Towards a Hierarchical Bayesian Model of Multi-View Anomaly Detection

Zhen Wang and **Chao Lan**
Department of Computer Science, University of Wyoming, WY, USA
{zwang10, clan}@uwyo.edu

Abstract

Traditional anomaly detectors examine a single view of instances and cannot discover multi-view anomalies, i.e., instances that exhibit inconsistent behaviors across different views. To tackle the problem, several multi-view anomaly detectors have been developed recently, but they are all transductive and unsupervised thus may suffer some challenges. In this paper, we propose a novel inductive semi-supervised Bayesian multi-view anomaly detector. Specifically, we first present a generative model for normal data. Then, we build a hierarchical Bayesian model, by first assigning priors to all parameters and latent variables, and then assigning priors over the priors. Finally, we employ variational inference to approximate the posterior of the model and evaluate anomalous scores of multi-view instances. In the experiment, we show the proposed Bayesian detector consistently outperforms state-of-the-art counterparts across several public data sets and three well-known types of multi-view anomalies. In theory, we prove the inferred Bayesian estimator is consistent and derive a proximate sample complexity for the proposed anomaly detector.

1 Introduction

Anomaly Detection (AD) is a fundamental task with broad applications, such as in clinical diagnosis, fraud transaction detection and cybersecurity [Chandola *et al.*, 2009]. Traditional detectors only examine a single view[1] of instances and cannot discover multi-view anomalies, i.e., instances that exhibit inconsistent behaviors across different views. One example is in web image analysis, an image can be described its category such as car or animal (view 1) and its web page such as cars.com or animals.com (view 2). If an image is assigned to the animal group in view 1 but car group in view 2, then it is natural to consider this image anomalous [Marcos Alvarez *et al.*, 2013]. Other examples can be found in digit recognition [Li *et al.*, 2018b] and movie recommendation on MovieLens dataset [Gao *et al.*, 2011]. How to effectively leverage multiple views to detect anomaly is an interesting and significant topic, often referred to as multi-view anomaly detection.

In the literature, a number of multi-view anomaly detectors have been developed. Some of them try to find samples that have inconsistent cross-view cluster membership. HOrizontal Anomaly Detection (HOAD) [Gao *et al.*, 2011] pioneers this branch of methods. In HOAD, the author first constructs a combined similarity graph based on the similarity matrices, and computes the key eigenvectors of the graph Laplacian of the combined matrix. Then anomalies are identified by computing cosine distance between the components of these eigenvectors. This idea is further studied by [Marcos Alvarez *et al.*, 2013] and [Liu and Lam, 2012] for different application tasks. Another successful group of methods is developed from a perspective of data representation [Li *et al.*, 2015; Zhao and Fu, 2015; Li *et al.*, 2018a]. The intuition in these works is that a normal sample usually serves as a good contributor in representing the other normal samples while the outliers do not. Low-rank matrix recovery is the technique which can exploit the intrinsic structure of data and explore the representation relationship of samples. Therefore, by calculating the representation coefficients in low-rank matrix recovery, the multi-view outliers can be identified. In addition, [Iwata and Yamada, 2016] utilizes a sophisticated statistical machine learning algorithm to detect anomalies. They design a probabilistic latent variable model to infer the consistent or inconsistent characteristics of multiple views for each object.

We note that above detection methods are unsupervised and transductive. In some applications, however, one can often get plenty of labeled normal data (e.g., one can collect or simulate normal network traffic data for a certain period of time). In these cases, it is natural to hypothesize these data will enable the detector to better capture normal behavior than unlabeled data. In addition, when an unseen testing instance arrives, above methods have to add it to the existing training set and rerun the detection algorithm (e.g., clustering or matrix factorization), which often causes inefficiency.

To lift above limitations, in this paper we propose a novel Bayesian model for semi-supervised multi-view anomaly detection. To be specific, we first present a generative model for normal data, assuming that different views of a normal instance are generated from a single latent factor through

[1] A view is a set of features that often have similar semantics, e.g., a webpage can be described by one view of its content and another view of its hyperlinks [Xu *et al.*, 2013].

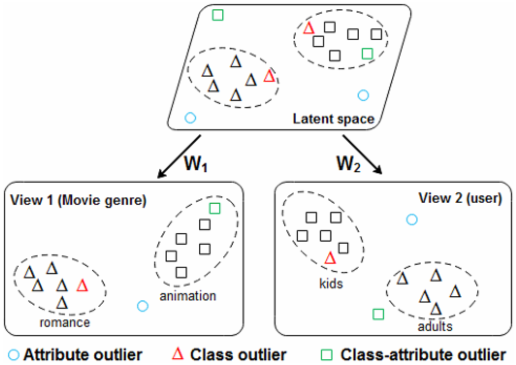


Figure 1: Illustration of three types of outliers in multi-view setting.

different projection matrices; and the views are independent conditioned on the factor [Blum and Mitchell, 1998; Dasgupta *et al.*, 2002]. Then we build a hierarchical Bayesian model, by first assigning priors on model parameters and then assigning priors over the priors. In particular, we assign the *automatic relevance determination* (ARD) prior [Neal, 2012] on the projection matrices to sparsify their columns for automatically determining the dimension of latent factor; we also place Student's t distributions on the latent factor prior and the likelihood to improve robustness of the estimator [Archambeau *et al.*, 2006; Gai *et al.*, 2008]. Finally, we employ *variational inference* to derive an analytical approximation to the posterior probability (of unobserved variables and parameters) of model. To detect multi-view anomalies, we propose to measure the outlier score by calculating the value of log marginal distribution of multiple observed views.

The contributions of this paper are summarized as: 1) To the best of our knowledge, this paper is the first attempt to detect multi-view outliers under semi-supervised scenario via a Bayesian model of inductive learning. 2) In theory, we proves the proposed estimator approaches the true model at a rate of $\mathcal{O}\big(\frac{\sum_v d^v m_0 \log(\sum_v d^v N)}{N}\big)$, and under mild conditions we derive a first sample complexity of $\mathcal{O}\big(\frac{1}{\epsilon^2\gamma^2}\big(V^2 \ln V + \ln\frac{1}{\delta}\big)\big)$ for a multi-view anomaly detector to achieve detection rate ϵ. 3) we experimentally evaluates the proposed method on both synthetic and real-life multi-view data. The competing results demonstrate the effectiveness of our model.

