Multi-view Anomaly Detection via Robust Probabilistic Latent Variable Models

Tomoharu Iwata
NTT Communication Science Laboratories
iwata.tomoharu@lab.ntt.co.jp

Makoto Yamada
Kyoto University
makoto.m.yamada@ieee.org

Abstract

We propose probabilistic latent variable models for multi-view anomaly detection, which is the task of finding instances that have inconsistent views given multi-view data. With the proposed model, all views of a non-anomalous instance are assumed to be generated from a single latent vector. On the other hand, an anomalous instance is assumed to have multiple latent vectors, and its different views are generated from different latent vectors. By inferring the number of latent vectors used for each instance with Dirichlet process priors, we obtain multi-view anomaly scores. The proposed model can be seen as a robust extension of probabilistic canonical correlation analysis for noisy multi-view data. We present Bayesian inference procedures for the proposed model based on a stochastic EM algorithm. The effectiveness of the proposed model is demonstrated in terms of performance when detecting multi-view anomalies.

1 Introduction

There has been great interest in multi-view learning, in which data are obtained from various information sources. In a wide variety of applications, data are naturally comprised of multiple views. For example, an image can be represented by color, texture and shape information; a web page can be represented by words, images and URLs occurring on in the page; and a video can be represented by audio and visual features. In this paper, we consider the task of finding anomalies in multi-view data. The task is called horizontal anomaly detection [13], or multi-view anomaly detection [16]. Anomalies in multi-view data are instances that have inconsistent features across multiple views.

Multi-view anomaly detection can be used for many applications, such as information disparity management [9], purchase behavior analysis [13], malicious insider detection [16], and user aggregation from multiple databases. In information disparity management, multiple views can be obtained from documents written in different languages such as Wikipedia. Multi-view anomaly detection tries to find documents that contain different information across different languages, which would be helpful for editors to select documents to be updated, or beneficial for cultural anthropologists to analyze social difference across different languages. In purchase behavior analysis, multiple views for each item can be defined as its genre and its purchase history, i.e. a set of users who purchased the item. Multi-view anomaly detection can find movies inconsistently purchased by users based on the movie genre, which would assist creating marketing strategies.

Multi-view anomaly detection is different from standard (single-view) anomaly detection. Single-view anomaly detection finds instances that do not conform to expected behavior [6]. Figure 1 (a) shows the difference between a multi-view anomaly and a single-view anomaly in a two-view data set. 'M' is a multi-view anomaly since 'M' belongs to different clusters in different views ('A–D' cluster in View 1 and 'E–J' cluster in View 2) and views of 'M' are not consistent. 'S' is a single-view anomaly since 'S' is located far from other instances in each view. However, both views of 'S' have the same relationship with the others (they are far from the other instances), and then 'S'
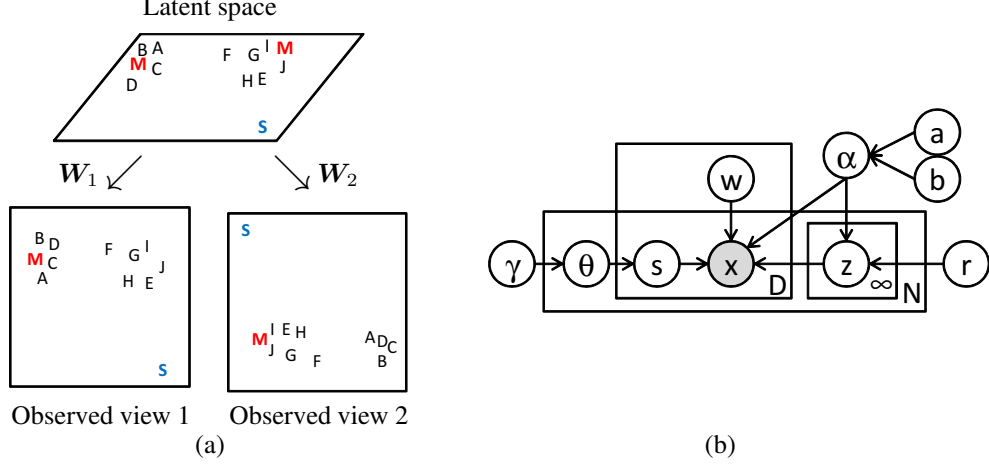
Figure 1: (a) A multi-view anomaly 'M' and a single-view anomaly 'S' in a two-view data set. Each letter represents an instance, and the same letter indicates the same instance. $\boldsymbol{W}_d$ is a projection matrix for view d. (b) Graphical model representation of the proposed model.

is *not* a multi-view anomaly. Single-view anomaly detection methods, such as one-class support vector machines [18] or tensor-based anomaly detection [11], consider that 'S' is anomalous. On the other hand, we would like to develop a multi-view anomaly detection method that detects 'M' as anomaly, but not 'S'. Note that although single-view anomalies are uncommon instances, multi-view anomalies can be majority if they are inconsistent across multiple views.

