neurons used for the global outlier dataset. Figures 4, 5 and 6 show the distributions of outlierness for each method in the first run under the best parameters for the clustered, local and image outlier datasets, respectively.

Table 4 Average measurements from reconstruction-based methods for three types of outliers

Type	Measure	RPCA	N-AE	RDA	RDA-Stbl	VAE	DRAE	L0-AE
Clustered (rate: 0.10)	AUC[%]	13.8	79.9	83.0	83.3	87.2	97.5	100.0
	O_{avg}^i	1.4E-04	5.7E-03	1.4E-02	6.2E-03	—	2.3E-03	3.3E-03
	O_{avg}^o	1.8E-06	3.8E-03	9.7E-03	3.7E-03	—	7.6E-01	1.9E+00
	O_{avg}^o / O_{avg}^i	0.01	0.68	0.71	0.59	—	336.02	565.35
	params	—	—	2.5E-03	5.0E-03	—	1.0E-03	0.25
Local (rate:0.025)	AUC[%]	37.4	82.8	83.7	84.7	70.2	85.1	85.9
	O_{avg}^i	1.6E-04	2.2E-02	2.4E-02	2.9E-02	—	2.5E-02	2.6E-02
	O_{avg}^o	4.0E-05	6.2E-02	8.5E-02	1.1E-01	—	8.6E-02	9.8E-02
	O_{avg}^o / O_{avg}^i	0.26	2.88	3.55	3.67	—	3.45	3.76
	params	—	—	1.0E-03	1.0E-04	—	1.0E-03	0.10
Image (rate: 0.05)	AUC[%]	91.5	91.8	92.8	91.8	83.1	93.2	93.8
	O_{avg}^i	3.3E+01	4.5E+01	7.1E+01	6.1E+01	—	4.4E+01	4.3E+01
	O_{avg}^o	9.8E+01	1.1E+02	1.8E+02	1.6E+02	—	1.9E+02	2.0E+02
	O_{avg}^o / O_{avg}^i	2.95	2.43	2.57	2.63	—	4.28	4.68
	params	—	—	1.0E-03	1.0E-05	—	2.5E-01	0.10

Bold indicates the best results for each dataset

Fig. 4 Distributions of outlierness for clustered outliers in the first run: sample IDs of inliers and outliers are 1~9000 and 9001~10,000, respectively **Fig. 5** Distributions of outlierness for local outliers in the first run: sample IDs of inliers and outliers are 1~9750 and 9751~10,000, respectively **Fig. 6** Distributions of outlierness for image outliers in the first run: sample IDs of inliers and outliers are 1~4750 and 4751~5000, respectively

Fig. 7 Parameter sensitivity of L0-AE for the clustered, local and image outlier datasets; each error bar represents a 95% confidence interval, gray vertical lines indicate the true outlier rates of each dataset, and gray horizontal lines indicate the AUCs of a normal AE ($C_p = 0$) for each dataset

For the clustered outlier dataset, we can see that L0-AE completely separated the distributions of outlieriness of inliers and outliers. This is likely to be because of the appropriate C_p settings and the l_0 -norm constraint, which allowed the manifold to be learned by excluding the effects of all clustered outliers. By contrast, the other methods had a lower average AUC than L0-AE because the learned manifolds were affected by the outliers as a result of the nonuse of the l_0 -norm constraint.

For the local and image outlier datasets, our L0-AE again had the best performance measures, as shown in Table 4, although the degree of improvement was smaller than that of the clustered outlieriness experiments. A possible reason for this is that there was little room for improvement for the other robustified AE-based methods because the network structure, which was set up to provide the highest accuracy of unconstrained N-AE for fairness, was not suitable for the other methods. Thus, we verified that L0-AE can learn nonlinear manifolds better than other reconstruction-based methods for all cases, avoiding the influence of outliers.

Next, we evaluated the parameter sensitivity of L0-AE for the three datasets (Figs. 1b–d). Figure 7c shows the AUCs with different C_p values for each dataset. As with the global outlier dataset, the AUCs were moderately robust to changes in C_p , and the maximum AUC values were achieved at C_p values moderately greater than the true outlier rates. A similar trend was observed for all well-known outliers, thereby confirming the utility of L0-AE. The development of an automatic optimal C_p search method under the l_0 -norm constraint without ground truth outlier rates or labels is important future work. For the image outlier dataset, the AUC did not decrease as C_p increased. Given the high AUC of RPCA, which is a linear method, for the image dataset, it is likely that the “4” images in the MNIST dataset were confined to a linear subspace and well separated from the images of the other classes. Therefore, even though C_p was high and the number of inlier samples to be reconstructed reduced, we believe that L0-AE can learn manifolds such that the entire inlier sample can be successfully reconstructed.