The rest of this paper is organized as follows. In Section II, we introduce basic notations; in Section III, we present the proposed Bayesian multi-view anomaly detection model; in Section IV, we theoretically analyze the model; in Section V, we show experimental results and discussions; in Section VI, we conclude the study.

2 Preliminaries and Problem Setup

Before going further, we explain some preliminary knowledge and notational conventions used throughout the paper.

To clarify, in anomaly detection applications, the term semi-supervised detection has been widely used to describe the scenario in which AD methods only incorporate the use of labeled normal samples to learn a model that compactly characterizes the "normal" class [Chalapathy and Chawla, 2019; Akcay *et al.*, 2018; Chandola *et al.*, 2009; Blanchard *et al.*, 2010; Muñoz-Marí *et al.*, 2010; Song *et al.*, 2017]. However, there are a few works [Das *et al.*, 2016; Siddiqui *et al.*, 2018; Görnitz *et al.*, 2013] having investigated the general semi-supervised setting where one also utilizes unlabelled data. In this work, we stick to the first AD setting.

Following the definition used in [Li *et al.*, 2018a], there are three kinds of outliers in multi-view setting. As shown in Figure 1, *Class-outlier* is an outlier that exhibits inconsistent characteristics (e.g. cluster membership) across different views. *Attribute-outlier* is an outlier that exhibits consistent abnormal behaviours in each view. *Class-Attribute-outlier* is an outlier that exhibits class outlier characteristics in some views while shows attribute outlier properties in the other views. Suppose we are given a data set $\mathcal{D}$ which consists of N instances, denoted by $n = 1, 2, ..., N$, described by V views with each view $v = 1, 2, ..., V$. The feature representation of instance n under view v is $\mathbf{x}_n^v \in \mathbb{R}^{d^v}$, where d^v is the dimensionality of view v. $X^v = [\mathbf{x}_1^v, \mathbf{x}_2^v, ..., \mathbf{x}_N^v] \in \mathbb{R}^{d^v \times N}$ is sample set observed in view v. In this way, the whole data set is denoted as $\mathcal{D} = \{\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^V\}$. Then, the multi-view anomaly detector computes an anomaly score for each instance and compares it to a threshold $\hat{\tau}_\zeta$ for finding the outlier in multi-view setting.

3 Bayesian Muli-View Anomaly Detector

In this section, we illustrate the proposed probabilistic model together with its estimation and the outlier score calculation.

3.1 Generative Process

To link the multiple views $\mathbf{x}^1, \mathbf{x}^2, ...,$ and $\mathbf{x}^V$ together, we introduce a common latent variable $\mathbf{z}$. The intuition here is that: generally, an normal instance can be sufficiently described by a single view for learning tasks. Therefore, it is reasonable to suppose that these different views share some common features or latent structure, then the problem is how to build a framework to learn these common structure or the correspondence between observed views and the unobserved space. To explore it, we proposed a probabilistic model which describes the generative process of multi-view instances whose views are linked via a single, reduced-dimensionality latent variable space. Specifically, We assume $\mathbf{x}^1, \mathbf{x}^2, ...,$ and $\mathbf{x}^V$ are generated from $\mathbf{z}$ by first choosing a value for the latent variable $\mathbf{z}$ and then sampling observed variables conditioned on this latent value. The $d^1, d^2, ..., d^V$-dimensional observed vectors $\mathbf{x}_n^1, \mathbf{x}_n^2, ..., \mathbf{x}_n^V$ are defined by a linear transformation governed by the matrix $\mathbf{W}_v \in \mathbb{R}^{d^v \times m}$ of the latent vector $\mathbf{z} \in \mathbb{R}^m$ plus a projection noise $\boldsymbol{\epsilon}_v \in \mathbb{R}^{d^v}$, so that

$$\mathbf{x}_n^v = \mathbf{W}_v \mathbf{z}_n + \boldsymbol{\mu}_v + \boldsymbol{\epsilon}_v \quad v = 1, 2, ..., V \tag{1}$$

where $\boldsymbol{\mu}_v \in \mathbb{R}^{d^v}$ is the data offset.

3.2 Model Specification

In the following, we introduce the probabilistic formulation, assign prior distribution on latent variables and parameters in Eq. (1), and assign priors on the priors. Specifically, we define Student's-t prior distribution over the latent variable $\mathbf{z}$

$$\mathbf{z} \sim \mathcal{S}(\mathbf{z} \mid \mathbf{0}, \mathbf{I}_m, \nu) = \int_0^{+} p(\mathbf{z}|u)p(u)du \tag{2}$$

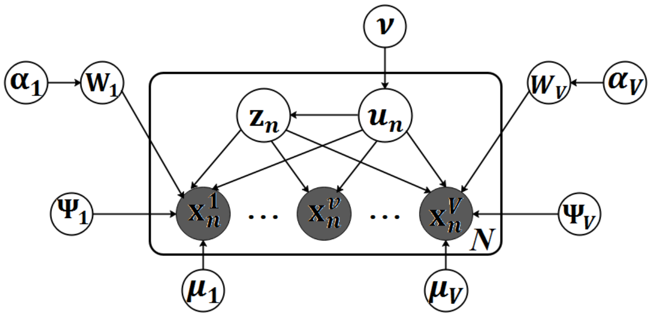


Figure 2: The proposed hierarchical Bayesian model for a data set of N observations (For the concision of graph, here we omit the dependence of $\mathbf{x}_n^v$ on $\mathbf{W}_v, \boldsymbol{\Psi}_v, \boldsymbol{\mu}_v$).