We propose a probabilistic latent variable model for multi-view anomaly detection. With the proposed model, there is a latent space that is shared across all views. We assume that all views of a non-anomalous (normal) instance are generated using a single latent vector. On the other hand, an anomalous instance is assumed to have multiple latent vectors, and its different views are generated using different latent vectors, which indicates inconsistency across different views of the instance. Figure 1 (a) shows an example of a latent space shared by the two-view data. Two views of every non multi-view anomaly can be generated from a latent vector using view-dependent projection matrices. On the other hand, since two views of multi-view anomaly 'M' are not consistent, two latent vectors are required to generate the two views using the projection matrices.

Since the number of latent vectors for each instance is unknown, we automatically infer it from the given data by using Dirichlet process priors. The inference of the proposed model is based on a stochastic EM algorithm. In the E-step, a latent vector is assigned for each view of each instance using collapsed Gibbs sampling while analytically integrating out latent vectors. In the M-step, projection matrices for mapping latent vectors into observations are estimated by maximizing the joint likelihood. By alternately iterating E- and M-steps, we infer the number of latent vectors used in each instance and calculate its anomaly score from the probability of using more than one latent vector.

2 Proposed Model

Suppose that we are given N instances with D views $\boldsymbol{X} = \{\boldsymbol{X}_n\}_{n=1}^N$, where $\boldsymbol{X}_n = \{\boldsymbol{x}_{nd}\}_{d=1}^D$ is a set of multi-view observation vectors for the nth instance, and $\boldsymbol{x}_{nd} \in \mathbb{R}^{M_d}$ is the observation vector of the dth view. The task is to find anomalous instances that have inconsistent observation features across multiple views. We propose a probabilistic latent variable model for this task. The proposed model assumes that each instance has potentially a countably infinite number of latent vectors $\boldsymbol{Z}_n = \{\boldsymbol{z}_{nj}\}_{j=1}^\infty$, where $\boldsymbol{z}_{nj} \in \mathbb{R}^K$. Each view of an instance $\boldsymbol{x}_{nd}$ is generated depending on a view-specific projection matrix $\boldsymbol{W}_d \in \mathbb{R}^{M_d \times K}$ and a latent vector $\boldsymbol{z}_{ns_{nd}}$ that is selected from a set of latent vectors $\boldsymbol{Z}_n$. Here, $s_{nd} \in \{1, \cdots, \infty\}$ is the latent vector assignment of $\boldsymbol{x}_{nd}$. When the instance is non-anomalous and all its views are consistent, all of the views are generated from a single latent vector. In other words, the latent vector assignments for all views are the same, $s_{n1} = s_{n2} = \cdots = s_{nD}$. When it is an anomaly and some views are inconsistent, different views

2

are generated from different latent vectors, and some latent vector assignments are different, i.e. $s_{nd} \neq s_{nd'}$ for some $d \neq d'$.

Specifically, the proposed model is an infinite mixture model, where the probability for the dth view of the nth instance is given by

$$p(\boldsymbol{x}_{nd}|\boldsymbol{Z}_n, \boldsymbol{W}_d, \boldsymbol{\theta}_n, \alpha) = \sum_{j=1}^{\infty} \theta_{nj} \mathcal{N}(\boldsymbol{x}_{nd}|\boldsymbol{W}_d \boldsymbol{z}_{nj}, \alpha^{-1}\boldsymbol{I}), \tag{1}$$

where $\boldsymbol{\theta}_n = \{\theta_{nj}\}_{j=1}^{\infty}$ are the mixture weights, θ_{nj} represents the probability of choosing the jth latent vector, α is a precision parameter, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{I}$ is the identity matrix. Information of non-anomalous instances that cannot be handled by a single latent vector is modeled in Gaussian noise which is controlled by α. Since we assume the same observation noise α across different views, the observations need to be normalized. We use a Dirichlet process for the prior of mixture weight $\boldsymbol{\theta}_n$. Its use enables us to automatically infer the number of latent vectors for each instance from the given data.

The complete generative process of the proposed model for multi-view instances $\boldsymbol{X}$ is as follows,

1. Draw a precision parameter $\alpha \sim \text{Gamma}(a, b)$
2. For each instance: $n = 1, \ldots, N$
   (a) Draw mixture weights $\boldsymbol{\theta}_n \sim \text{Stick}(\gamma)$
   (b) For each latent vector: $j = 1, \ldots, \infty$
       i. Draw a latent vector $\boldsymbol{z}_{nj} \sim \mathcal{N}(\boldsymbol{0}, (\alpha r)^{-1}\boldsymbol{I})$
   (c) For each view: $d = 1, \ldots, D$
       i. Draw a latent vector assignment $s_{nd} \sim \text{Discrete}(\boldsymbol{\theta}_n)$
       ii. Draw an observation vector $\boldsymbol{x}_{nd} \sim \mathcal{N}(\boldsymbol{W}_d \boldsymbol{z}_{ns_{nd}}, \alpha^{-1}\boldsymbol{I})$

Here, $\text{Stick}(\gamma)$ is the stick-breaking process [19] that generates mixture weights for a Dirichlet process with concentration parameter γ, and r is the relative precision for latent vectors. α is shared for observation and latent vector precision because it makes it possible to analytically integrate out α as shown in (4). Figure 1 (b) shows a graphical model representation of the proposed model, where the shaded and unshaded nodes indicate observed and latent variables, respectively.