Table 5 Comparison of AUCs [%] with standard deviation

Methods	Cardio	Cover	Kdd99	Mnist	Musk	Satel.	Satim.	Seismic	Shuttle	Smtp	Thyroid	Vowels	Avg. AUC	Avg. Rank	Avg. Time [s]
RPCA	93.8±0.0	95.9 ±0.0	25.7±0.0	82.3±0.0	97.0±0.0	61.9±0.0	98.0±0.0	68.0±0.0	97.7±0.0	80.0±0.0	93.1±0.0	88.8±0.0	81.86	4.92	3.4
N-AE	80.7±6.9	87.9±9.9	14.5±4.1	83.7±1.0	67.7±12.4	64.6±1.1	78.1±2.3	70.8±2.5	77.0±16.9	75.9±6.6	89.1±2.2	90.1±2.5	73.34	6.75	57.6
RDA	83.4±6.9	88.6±8.0	13.3±2.6	82.7±3.8	70.9±15.5	63.2±2.3	77.7±8.9	72.1 ±2.1	80.1±17.9	77.9±6.7	89.9±5.3	90.8±3.3	74.21	6.08	66.4
RDA-Stbl	86.0±8.1	87.1±11.6	12.7±3.4	82.6±1.3	68.0±11.2	63.7±1.6	77.6±3.0	71.8±1.3	74.6±18.2	74.7±5.2	90.0±3.0	89.1±2.5	73.13	7.25	77.5
VAE	72.1±4.4	81.5±4.6	55.1±23.3	81.7±14.7	34.3±33.8	67.7±5.5	94.1±5.0	69.5±3.1	98.3±3.0	87.1±4.0	97.0±2.1	83.9±3.3	76.86	5.92	139.8
DRAE	89.5±6.9	91.5±1.1	85.4±25.8	84.4±2.6	100.0 ±0.0	65.8±3.2	90.1±12.6	58.6±10.8	78.9±14.5	83.5±6.2	94.2±2.5	90.1±3.4	84.35	4.50	2869.3
L0-AE	94.0 ±3.7	87.2±6.4	96.2 ±1.2	86.2 ±2.2	100.0 ±0.0	75.6 ±3.1	99.4 ±2.4	66.9±3.0	97.7±6.1	89.0±5.2	93.8±3.6	90.9±2.9	89.74	2.92	74.5
OC-SVM	93.0±0.0	91.8±0.0	81.5±0.0	82.0±0.0	93.1±0.0	59.9±0.0	98.0±0.0	59.3±0.0	98.3±0.0	76.9±0.0	85.0±0.0	57.3±0.0	81.35	6.00	668.6
LOF	85.1±0.0	60.1±0.0	36.3±0.0	80.3±0.0	84.0±0.0	56.7±0.0	96.3±0.0	59.2±0.0	56.8±0.0	87.6±0.0	96.3±0.0	93.4 ±0.0	74.33	6.58	46.6
IForest	92.2±1.3	87.2±3.2	78.1±2.9	79.8±2.0	99.9±0.1	70.5±1.8	99.3±0.1	67.3±0.6	99.7 ±0.1	90.7 ±0.6	97.7 ±0.3	75.4±2.4	86.49	4.08	2.3

Bold indicates the best results for each dataset

Fig. 8 Parameter sensitivity of L0-AE for real datasets; each error bar represents a 95% confidence interval, gray vertical lines indicate the true outlier rates of each dataset, and gray horizontal lines indicate the AUCs of a normal AE ($C_p = 0$) for each dataset

Fig. 9 Examples of the transition of the values of the objective functions from RDA and L0-AE

Table 6 Number of epochs in which the objective function value increased over the previous epoch (summed over all datasets)

N-AE	L0-AE $C_p =$					RDA $\lambda =$					
	0.1	0.2	0.3	0.4	0.5	0.000025	0.0001	0.0005	0.0025	0.01	
0	0	0	9	0	0	3152	2031	1323	777	15	

5.4 Evaluation results for real datasets

First, we compared the AUCs for the real datasets. The AUC values were averaged over 50 trials with different random seeds. For the RDA, we used $\lambda = 5.0 \times 10^{-5}$; for RDA-stbl, $\lambda = 5.0 \times 10^{-4}$; for the DRAE, $\lambda = 1.0 \times 10^{-1}$; and for L0-AE, $C_p = 0.3$. Table 5 presents the average AUCs for each dataset; *Avg. AUC*, *Avg. rank* and *Avg. time* refer to the average AUC, average rank over the datasets and average run-time, respectively.