In Eq. (2), we adopt Student's-t distribution's equivalent form according to [Liu and Rubin, 1995; Archambeau *et al.*, 2006]. $u > 0$ is a latent scale variable. Its Gamma prior $p(u)$ and the associated Gaussian condition $p(\mathbf{z}|u)$ are defined as:

$$u \sim \mathcal{G}\big(u \mid \frac{\nu}{2}, \frac{\nu}{2}\big), \qquad \mathbf{z} \mid u \sim \mathcal{N}\left(\mathbf{z} \mid \mathbf{0}, u^{\text{-}1}\mathbf{I}_m\right) \tag{3}$$

Similarly, noise $\boldsymbol{\epsilon}_v$ is a d^v dimensional zero-mean Student's-t variable with precision $\boldsymbol{\Psi}_v$ and degree of freedom ν

$$\boldsymbol{\epsilon}_v \sim \mathcal{S}(\boldsymbol{\epsilon}_v \mid \mathbf{0}, \boldsymbol{\Psi}_v, \nu) \tag{4}$$

By the property of affine transformation of random variable, combining (1), (2), (3) and (4) gives the conditional distributions of observed variables $\mathbf{x}^v$

$$\mathbf{x}^v \mid \mathbf{z} \sim \mathcal{S}(\mathbf{x}^v \mid \mathbf{W}_v\mathbf{z} + \boldsymbol{\mu}_v, \boldsymbol{\Psi}_v, \nu) \tag{5}$$

$$\mathbf{x}^v \mid \mathbf{z}, u \sim \mathcal{N}\big(\mathbf{x}^v \mid \mathbf{W}_v\mathbf{z} + \boldsymbol{\mu}_v, (u\boldsymbol{\Psi}_v)^{\text{-}1}\big) \tag{6}$$

Next, we place priors on parameters $\mathbf{W}_v$, $\boldsymbol{\mu}_v$ and $\boldsymbol{\Psi}_v$. Let $\mathbf{w}_{vi} \in \mathbb{R}^m$ be the i_{th} row of $\mathbf{W}_v$, we first employ ARD prior on $\mathbf{W}_v$ to automatically sparsify its columns

$$\mathbf{W}_v \mid \boldsymbol{\alpha}_v \sim \textstyle\prod_{i=1}^{d^v} \mathcal{N}\big(\mathbf{w}_{vi} \mid \mathbf{0}, \big(diag(\boldsymbol{\alpha}_v)\big)^{\text{-}1}\big) \tag{7}$$

Then we parameterize the distributions over $\boldsymbol{\mu}_v$ and $\boldsymbol{\Psi}_v$ by defining

$$\boldsymbol{\mu}_v \sim \mathcal{N}(\boldsymbol{\mu}_v \mid \mathbf{0}, \beta_v^{\text{-}1}\mathbf{I}_{d^v}), \qquad \boldsymbol{\Psi}_v \sim \mathcal{W}(\boldsymbol{\Psi}_v \mid \mathbf{K}_v^{\text{-}1}, \nu_v) \tag{8}$$

where $\mathcal{W}(\boldsymbol{\Psi}_v|\mathbf{K}_v^{\text{-}1}, \nu_v) \propto |\boldsymbol{\Psi}_v|^{\frac{(\nu_v - d^v - 1)}{2}} \exp\big(\text{-}\frac{1}{2}Tr(\mathbf{K}_v\boldsymbol{\Psi}_v)\big)$ denotes Wishart distribution. Finally, we complete the specification of priors (ν,$\boldsymbol{\alpha}_v$) over prior distributions of variables u and $\mathbf{w}_{vi}$ respectively

$$\nu \sim \mathcal{G}(\nu \mid a_\nu, b_\nu), \qquad \boldsymbol{\alpha}_v \sim \textstyle\prod_{j=1}^{m} \mathcal{G}(\alpha_{vj} \mid a_\alpha, b_\alpha) \tag{9}$$

where $\boldsymbol{\alpha}_v$ controls the magnitude of $\mathbf{W}_v$. If certain α_{vj} is large, the jth column of $\mathbf{W}_v$ will tend to take value zero and become little importance.

The graphical representation of Bayesian model over a data set of N instances is illustrated by Figure 2 in which arrows represent conditional dependencies between random variables. Since we have no further knowledge about the hyperparameters of priors, we choose broad ones by setting $a_\alpha = b_\alpha = \beta_v = 10^{\text{-}3}$, $\mathbf{K}_v = 10^{\text{-}3}\mathbf{I}_{d^v}$, $\nu_v = d^v + 1$, $a_\nu = 2$ and $b_\nu = 0.1$, $m = \min\{d^v - 1; v = 1, \dots, V\}$.

3.3 Model Inference

The goal of model inference is to learn posterior distributions of latent variables and parameters. Based on Figure 2, the joint probability of data set $\mathcal{D}$, latent components $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$, $U = \{u_1, \dots, u_N\}$, and parameters $\boldsymbol{\Theta} = \{\{\mathbf{W}_v, \boldsymbol{\alpha}_v, \boldsymbol{\mu}_v, \boldsymbol{\Psi}_v\}_{v=1}^V, \nu\}$ can be written as

$$\begin{aligned} &p(\mathbf{X}^1, \dots, \mathbf{X}^V, \mathbf{Z}, U, \boldsymbol{\Theta}) = p(\nu)\times \\ &\textstyle\prod_{v=1}^V p(\mathbf{W}_v|\boldsymbol{\alpha}_v)p(\boldsymbol{\alpha}_v)p(\boldsymbol{\mu}_v)p(\boldsymbol{\Psi}_v)\times \\ &\textstyle\prod_{n=1}^N\prod_{v=1}^V p(\mathbf{x}_n^v|\mathbf{z}_n,u_n,\mathbf{W}_v,\boldsymbol{\mu}_v,\boldsymbol{\Psi}_v)p(\mathbf{z}_n|u_n)p(u_n|\nu) \end{aligned} \tag{10}$$