The joint probability of the data $\boldsymbol{X}$ and the latent vector assignments $\boldsymbol{S} = \{\{s_{nd}\}_{d=1}^{D}\}_{n=1}^{N}$ is given by

$$p(\boldsymbol{X}, \boldsymbol{S}|\boldsymbol{W}, a, b, r, \gamma) = p(\boldsymbol{S}|\gamma)p(\boldsymbol{X}|\boldsymbol{S}, \boldsymbol{W}, a, b, r), \tag{2}$$

where $\boldsymbol{W} = \{\boldsymbol{W}_d\}_{d=1}^{D}$. Because we use conjugate priors, we can analytically integrate out mixture weights $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_n\}_{n=1}^{N}$, latent vectors $\boldsymbol{Z}$, and precision parameter α. Here, we use a Dirichlet process prior for multinomial parameter $\boldsymbol{\theta}_n$, and a Gaussian-Gamma prior for latent vector $\boldsymbol{z}_{nj}$. By integrating out mixture weights $\boldsymbol{\Theta}$, the first factor is calculated by

$$p(\boldsymbol{S}|\gamma) = \prod_{n=1}^{N} \frac{\gamma^{J_n} \prod_{j=1}^{J_n}(N_{nj}-1)!}{\gamma(\gamma+1)\cdots(\gamma+D-1)}, \tag{3}$$

where N_{nj} represents the number of views assigned to the jth latent vector in the nth instance, and J_n is the number of latent vectors of the nth instance for which $N_{nj} > 0$. By integrating out latent vectors $\boldsymbol{Z}$ and precision parameter α, the second factor of (2) is calculated by

$$p(\boldsymbol{X}|\boldsymbol{S}, \boldsymbol{W}, a, b, r) = (2\pi)^{-\frac{N\sum_d M_d}{2}} r^{\frac{K\sum_n J_n}{2}} \frac{b^a}{b'^{a'}} \frac{\Gamma(a')}{\Gamma(a)} \prod_{n=1}^{N} \prod_{j=1}^{J_n} |\boldsymbol{C}_{nj}|^{\frac{1}{2}}, \tag{4}$$

where

$$a' = a + \frac{N\sum_{d=1}^{D} M_d}{2}, \quad b' = b + \frac{1}{2}\sum_{n=1}^{N}\sum_{d=1}^{D} \boldsymbol{x}_{nd}^{\top}\boldsymbol{x}_{nd} - \frac{1}{2}\sum_{n=1}^{N}\sum_{j=1}^{J_n} \boldsymbol{\mu}_{nj}^{\top}\boldsymbol{C}_{nj}^{-1}\boldsymbol{\mu}_{nj}, \tag{5}$$

$$\boldsymbol{\mu}_{nj} = \boldsymbol{C}_{nj} \sum_{d:s_{nd}=j} \boldsymbol{W}_d^\top \boldsymbol{x}_{nd}, \quad \boldsymbol{C}_{nj}^{-1} = \sum_{d:s_{nd}=j} \boldsymbol{W}_d^\top \boldsymbol{W}_d + r\boldsymbol{I}. \tag{6}$$

The posterior for the precision parameter α and that for the latent vector $\boldsymbol{z}_{nj}$ are given by

$$p(\alpha|\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{W}, a, b) = \mathrm{Gamma}(a', b'), \quad p(\boldsymbol{z}_{nj}|\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{W}, r) = \mathcal{N}(\boldsymbol{\mu}_{nj}, \alpha^{-1}\boldsymbol{C}_{nj}), \tag{7}$$

respectively.

3 Inference

We describe inference procedures for the proposed model based on a stochastic EM algorithm, in which collapsed Gibbs sampling of latent vector assignments $\boldsymbol{S}$ and the maximum joint likelihood estimation of projection matrices $\boldsymbol{W}$ are alternately iterated while analytically integrating out the latent vectors $\boldsymbol{Z}$, mixture weights $\boldsymbol{\Theta}$ and precision parameter α. By integrating out latent vectors, we do not need to explicitly infer the latent vectors, leading to a robust and fast-mixing inference.