L0-AE achieved the highest average AUC and average rank. Among the reconstruction-based methods, L0-AE achieved the highest AUCs for eight out of 12 datasets. Particularly on kdd99_rev, the AUC of L0-AE was considerably higher than those of the other AE-based methods. Because kdd99_rev had a high rate of outliers and they were distributed close to each other, the methods with l_1 -norm regularization and no regularization could not avoid reconstructing the outliers, whereas L0-AE almost completely avoided reconstruction because of its l_0 -norm constraint. Furthermore, we observed that the AUCs of the RDA and

RDA-Stbl were nearly equal. This shows the importance of the l_0 -norm constraint. L0-AE outperformed the DRAE on average; we consider that L0-AE selectively reconstructs only the inliers, whereas the DRAE reconstructs inliers and reduces the variance of each label, thereby allowing outliers to affect manifolds. Additionally, the computational cost of the DRAE was higher than that of L0-AE, because of the calculation of the threshold. For VAE, training was unstable for some datasets. One possible reason is that VAE involves random sampling in the reparametrization trick, which increased the randomness of the results under these experimental settings. By contrast, among the AE-based methods, L0-AE achieved stable results. The RPCA results were relatively good for some datasets, which suggests that these datasets had linear features and l_0 -norm regularization worked; L0-AE performed well by capturing nonlinear features, even for the other datasets. The reason that RPCA outperformed some AE-based methods on average is that RPCA automatically detects the rank of the inlier, whereas the AE-based methods have a fixed latent dimension (there is no known method for

obtaining an appropriate latent dimension in an unsupervised setting).

Next, we evaluated the parameter sensitivity of L0-AE using real datasets. Figure 8 shows the AUCs with different C_p values for L0-AE (averaged over 50 trials). For most C_p , the AUCs of L0-AE were higher than those of the normal AE ($C_p = 0$), that is, the baseline of the non-robustified method. Therefore, even when the appropriate C_p is unknown, L0-AE is still practical enough to be used. Additionally, as in the case of the artificial datasets, the maximum AUC values essentially occurred at C_p values moderately greater than the true outlier rates. Therefore, in practical terms, it is easier to select an appropriate C_p if the true outlier rate can be estimated.

5.5 Evaluation of convergence

We compared the convergence of L0-AE with that of the RDA. Here, we did not use mini-batch training to remove the effect of randomness. Table 6 presents the sum of the number of epochs in which the value of the objective function increased over the previous epoch during 20 trials (96,000 epochs in total). The results of N-AE are also included for reference. Additionally, Fig. 9 shows two transition examples of the values of the objective functions of the RDA and L0-AE. Among them, Fig. 9d shows the results of the only trial in which the objective function increased in L0-AE; the epochs in which the objective function increased were 294–302 when $C_p = 0.3$, with an average increase of 0.23, which is considerably less than the value of the objective function. In Table 6 and Fig. 9, we observe that L0-AE converged well, regardless of the parameter C_p , unlike the RDA. This empirically demonstrates the validity of Theorem 2, which states that our alternating optimization algorithm converges when the gradient-based optimization behaves ideally. For the RDA, when λ was small, the value of the objective function was unstable, but when λ was large, it was stable. We believe that this is because the sparsity of S improves as λ increases, thus reducing the gap in the objective function between phases of alternating optimization. We observe that, with N-AE, the values of the objective function did not increase, which implies that our gradient-based optimization generally satisfies the assumption in Theorem 2.

6 Conclusion

In this paper, we proposed L0-AE for unsupervised outlier detection. L0-AE decomposes data nonlinearly into low-dimensional manifolds that capture the nonlinear features of the data and a sparse error matrix under the l_0 -norm constraint. We proposed an efficient alternating optimization algorithm for training L0-AE and proved that this algorithm

converges under a mild condition. We conducted extensive experiments with real and artificial datasets and confirmed that L0-AE is highly robust to outliers. We also confirmed that this high robustness leads to higher outlier detection accuracy, defined by the AUC, than those of existing unsupervised outlier detection methods.