It is analytically intractable to derive the posterior distribution $p(\mathbf{Z}, U, \boldsymbol{\Theta}|\mathcal{D})$ from Eq. (10) directly. Therefore, we adopt variational inference for approxiamting the posterior by a facterized distribution

$$\begin{aligned} &q(\mathbf{Z}, U, \boldsymbol{\Theta}) = \textstyle\prod_{n=1}^N q(\mathbf{z}_n)\prod_{n=1}^N q(u_n)\times \\ &\quad q(\nu)\textstyle\prod_{v=1}^V\big(q(\boldsymbol{\Psi}_v)q(\boldsymbol{\mu}_v)\prod_{j=1}^m q(\alpha_{vj})\prod_{i=1}^{d^v} q(\mathbf{w}_{vi})\big) \end{aligned} \tag{11}$$

The distribution q is found by maximizing the lower bound

$$\begin{aligned} \mathcal{L}_{q(\mathbf{Z},U,\boldsymbol{\Theta})} &= \log p(\mathcal{D}) - KL(q(\mathbf{Z},U,\boldsymbol{\Theta})||p(\mathbf{Z},U,\boldsymbol{\Theta}|\mathcal{D})) \\ &= \iiint q(\mathbf{Z},U,\boldsymbol{\Theta}) \log\frac{p(\mathcal{D},\mathbf{Z},U,\boldsymbol{\Theta})}{q(\mathbf{Z},U,\boldsymbol{\Theta})} d\mathbf{Z}dU d\boldsymbol{\Theta} \end{aligned} \tag{12}$$

Since $\log p(\mathcal{D})$ is constant, maximizing the low bound is equivalent to minimizing the KL divergence between $q(\mathbf{Z},U,\boldsymbol{\Theta})$ and $p(\mathbf{Z},U,\boldsymbol{\Theta}|\mathcal{D})$. By substituting factor distributions in Eq. (11) into (12) and dissecting out the dependence on one of the factors $q_l(\boldsymbol{\Omega}_l)$, we have following result

$$\log q_l(\boldsymbol{\Omega}_l) = \mathbb{E}_{q_k(\boldsymbol{\Omega}_k), k\neq l}[\log p(\mathcal{D}, \mathbf{Z}, U, \boldsymbol{\Theta})] + const \tag{13}$$

where $\boldsymbol{\Omega} = \{\{\mathbf{z}_n, u_n\}_{n=1}^N, \{\mathbf{W}_v, \{\alpha_{vj}\}_{j=1}^m, \boldsymbol{\mu}_v, \boldsymbol{\Psi}_v\}_{v=1}^V, \nu\}$ refers all latent components and parameters of model, $\mathbb{E}.[\cdot]$ represents an expectation *w.r.t.* distribution $q_k(\boldsymbol{\Omega}_k)$ for all $k \neq l$. Combining (13), (10) with the distributions defined in section 3.2, we obtain the following factor distributions

$$q(\nu) = \mathcal{G}(\nu|\hat{a}_\nu, \hat{b}_\nu), \qquad q(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_{\mathbf{z}_n}, \boldsymbol{\Sigma}_{\mathbf{z}_n}) \tag{14}$$

$$q(u_n) = \mathcal{G}(u_n|\alpha_{u_n}, \beta_{u_n}), \ \ q(\boldsymbol{\Psi}_v) = \mathcal{W}(\boldsymbol{\Psi}_v|\hat{\mathbf{K}}_v^{\text{-}1}, \hat{\nu}_v) \tag{15}$$

$$q(\boldsymbol{\mu}_v) = \mathcal{N}(\boldsymbol{\mu}_v|\boldsymbol{\mu}_{\boldsymbol{\mu}_v}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_v}), \ \ q(\alpha_{vj}) = \mathcal{G}(\alpha_{vj}|\hat{a}_\alpha, \hat{b}_\alpha) \tag{16}$$

$$q(\mathbf{w}_{vi}) = \mathcal{N}(\mathbf{w}_{vi}|\boldsymbol{\mu}_{\mathbf{w}_{vi}}, \boldsymbol{\Sigma}_{\mathbf{w}_{vi}}) \qquad i = 1, \dots, d^v \tag{17}$$

where $n = 1, \dots, N$, $v = 1, \dots, V$, $j = 1, \dots, m$,

$$\hat{a}_\nu = a_\nu + \frac{N}{2}, \ \hat{b}_\nu = b_\nu - \frac{1}{2}\Big(N + \textstyle\sum_n\big(\langle\log u_n\rangle - \langle u_n\rangle\big)\Big) \tag{18}$$

$$\boldsymbol{\mu}_{\mathbf{z}_n} = \boldsymbol{\Sigma}_{\mathbf{z}_n}\Big[\textstyle\sum_v\langle\mathbf{W}_v^T\rangle\langle u_n\rangle\langle\boldsymbol{\Psi}_v\rangle\big(\mathbf{x}_n^v - \langle\boldsymbol{\mu}_v\rangle\big)\Big]$$

$$\boldsymbol{\Sigma}_{\mathbf{z}_n} = \big(\textstyle\sum_v\langle\mathbf{W}_v^T u_n\boldsymbol{\Psi}_v\mathbf{W}_v\rangle + \langle u_n\rangle\mathbf{I}_m\big)^{\text{-}1} \tag{19}$$

$$\hat{a}_\alpha = a_\alpha + \frac{d^v}{2}, \quad \hat{b}_\alpha = b_\alpha + \frac{\langle\|\mathbf{W}_{v,:j}\|^2\rangle}{2} \tag{20}$$

$$\begin{aligned}\beta_{u_n} = \frac{\langle\nu\rangle + \langle\mathbf{z}_n^T\mathbf{z}_n\rangle}{2} + \frac{1}{2}\textstyle\sum_v \big(&\mathbf{x}_n^{v\,T}\langle\mathbf{\Psi}_v\rangle\mathbf{x}_n^v - 2\mathbf{x}_n^{v\,T}\langle\mathbf{\Psi}_v\rangle\langle\boldsymbol{\mu}_v\rangle \\ &-2\mathbf{x}_n^{v\,T}\langle\mathbf{\Psi}_v\rangle\langle\mathbf{W}_v\rangle\langle\mathbf{z}_n\rangle + 2\langle\mathbf{z}_n^T\rangle\langle\mathbf{W}_v^T\rangle\langle\mathbf{\Psi}_v\rangle\langle\boldsymbol{\mu}_v\rangle \\ &+ \langle\boldsymbol{\mu}_v^T\mathbf{\Psi}_v\boldsymbol{\mu}_v\rangle + Tr[\langle\mathbf{W}_v\mathbf{z}_n\mathbf{z}_n^T\mathbf{W}_v^T\rangle\langle\mathbf{\Psi}_v\rangle]\big)\end{aligned}$$