Let $\ell = (n, d)$ be the index of the dth view of the nth instance for notational convenience. In the E-step, given the current state of all but one latent assignment s_ℓ, a new value for s_ℓ is sampled from $\{1, \cdots, J_{n\backslash\ell} + 1\}$ according to the following probability,

$$p(s_\ell = j|\boldsymbol{X}, \boldsymbol{S}_{\backslash\ell}, \boldsymbol{W}, a, b, r, \gamma) \propto \frac{p(s_\ell = j, \boldsymbol{S}_{\backslash\ell}|\gamma)}{p(\boldsymbol{S}_{\backslash\ell}|\gamma)} \cdot \frac{p(\boldsymbol{X}|s_\ell = j, \boldsymbol{S}_{\backslash\ell}, \boldsymbol{W}, a, b, r)}{p(\boldsymbol{X}_{\backslash\ell}|\boldsymbol{S}_{\backslash\ell}, \boldsymbol{W}, a, b, r)}, \tag{8}$$

where $\backslash\ell$ represents a value or set excluding the dth view of the nth instance. The first factor is given by

$$\frac{p(s_\ell = j, \boldsymbol{S}_{\backslash\ell}|\gamma)}{p(\boldsymbol{S}_{\backslash\ell}|\gamma)} = \begin{cases} \frac{N_{nj\backslash\ell}}{D-1+\gamma} & \text{if } j \leq J_{n\backslash\ell} \\ \frac{\gamma}{D-1+\gamma} & \text{if } j = J_{n\backslash\ell} + 1, \end{cases} \tag{9}$$

using (3), where $j \leq J_{n\backslash\ell}$ is for existing latent vectors, and $j = J_{n\backslash\ell} + 1$ is for a new latent vector. By using (4), the second factor is given by

$$\frac{p(\boldsymbol{X}|s_\ell = j, \boldsymbol{S}_{\backslash\ell}, \boldsymbol{W}, a, b, r)}{p(\boldsymbol{X}_{\backslash\ell}|\boldsymbol{S}_{\backslash\ell}, \boldsymbol{W}, a, b, r)} = (2\pi)^{-\frac{M_d}{2}} r^{I(j=J_{n\backslash\ell}+1)\frac{K}{2}} \frac{b_{\backslash\ell}'^{a_{\backslash\ell}'}}{b_{s_\ell=j}'^{a_{s_\ell=j}'}} \frac{\Gamma(a_{s_\ell=j}')}{\Gamma(a_{\backslash\ell}')} \frac{|\boldsymbol{C}_{j,s_\ell=j}|^{\frac{1}{2}}}{|\boldsymbol{C}_{j\backslash\ell}|^{\frac{1}{2}}}, \tag{10}$$

where $I(\cdot)$ represents the indicator function, i.e. $I(A) = 1$ if A is true and 0 otherwise, and subscript $s_\ell = j$ indicates the value when $\boldsymbol{x}_\ell$ is assigned to the jth latent vector as follows,

$$b_{s_\ell=j}' = b_{\backslash\ell}' + \frac{1}{2}\boldsymbol{x}_\ell^\top \boldsymbol{x}_\ell + \frac{1}{2}\boldsymbol{\mu}_{nj\backslash\ell}^\top \boldsymbol{C}_{nj\backslash\ell}^{-1} \boldsymbol{\mu}_{nj\backslash\ell} - \frac{1}{2}\boldsymbol{\mu}_{nj,s_\ell=j}^\top \boldsymbol{C}_{nj,s_\ell=j}^{-1} \boldsymbol{\mu}_{nj,s_\ell=j}, \tag{11}$$

$$a_{s_\ell=j}' = a', \quad \boldsymbol{\mu}_{nj,s_\ell=j} = \boldsymbol{C}_{nj,s_\ell=j}(\boldsymbol{W}_d^\top \boldsymbol{x}_\ell + \boldsymbol{C}_{nj\backslash\ell}^{-1}\boldsymbol{\mu}_{nj\backslash\ell}), \tag{12}$$

$$\boldsymbol{C}_{nj,s_\ell=j}^{-1} = \boldsymbol{W}_d^\top \boldsymbol{W}_d + \boldsymbol{C}_{nj\backslash\ell}^{-1}. \tag{13}$$

Intuitively, if the current view cannot be modeled well by existing latent vectors, a new latent vector is used, which indicates that the view is inconsistent with the other views.

In the M-step, the projection matrices $\boldsymbol{W}$ are estimated by maximizing the logarithm of the joint likelihood (2) while fixing cluster assignment variables $\boldsymbol{S}$. By setting the gradient of the joint log likelihood with respect to $\boldsymbol{W}$ equal to zero, an estimate of $\boldsymbol{W}$ is obtained as follows,

$$\boldsymbol{W}_d = \Big(\frac{a'}{b'}\sum_{n=1}^N \boldsymbol{x}_{nd}\boldsymbol{\mu}_{ns_{nd}}^\top\Big)\Big(\sum_{n=1}^N\sum_{j=1}^{J_n}\boldsymbol{C}_{nj} + \frac{a'}{b'}\sum_{n=1}^N \boldsymbol{\mu}_{ns_{nd}}\boldsymbol{\mu}_{ns_{nd}}^\top\Big)^{-1}. \tag{14}$$

When we iterate the E-step that samples the latent vector assignment s_{nd} by employing (8) for each view $d = 1, \ldots, D$ in each instance $n = 1, \ldots, N$, and the M-step that maximizes the joint likelihood using (14) with respect to the projection matrix $\boldsymbol{W}_d$ for each view $d = 1, \ldots, D$, we obtain an estimate of the latent vector assignments and projection matrices.