Data Availability All data used in the paper are clearly marked with the sources or methods of preparation.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Aggarwal, C.C.: Outlier analysis. data mining. Springer, Cham (2015)
2. Amaldi, E., Kann, V.: On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theor. Computer Sci.* **209**(1–2), 237–260 (1998)
3. An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. Tech. Rep. 1, Special Lecture on IE (2015)
4. Bibby, J.: Axiomatisations of the average and a further generalisation of monotonic sequences. *Glasgow Math. J.* **15**(1), 63–65 (1974)
5. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. *ACM Sigmod Record ACM* **29**, 93–104 (2000)
6. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? *J. ACM* **58**(3), 895 (2011)
7. Capelleveen, G., Poel, M., Mueller, R., Thornton, D., Hillegersberg, J.: Intrusion detection system (ids): anomaly detection using outlier detection approach. *Proc. Computer Sci.* **21**, 18–31 (2016)
8. Chainer: A flexible framework for neural networks. <https://chainer.org/> (2019)
9. Chen, J., Sathe, S., Aggarwal, C., Turaga, D.: Outlier detection with autoencoder ensembles. In: *ICDM'17*. pp. 90–98. SIAM (2017)
10. Dai, B., Wang, Y., Aston, J., Hua, G., Wipf, D.: Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *J. Mach. Learn. Res.* **19**(1), 1573–1614 (2018)
11. Ganguli, D.: Github - dganguli/robust-pca: A simple python implementation of r-pca. <https://github.com/dganguli/robust-pca> (2019)

12. Goldstein, M., Uchida, S.: A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one* **11**(4), 897 (2016)
13. Goldstein, M., Uchida, S.: A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one* **11**(4), 859 (2016)
14. Guo, J., Huang, W., Williams, B.M.: Real time traffic flow outlier detection using short-term traffic conditional variance prediction. *Transp. Res. Part C: Emerg. Technol.* **50**, 160–172 (2015)
15. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006)
16. Hodge, V., Austin, J.: A survey of outlier detection methodologies. *Artif. Intell. Rev.* **22**(2), 85–126 (2004)
17. Huang, H., Qi, H., Shinjae, Y., Dantong, Y.: Local anomaly descriptor: a robust unsupervised algorithm for anomaly detection based on diffusion space. In: *Proceedings of the 21st ACM international conference on Information and knowledge management*. pp. 405–414 (2012)
18. Ishii, Y., Koide, S., Hayakawa, K.: L0-norm constrained autoencoders for unsupervised outlier detection. In: *PAKDD'20*. pp. 674–687. Springer, Cham (2020)
19. Ishii, Y., Takanashi, M.: Low-cost unsupervised outlier detection by autoencoders with robust estimation. *J. Inf. Process.* **27**, 335–339 (2019)
20. Jabez, J., Muthukumar, B.: Intrusion detection system (ids): anomaly detection using outlier detection approach. *Proc. Computer Sci.* **48**, 338–346 (2015)
21. Jolliffe, I.: *Principal component analysis*. Springer, Berlin Heidelberg (2011)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
23. Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Loop: local outlier probabilities. In: *CIKM'09*. pp. 1649–1652. ACM (2009)
24. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: *ICDM'08*. pp. 413–422. IEEE (2008)
25. Rayana, S.: Odds library. <http://odds.cs.stonybrook.edu> (2019)
26. Rousseeuw, P.J., Leroy, A.M.: *Robust regression and outlier detection*. Wiley, New York (1987)
27. Ruppert, D., Carroll, R.J.: Trimmed least squares estimation in the linear model. *J. Am. Stat. Assoc.* **75**(372), 828–838 (1980)
28. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: *IPMI*. pp. 146–157. Springer (2017)
29. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. *Neural Comput.* **13**(7), 1443–1471 (2001)
30. Xia, Y., Cao, X., Wen, F., Hua, G., Sun, J.: Learning discriminative reconstructions for unsupervised outlier removal. In: *IEEE ICCV'15*. pp. 1511–1519 (2015)
31. Zhai, S., Cheng, Y., Lu, W., Zhang, Z.: Deep structured energy based models for anomaly detection. *arXiv preprint arXiv:1605.07717* (2016)
32. Zhang, J., Zulkernine, M.: Anomaly based network intrusion detection with unsupervised outlier detection. In: *IEEE International Conference on Communications*. vol. 5 (2006)
33. Zhao, Y.: pyod. <http://pyod.readthedocs.io/en/latest/> (2019)
34. Zhou, C., Paffenroth, R.C.: Anomaly detection with robust deep autoencoders. In: *KDD'17*. pp. 665–674 (2017)
35. Zhou, Z., Li, X., Wright, J., Candes, E., Ma, Y.: Stable principal component pursuit. In: *ISIT'10*. pp. 1518–1522. IEEE (2010)
36. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: *ICLR'18* (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.