$$\alpha_{u_n} = \frac{1}{2}(\langle\nu\rangle + m) + \textstyle\sum_v \frac{d^v}{2} \tag{21}$$

$$\begin{aligned}\hat{\mathbf{K}}_v = \mathbf{K}_v + \textstyle\sum_n \Big(&\mathbf{x}_n^v\mathbf{x}_n^{v\,T} + \langle\boldsymbol{\mu}_v\boldsymbol{\mu}_v^T\rangle - \langle\boldsymbol{\mu}_v\rangle\mathbf{x}_n^{v\,T} - \mathbf{x}_n^v\langle\boldsymbol{\mu}_v^T\rangle \\ &- \langle\mathbf{W}_v\rangle\langle\mathbf{z}_n\rangle\mathbf{x}_n^{v\,T} - \mathbf{x}_n^v\langle\mathbf{z}_n^T\rangle\langle\mathbf{W}_v^T\rangle + \langle\mathbf{W}_v\rangle\langle\mathbf{z}_n\rangle\langle\boldsymbol{\mu}_v^T\rangle \\ &+ diag\big([Tr(\mathbf{\Sigma}_{\mathbf{w}_{v1}}\langle\mathbf{z}_n\mathbf{z}_n^T\rangle), \ldots, Tr(\mathbf{\Sigma}_{\mathbf{w}_{vd^v}}\langle\mathbf{z}_n\mathbf{z}_n^T\rangle)]\big) \\ &+ \langle\boldsymbol{\mu}_v\rangle\langle\mathbf{z}_n^T\rangle\langle\mathbf{W}_v^T\rangle + \langle\mathbf{W}_v\rangle\langle\mathbf{z}_n\mathbf{z}_n^T\rangle\langle\mathbf{W}_v^T\rangle\Big)\langle u_n\rangle\end{aligned}$$

$$\hat{\nu}_v = \nu_v + N \tag{22}$$

$$\boldsymbol{\mu}_{\boldsymbol{\mu}_v} = \mathbf{\Sigma}_{\boldsymbol{\mu}_v}\textstyle\sum_n\langle u_n\rangle\langle\mathbf{\Psi}_v\rangle\big(\mathbf{x}_n^v - \langle\mathbf{W}_v\rangle\langle\mathbf{z}_n\rangle\big)$$

$$\mathbf{\Sigma}_{\boldsymbol{\mu}_v} = \big(\textstyle\sum_n\langle u_n\rangle\langle\mathbf{\Psi}_v\rangle + \beta_v\mathbf{I}_{d^v}\big)^{\text{-}1} \tag{23}$$

$$\mathbf{\Sigma}_{\mathbf{w}_{vi}} = \big(diag(\langle\boldsymbol{\alpha}_v\rangle) + \textstyle\sum_n\langle\mathbf{z}_n\mathbf{z}_n^T\rangle\langle u_n\mathbf{\Psi}_v\rangle_{ii}\big)^{\text{-}1}$$

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{w}_{vi}} = \mathbf{\Sigma}_{\mathbf{w}_{vi}}\textstyle\sum_n\big(&\langle\mathbf{z}_n\rangle\langle u_n\mathbf{\Psi}_v\rangle_{,:i}{}^T(\mathbf{x}_n^v - \langle\boldsymbol{\mu}_v\rangle) \\ &- \langle\mathbf{z}_n\mathbf{z}_n^T\rangle\textstyle\sum_{i'=1,i'\neq i}^{d^v}\langle u_n\mathbf{\Psi}_v\rangle_{i'i}\langle\mathbf{w}_{vi'}\rangle\big)\end{aligned} \tag{24}$$

where $M_{,:i}$ is the i_{th} column of matrix M, $\langle\cdot\rangle$ denotes the expectation, and we use Stirling's approximation $\Gamma(x) \sim \sqrt{2\pi}\exp(-x)x^{x-\frac{1}{2}}$ for $\log\big(\Gamma(\nu/2)\big)$ when deriving the factor $q(\nu)$. The equations (14-24) represent a set of consistency conditions for the maximum of the lower bound subject to the factorization constraint. We can find an optimal solution by first initializing all $q_k(\mathbf{\Omega}_k)$ properly and then cycling through the factors and re-estimating each distribution in turn using the updated moments of other factors. We monitor the convergence of optimization by evaluating the lower bound.

3.4 Outlier Score Measurement

The distribution $p(\mathbf{x}^1, ..., \mathbf{x}^V)$ of V-view observed variable $(\mathbf{x}^1, ..., \mathbf{x}^V)$ is expressed, from the sum and product rules of probability, in the form

$$\begin{aligned}p(\mathbf{x}^1, ..., \mathbf{x}^V) &= \iint\prod_{v=1}^{V} p(\mathbf{x}^v|\mathbf{z}, u)p(\mathbf{z}|u)p(u)d\mathbf{z}du \\ &= \mathcal{S}(\mathbf{x}^1, ..., \mathbf{x}^V|\mathbf{M}^{(\boldsymbol{\mu})}, \mathbf{\Lambda}^{(\mathbf{W},\mathbf{\Psi})}, \nu)\end{aligned} \tag{25}$$

by integrating out $\mathbf{z}$ and u, it gives a Student's-t distribution for $p(\mathbf{x}^1, ..., \mathbf{x}^V)$, where

$$\mathbf{M}^{(\boldsymbol{\mu})} = [\boldsymbol{\mu}_1; \boldsymbol{\mu}_2; \ldots; \boldsymbol{\mu}_V] \tag{26}$$

and $\mathbf{\Lambda}^{(\mathbf{W},\mathbf{\Psi})}$ is a $\sum_v d^v$-by-$\sum_v d^v$ precision matrix

$$\Lambda_{vv'}^{(\mathbf{W},\mathbf{\Psi})} = \begin{cases}\mathbf{W}_v\mathbf{W}_{v'}^T + \mathbf{\Psi}_v^{\text{-}1} & \text{if } v = v' \\ \mathbf{W}_v\mathbf{W}_{v'}^T & \text{if } v \neq v'\end{cases} \tag{27}$$