4

In Section 2, we defined that an instance is an anomaly when its different views are generated from different latent vectors. Therefore, for an anomaly score, we use the probability that the instance uses more than one latent vector. It is estimated by using the samples obtained in the inference as follows, $v_n = \frac{1}{H} \sum_{h=1}^{H} I(J_n^{(h)} > 1)$, where $J_n^{(h)}$ is the number of latent vectors used by the nth instance in the hth iteration of the Gibbs sampling after the burn-in period, and H is the number of the iterations. The output of the proposed method is a ranked list of anomalies based on their anomaly scores. An analyst would investigate top few anomalies, or use a threshold to select the anomalies [6]. The threshold can be determined based on a targeted false alarm and detection rate.

We can use cross-validation to select an appropriate dimensionality for the latent space K. With cross-validation, we assume that some features are missing from the given data, and infer the model with a different K. Then, we select the smallest K value that has performed the best at predicting missing values.

4 Related Work

Anomaly detection has had a wide variety of applications, such as credit card fraud detection [1], intrusion detection for network security [17], and analysis for healthcare data [3]. However, most existing anomaly detection techniques assume data with a single view, i.e. a single observation feature set.

A number of anomaly detection methods for two-view data have been proposed [12, 20–22, 24]. However, they cannot be used for data with more than two views. Gao et al. [13] proposed a HOrizontal Anomaly Detection algorithm (HOAD) for finding anomalies from multi-view data. In HOAD, there are hyperparameters including a weight for the constraint that require the data to be labeled as anomalous or not for tuning, and the performance is sensitive to the hyperparameters. On the other hand, the parameters with the proposed model can be estimated from the given multi-view data without label information by maximizing the likelihood. In addition, because the proposed model is a probabilistic generative model, we can extend it in a probabilistically principled manner, for example, for handling missing data and combining with other probabilistic models.

Liu and Lam [16] proposed multi-view anomaly detection methods using consensus clustering. They found anomalies based on the inconsistency of clustering results across multiple views. Therefore, they cannot find inconsistency within a cluster. Christoudias et al. [8] proposed a method for filtering instances that are corrupted by background noise from multi-view data. The multi-view anomalies considered in this paper include not only instances corrupted by background noise but also instances categorized into different foreground classes across views, and instances with inconsistent views even if they belong to the same cluster. Recently, Alvarez et al. [2] proposed a multi-view anomaly detection method. However, since the method is based on clustering, it cannot find anomalies when there are no clusters in the given data.

The proposed model is a generalization of either probabilistic principal component analysis (PPCA) [23] or probabilistic canonical correlation analysis (PCCA) [5]. When all views are generated from different latent vectors for every instance, the proposed model corresponds to PPCA that is performed independently for each view. When all views are generated from a single latent vector for every instance, the proposed model corresponds to PCCA with spherical noise.

PCCA, or canonical correlation analysis (CCA), can be used for multi-view anomaly detection. With PCCA, a latent vector that is shared by all views for each instance and a linear projection matrix for each view are estimated by maximizing the likelihood, or minimizing the reconstruction error of the given data. The reconstruction error for each instance can be used as an anomaly score. However, the reconstruction errors are not reliable because they are calculated from parameters that are estimated using data with anomalies by assuming that all of the instances are non-anomalous. On the other hand, because the proposed model simultaneously estimates the parameters and infers anomalies, the estimated parameters are not contaminated by the anomalies. With PPCA and PCCA, Gaussian distributions are used for observation noise, which are sensitive to atypical observations. Robust PPCA and PCCA [4] use Student-t distributions instead of Gaussian distributions, which are stable to data containing single-view anomalies. The proposed model assumes Gaussian observation noise, and its precision is parameterized by a Gamma distributed variable α. Since we marginalize out α in the inference as written in (4), the observation noise becomes a Student-t distribution. Therefore, the proposed model is robust to single-view anomalies.

With some CCA-related methods, each latent vector is factorized into shared and private components across different views [10]. They assume that every instance has shared and private parts that are the same dimensionality for all instances. In contrast, the proposed model assumes that non-anomalous instances have only shared latent vectors, and anomalies have private latent vectors. The proposed model can be seen as CCA with private latent vectors, where latent vectors across views are clustered for each instance. When CCA with private latent vectors are inferred without clustering, the inferred private latent vectors do not become the same even if it is generated from a single latent vector, because switching latent dimension or rotating the latent space does not change the likelihood. Therefore, difference of the latent vectors cannot be used for multi-view anomaly detection.