From Eq. (12), we know that the log marginal likelihood can be approximated by the evidence lower bound (ELBO). Through maximizing the ELBO, we find an optimum estimate of the data distribution. According to theorem 1 in section 4.1, this estimated distribution is close to the real model distribution with theoretical guarantee. Since all parameters of estimated data distribution are learned on the normal sample, thus it reasonably concludes that normal instance will have bigger value in Eq. (25). By this insight, we formulate the outlier score $s(\mathbf{x}_i)$ of an instance $\mathbf{x}_i = [\mathbf{x}_i^1; \mathbf{x}_i^2; \ldots; \mathbf{x}_i^V]$ as the negative unscaled Student's-t density

$$s(\mathbf{x}_i) := -\Big[1 + \frac{(\mathbf{x}_i - \mathbf{M}^{(\boldsymbol{\mu})})^T\mathbf{\Lambda}^{(\mathbf{W},\mathbf{\Psi})}(\mathbf{x}_i - \mathbf{M}^{(\boldsymbol{\mu})})}{\nu}\Big]^{\text{-}\frac{\nu + \sum_v d^v}{2}} \tag{28}$$

Conceptually, a negative outlier score measures the probability that the sample is generated from the multi-view distribution defined by normal data, therefore bigger value in Eq. (28) means this sample is less likely to be the normal class.

4 Theoretical Analysis

In this section, we show theoretical analysis for the proposed model (Due to space limitation, detailed proofs are omitted).

4.1 Consistency of the Bayesian Estimator

Let $\mathbf{X}_V^N := \{\{\mathbf{x}_n^v\}_{v=1}^V\}_{n=1}^N$ be a sample of *i.i.d* multi-view random variables collected from distribution p_0. We consider statistical models $\mathcal{M}_m = \{p_{\boldsymbol{\theta}_m}|\boldsymbol{\theta}_m \in \mathbf{\Theta}_m\}$ with the countable collection $\big\{\mathcal{M}_m|1 \leq m \leq \min\{d^v; v = 1, \ldots, V\}\big\}$, where $\mathbf{\Theta}_m$ is the parameter set associated with m latent components model. Let $\mathcal{F}^+(\mathbf{\Theta}_m)$ be the set of all possible distributions over $\mathbf{\Theta}_m$. Now we assume that there exists a true model $\mathcal{M}_{m_0}$ that contains the true data distribution $p_{\boldsymbol{\theta}^*_{m_0}}$ (i.e., there exists m_0 and $\boldsymbol{\theta}^*_{m_0} \in \mathbf{\Theta}_{m_0}$ satisfying $p_0 = p_{\boldsymbol{\theta}^*_{m_0}}$).

Assumption 1. *There exists $g(N)$ for which there is a distribution $\rho_{m_0,N,V} \in \mathcal{F}^+(\mathbf{\Theta}_{m_0})$ such that*

$$\int KL(p_{\boldsymbol{\theta}^*_{m_0}}, p_{\boldsymbol{\theta}_{m_0}})\rho_{m_0,N,V}d\boldsymbol{\theta}_{m_0} \leq g(N), \tag{29}$$

$$KL(\rho_{m_0,N,V}, \pi_{m_0}(\boldsymbol{\theta}_{m_0})) \leq N \cdot g(N) \tag{30}$$

where $\pi_{m_0}(\cdot)$ on $\boldsymbol{\theta}_{m_0} \in \mathbf{\Theta}_{m_0}$ is a prior over model $\mathcal{M}_{m_0}$.

Theorem 1. *Given assumption 1, for any $\alpha \in (0, 1)$, if there exists a true model $\mathcal{M}_{m_0}$ such that $p_0 = p_{\mathbf{W}^*_{m_0}}$ and the coefficients of $\mathbf{W}^*_{m_0} = [\mathbf{W}^*_{1,m_0}; \mathbf{W}^*_{2,m_0}; \ldots; \mathbf{W}^*_{V,m_0}] \in \mathbb{R}^{\sum d^v \times m_0}$ are bounded, then*

$$\begin{aligned}&\mathbb{E}\Big[\int D_\alpha(p_{\mathbf{W}_m}, p_{\mathbf{W}^*_{m_0}}) \cdot \hat{\pi}^m_{\alpha,N,V}(\mathbf{W}_m|\mathbf{X}_V^N)\,d\mathbf{W}_m\Big] \\ &\qquad = \mathcal{O}\Big(\frac{\sum_{v=1}^V d^v m_0\log(\sum_{v=1}^V d^v N)}{N}\Big)\end{aligned} \tag{31}$$

where $\hat{\pi}^m_{\alpha,N,V}(\mathbf{W}_m|\mathbf{X}_V^N)$ is the approximate posterior distribution derived by variational inference.

From Eq. 31, we see that, in expectation *w.r.t* the random variables $\mathbf{X}_V^N$ under distribution $p_{\mathbf{W}^*_{m_0}}$, the average α-Renyi loss (D_α) [Van Erven and Harremos, 2014; Chérief-Abdellatif, 2018] between a distribution in the selected model and the true distribution over $\hat{\pi}^m_{\alpha,N,V}(\mathbf{W}_m|\mathbf{X}_V^N)$ goes to zero as $n \to +\infty$. This shows the Bayesian estimator of our model is consistent.

Algorithm 1 Compute Optimal Threshold

```
Input: Data $\{\mathcal{X}, \mathcal{X}'\}$, Swapping Rate $\gamma$, Detection Rate $\zeta$
Output: Detection Threshold $\hat{\tau}_\zeta$
1: Generate mixture set $\mathcal{X}^\gamma$ via swapping views randomly.
2: Compute anomaly scores for all points in $\mathcal{X}$ and $\mathcal{X}^\gamma$ via
   Eq. (28), and denote them as $\mathcal{S}$ and $\mathcal{S}^\gamma$ respectively.
3: Calculate empirical CDF $\hat{F}_a$.
4: Optimize threshold by
   $\hat{\tau}_\zeta = \max\left\{s(\mathbf{x}) \in \{\mathcal{S}, \mathcal{S}^\gamma\} \middle| \hat{F}_a(s(\mathbf{x})) \leq 1 - \zeta\right\}$
```

4.2 Sample Complexity of Multi-View Detector

Let $\{\mathcal{X}, \mathcal{X}'\}$ be two "clean" nominal sets (both containing k *i.i.d.* multi-view draws from p_0). We take $\mathcal{X}$ as training input. For $\mathcal{X}'$, given a swapping rate γ, we use it to generate a 'mixture' dataset $\mathcal{X}^\gamma$. In this case, the mixture data $\mathcal{X}^\gamma$ can be approximately treated as k points drawn from a mixture distribution p_γ, which generates a multi-view outlier and nominal point with probability 2γ and 1–2γ respectively. Our multi-view semi-supervised anomaly detector is trained on $\mathcal{X}$ and assigns anomaly scores to all data points in $\mathcal{X}$ and $\mathcal{X}^\gamma$. Intuitively, an ideal detector would rate all alien data points higher than all nominals (higher score means more anomalous). The key challenge in practice is to select a threshold for anomaly score that gives the guarantee on achieving the desired outlier detection rate. Motivated by [Liu *et al.*, 2018], our approach to obtaining a theoretical guarantee is based on considering the cumulative distribution function (CDF) over anomaly scores.

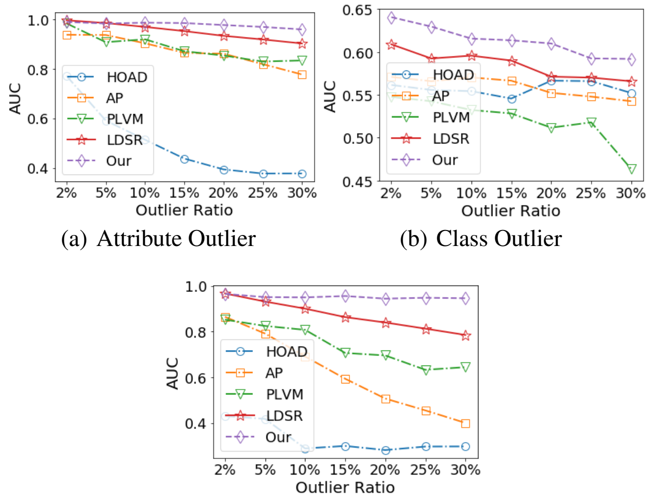
Let $\hat{F}$ and $\hat{F}_\gamma$ be the empirical CDFs of anomaly scores of samples from p_0 and p_γ respectively. The empirical CDF for an abnormal sample can be derived as $\hat{F}_a(s(\mathbf{x})) = \big(\hat{F}_\gamma(s(\mathbf{x})) - (1-2\gamma)\hat{F}(s(\mathbf{x}))\big)/2\gamma$. With sufficient data and knowledge of γ, empirical CDFs $\hat{F}$, $\hat{F}_\gamma$ and $\hat{F}_a$ will convergence to the ground truth F, F_γ and F_a. After deriving F_a, a detector can achieve an outlier detection rate of ζ by selecting an anomaly score threshold τ_ζ that is the $1-\zeta$ quantile of F_a and raises an alarm on the testing point whose anomaly score is greater than τ_ζ. Alg. 1 summarizes the steps for finding a reasonable threshold achieving the desired outlier detection.

Theorem 2. *Let $\mathcal{X}$ and $\mathcal{X}'$ be the nominal data sets containing k i.i.d V-view instances drawn from distribution p_0. Given a swapping rate γ, let $\mathcal{X}^\gamma$ be the mixture set generated from $\mathcal{X}'$ over the randomness of view selecting and view swapping. For any $\epsilon \in (0, \zeta)$ and $\delta \in (0, 1)$, if*

$$k > \frac{(1-\gamma)^2}{2\epsilon^2\gamma^2} \ln \frac{2}{1 - \sqrt{1 - g(\delta, V)}} \tag{32}$$

where $\frac{1}{g(\delta,V)} = \frac{1}{\delta}\sum_{\tilde{V}=2}^{V} \frac{V!}{\tilde{V}!(V-\tilde{V})!}\left\lfloor \frac{\tilde{V}!}{e} \right\rfloor$, $\lfloor\cdot\rfloor$ is the floor function, and e is Euler's Number, then with probability at least $1-\delta$, Algorithm 1 will output a threshold $\hat{\tau}_\zeta$ that achieves an multi-view outlier detection rate of at least $1-\eta$, where $\eta = 1 - \zeta + \epsilon$.

Theorem 2 provides a value for the sample size k that guarantees at least $\zeta - \epsilon$ fraction of outliers in the test points will be detected (an additional error term ϵ is introduced here because of the finite sample size). By using Stirling's formula for approximating factorials (e.g. $\tilde{V}!$, $(V-\tilde{V})!$), above guarantee is approximately polynomial since k grows as $\mathcal{O}\big(\frac{1}{\epsilon^2\gamma^2}\big(V^2 \ln V + \ln\frac{1}{\delta}\big)\big)$. We believe theorem 2 is the first PAC-style guarantee for multi-view anomaly detection.



(a) Attribute Outlier

(b) Class Outlier

(c) Class-Attribute Outlier

Figure 3: The variation curves of AUC W.R.T outlier ratio.