5 Experiments

Data   We evaluated the proposed model quantitatively by using 11 data sets, which we obtained from the LIBSVM data sets [7]. We generated two views by randomly splitting the features, where each feature can belong to only a single view, and anomalies were added by swapping views of two randomly selected instances regardless of their class labels for each view. Splitting data does not generate anomalies. Therefore, we can evaluate methods while controlling the anomaly rate properly. By swapping, although single-view anomalies cannot be created since the distribution for each view does not change, multi-view anomalies are created.

Comparing methods   We compared the proposed model with probabilistic canonical correlation analysis (PCCA), horizontal anomaly detection (HOAD) [13], consensus clustering based anomaly detection (CC) [16], and one-class support vector machine (OCSVM) [18]. For PCCA, we used the proposed model in which the number of latent vectors was fixed at one for every instance. The anomaly scores obtained with PCCA were calculated based on the reconstruction errors. HOAD requires to select an appropriate hyperparameter value for controlling the constraints whereby different views of the same instance are embedded close together. We ran HOAD with different hyperparameter settings $\{0.1, 1, 10, 100\}$, and show the results that achieved the highest performance for each data set. For CC, first we clustered instances for each view using spectral clustering. We set the number of clusters at 20, which achieved a good performance in preliminary experiments. Then, we calculated anomaly scores by the likelihood of consensus clustering when an instance was removed since it indicates inconsistency of the instance across different views. OCSVM is a representative method for single-view anomaly detection. To investigate the performance of a single-view method for multi-view anomaly detection, we included OCSVM as a comparison method. For OCSVM, multiple views are concatenated in a single vector, then use it for the input. We used Gaussian kernel. In the proposed model, we used $\gamma = 1$, $a = 1$, and $b = 1$ for all experiments. The number of iterations for the Gibbs sampling was 500, and the anomaly score was calculated by averaging over the multiple samples.

Multi-view anomaly detection   For the evaluation measurement, we used the area under the ROC curve (AUC). A higher AUC indicates a higher anomaly detection performance. Figure 2 shows AUCs with different rates of anomalies using 11 two-view data sets, which are averaged over 50 experiments. For the dimensionality of the latent space, we used $K = 5$ for the proposed model, PCCA, and HOAD. In general, as the anomaly rate increases, the performance decreases. The proposed model achieved the best performance with eight of the 11 data sets. This result indicates that the proposed model can find anomalies effectively by inferring a number of latent vectors for each instance. The performance of CC was low because it assumes that there are clusters for each view, and it cannot find anomalies within clusters. The AUC of OCSVM was low, because it is a single-view anomaly detection method, which considers instances anomalous that are different from others within a single view. Multi-view anomaly detection is the task to find instances that have inconsistent features across views, but not inconsistent features within a view. The computational time needed for PCCA was 2 sec, and that needed for the proposed model was 35 sec with wine data.

Figure 3 shows AUCs with different dimensionalities of latent vectors using data sets whose anomaly rate is 0.4. When the dimensionality was very low ($K = 1$ or 2), the AUC was low in most of the data sets, because low-dimensional latent vectors cannot represent the observation vectors well. With all the methods, the AUCs were relatively stable when the latent dimensionality was higher than four.
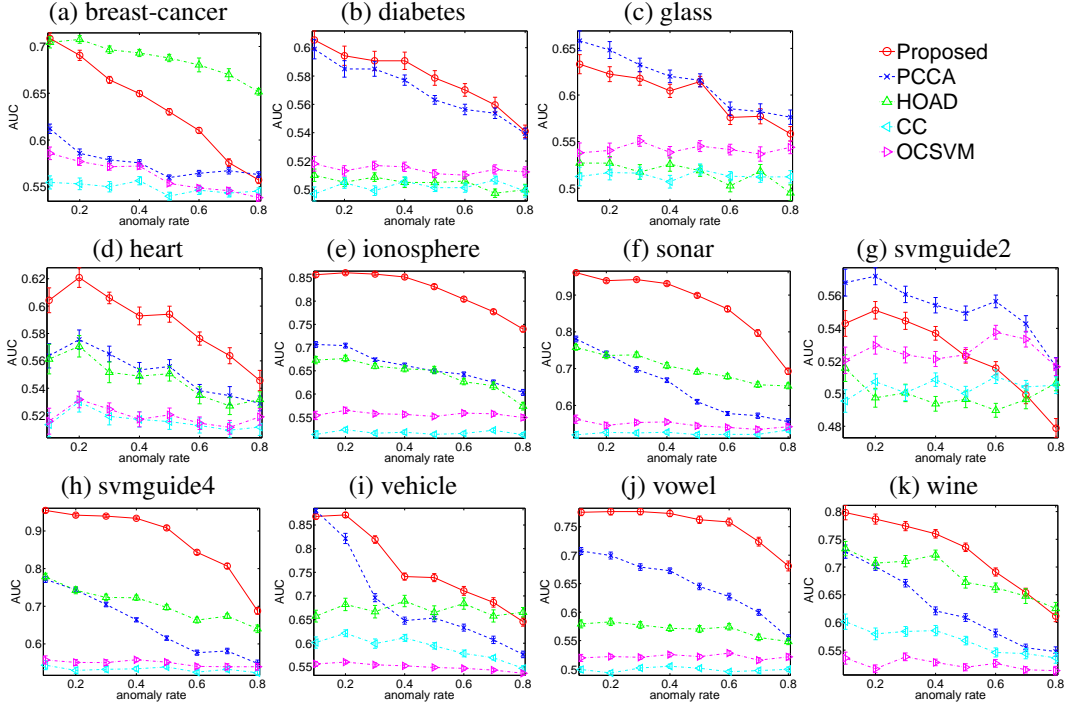
Figure 2: Average AUCs with different anomaly rates, and their standard errors. A higher AUC is better.
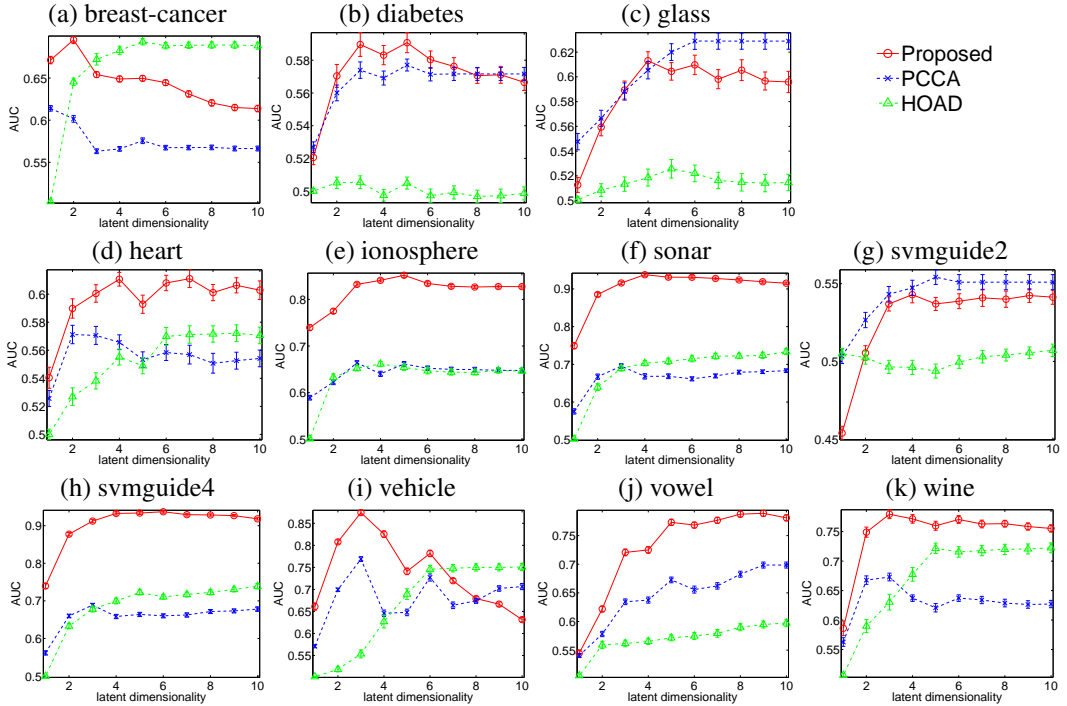

Figure 3: Average AUCs with different dimensionalities of latent vectors, and their standard errors.

Single-view anomaly detection We would like to find multi-view anomalies, but woul not like to detect single-view anomalies. We illustrated that the proposed model does not detect single-view anomalies using synthetic single-view anomaly data. With the synthetic data, latent vectors for

7

Table 1: Average AUCs for single-view anomaly detection.

Proposed	PCCA	OCSVM
0.117 ± 0.098	0.174 ± 0.095	0.860 ± 0.232

Table 2: High and low anomaly score movies calculated by the proposed model.

Title	Score	Title	Score
The Full Monty	0.98	Star Trek VI	0.04
Liar Liar	0.93	Star Trek III	0.04
The Professional	0.91	The Saint	0.04
Mr. Holland's Opus	0.88	Heat	0.03
Contact	0.87	Conspiracy Theory	0.03

single-view anomalies were generated from $\mathcal{N}(\mathbf{0}, \sqrt{10}\mathbf{I})$, and those for non-anomalous instances were generated from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Since each of the anomalies has only one single latent vector, it is not a multi-view anomaly. The numbers of anomalous and non-anomalous instances were 5 and 95, respectively. The dimensionalities of the observed and latent spaces were five and three, respectively. Table 1 shows the average AUCs with the single-view anomaly data, which are averaged over 50 different data sets. The low AUC of the proposed model indicates that it does not consider single-view anomalies as anomalies. On the other hand, the AUC of the one-class SVM (OCSVM) was high because OCSVM is a single-view anomaly detection method, and it leads to low multi-view anomaly detection performance.

Application to movie data   For an application of multi-view anomaly detection, we analyzed inconsistency between movie rating behavior and genre in MovieLens data [14]. An instance corresponds to a movie, where the first view represents whether the movie is rated or not by users, and the second view represents the movie genre. Both views consist of binary features, where some movies are categorized in multiple genres. We used 338 movies, 943 users and 19 genres. Table 2 shows high and low anomaly score movies when we analyzed the movie data by the proposed method with $K = 5$. 'The Full Monty' and 'Liar Liar' were categorized in 'Comedy' genre. They are rated by not only users who likes 'Comedy', but also who likes 'Romance' and 'Action-Thriller'. 'The Professional' was anomaly because it was rated by two different user groups, where a group prefers 'Romance' and the other prefers 'Action'. Since 'Star Trek' series are typical Sci-Fi and liked by specific users, its anomaly score was low.

6   Conclusion

We proposed a generative model approach for multi-view anomaly detection, which finds instances that have inconsistent views. In the experiments, we confirmed that the proposed model could perform much better than existing methods for detecting multi-view anomalies. There are several avenues that can be pursued for future work. Since the proposed model assumes the linearity of observations with respect to their latent vectors, it cannot find anomalies when different views are in a nonlinear relationship. We can relax this assumption by using Gaussian processes [15]. We can also relax the assumption that non-anomalous instances have the same latent vector across all views by introducing private latent vectors [10]. The proposed model assumes Gaussian observation noise. Our framework can be extended for binary or count data by using Bernoulli or Poisson distributions instead of Gaussian.

References

[1] E. Aleskerov, B. Freisleben, and B. Rao. Cardwatch: A neural network based database mining system for credit card fraud detection. In *Proceedings of the IEEE/IAFE Computational Intelligence for Financial Engineering*, pages 220–226, 1997.

[2] A. M. Alvarez, M. Yamada, A. Kimura, and T. Iwata. Clustering-based anomaly detection in multi-view data. In *Proceedings of ACM International Conference on Information and Knowledge Management*, CIKM, 2013.

[3] M.-L. Antonie, O. R. Zaiane, and A. Coman. Application of data mining techniques for medical image classification. *MDM/KDD*, pages 94–101, 2001.

[4] C. Archambeau, N. Delannay, and M. Verleysen. Robust probabilistic projections. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 33–40, 2006.

[5] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[6] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[7] C. Chang and C. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[8] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. In *Proceedings of the 24th Conference on Unvertainty in Artificial Intelligence*, UAI, 2008.

[9] K. Duh, C.-M. A. Yeung, T. Iwata, and M. Nagata. Managing information disparity in multilingual document collections. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(1):1, 2013.

[10] C. H. Ek, J. Rihan, P. H. Torr, G. Rogez, and N. D. Lawrence. Ambiguity modeling in latent spaces. In *Machine Learning for Multimodal Interaction*, pages 62–73. Springer, 2008.

[11] H. Fanaee-T and J. a. Gama. Tensor-based anomaly detection. *Know.-Based Syst.*, 98(C):130–147, 2016.

[12] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han. On community outliers and their efficient detection in information networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 813–822. ACM, 2010.

[13] J. Gao, W. Fan, D. Turaga, S. Parthasarathy, and J. Han. A spectral framework for detecting inconsistency across multi-source object relationships. In *IEEE 11th International Conference on Data Mining (ICDM)*, pages 1050–1055. IEEE, 2011.

[14] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 230–237. ACM, 1999.

[15] N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, 16(3):329–336, 2004.

[16] A. Y. Liu and D. N. Lam. Using consensus clustering for multi-view anomaly detection. In *2012 IEEE Symposium on Security and Privacy Workshops (SPW)*, pages 117–124. IEEE, 2012.

[17] L. Portnoy, E. Eskin, and S. Stolfo. Intrusion detection with unlabeled data using clustering. In *Proceedings of ACM CSS Workshop on Data Mining Applied to Security*, 2001.

[18] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[19] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[20] S. Shekhar, C.-T. Lu, and P. Zhang. Detecting graph-based spatial outliers. *Intelligent Data Analysis*, 6 (5):451–468, 2002.

[21] X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):631–645, 2007.

[22] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Proceedings of the 5th IEEE International Conference on Data Mining*, pages 418–425. IEEE, 2005.

[23] M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[24] X. Wang and I. Davidson. Discovering contexts and contextual outliers using random walks in graphs. In *Proceedings of the 9th IEEE International Conference on Data Mining*, pages 1034–1039. IEEE, 2009.