5 Experimental Evaluations

We now show the effectiveness of proposed method on public Outlier Detection Datasets (ODDS)[2], WebKB dataset[3] and MovieLens dataset[4]. We compared the proposed model with representative and cutting edge multi-view anomaly detectors: HOrizontal Anomaly Detection (HOAD) [Gao *et al.*, 2011], Affinity Propagation (AP) [Marcos Alvarez *et al.*, 2013], Probabilistic Latent Variable Model (PLVM) [Iwata and Yamada, 2016] and Latent Discriminant Subspace Representation (LDSR) [Li *et al.*, 2018a]. For AD problems, the most widely used performance evaluation metrics are ROC curve and AUC score.

5.1 Evaluation on Synthetic Multi-View Settings

We employ 9 data sets, namely *thyroid, annthyroid, forestcover, vowels, pima, vertebral, lympho, wine* and *glass*, which are obtained from the ODDS library [Rayana, 2016]. We generate multiple views by randomly splitting the features, where each feature can belong to only one view. To generate three types of multi-view outliers, we follow the strategy in previous works (e.g. [Li *et al.*, 2018a]) for fair comparison. After the outlier generation stage, we equivalently split all normal instances into two parts, and use one of them as the training set to train the proposed model. Then we verify the outlier detection performance on the test set which consists of the remaining normal data and generated outliers.

[2] http://odds.cs.stonybrook.edu

[3] http://lig-membres.imag.fr/grimal/data.html

[4] https://grouplens.org/datasets/movielens/latest

	Model	Thyroid	Annthyroid	ForestCover	Vowels	Pima	Vertebral	Lympho	Wine	Glass
Attribute Outlier	HOAD	.5202±.0864	.5078±.0724	.6801±.0866	.8540±.0691	.5921±.0768	.8338±.0972	.5714±.1648	.6503±.1574	.7083±.1410
	AP	.6737±.1164	.5747±.0669	.6774±.0739	.7062±.1125	.9376±.0293	.8586±.0604	.5369±.1539	.6947±.1078	.7497±.1117
	PLVM	.8989±.0091	.8904±.0363	.4870±.0126	.5481±.0067	.9086±.0083	.7564±.0061	.5413±.0251	.4058±.0481	.4087±.0246
	LDSR	.9751±.0074	.9876±.0022	.9983±.0005	.9181±.0153	.9858±.0057	.9793±.0200	**.9362**±.0053	**.9932**±.0009	**.9940**±.0040
	Our	**.9877**±.0056	**.9979**±.0011	**.9995**±.0027	**.9875**±.0071	**.9877**±.0044	**.9958**±.0074	.9225±.0177	.9417±.0450	.9530±.0292
Class Outlier	HOAD	.5393±.0303	.5849±.0348	.6872±.0337	.3818±.0384	.5557±.0310	.5209±.0812	.6058±.0715	.7124±.0638	.4277±.0932
	AP	.5847±.0227	.5265±.0350	.7906±.0332	.7520±.0513	.5659±.0365	.5272±.0449	.7402±.0498	.5629±.0933	.5576±.0518
	PLVM	.5676±.0093	.4087±.0176	.6035±.0044	.5479±.0282	.5425±.0138	.4444±.0416	.5254±.0061	.4860±.0040	.5433±.0104
	LDSR	.8631±.0217	.7128±.0418	.7551±.0293	.9245±.0173	.5924±.0543	.6070±.0568	.8228±.0762	.5889±.0916	.7098±.0498
	Our	**.8744**±.0205	**.7383**±.0450	**.8672**±.0197	**.9360**±.0158	**.6354**±.0400	**.8891**±.1171	**.8825**±.0410	**.8373**±.0424	**.7613**±.0570
Class-Attribute Outlier	HOAD	.4934±.0270	.4976±.0311	.4342±.0468	.5994±.1342	.4181±.0260	.7386±.0700	.7085±.0609	.5798±.0615	.5598±.0652
	AP	.6380±.0723	.5647±.0819	.8054±.0373	.8511±.0713	.7916±.0555	.7277±.0524	.5481±.0918	.5481±.1173	.7308±.0676
	PLVM	.7122±.0191	.8933±.0134	.8184±.0087	.6390±.0223	.8249±.0063	.6913±.0261	.6120±.0195	.7094±.0145	.9555±.0092
	LDSR	.9344±.0179	.9122±.0220	.9845±.0049	.9642±.0064	.9315±.0146	.9185±.0371	**.9765**±.0135	**1**±0	.9900±.0026
	Our	**.9863** ±.0075	**.9842**±.0076	**.9857**±.0095	**.9757**±.0082	**.9510**±.0169	**.9836**±.0198	.9571±.0536	.9201±.0470	**.9984**±.0023

Table 1: AUC values (mean ± std) on nine UCI datasets with outlier ratio 5%.

On each dataset, we repeat the random outlier generation procedure 20 times and at each time, we perturb 2.5% of the data in that procedure. We average their performance and report AUC results (mean ± standard deviation) in Table 1. From table 1, we can observe that the proposed method consistently outperforms all competing counterparts on almost nine data sets for all kinds of multi-view outliers. The superiority of proposed method is expected, because it uses the semi-supervised anomaly detection technique, which can maximally capture the nature and property of normal instances. This, in turn, can help the learned model to better distinguish whether the test instance is normal or not, thus improving the detection performance.

To investigate how the number of outliers affects the performance of different models, we experiment on data corrupted by progressively higher percentages of outliers. The Figure 3 shows the variation of AUCs on data set *pima* with outlier ratio of 2%, 5%, 10%, 15%, 20%, 25% and 30% for three types of outliers. We see that, in general, as the anomaly rate increases, the performance decreases. And the proposed method is comparatively robust and outperforms other compared ones with all outlier ratio settings.

5.2 Evaluation on Real World Multi-View Data

Further, we compare them on the *WebKB* dataset [Blum and Mitchell, 1998] which has been widely used for evaluating multi-view learning algorithms [Guo, 2013; Li *et al.*, 2014]. We use its *Cornell* subset in our experiment. It contains 195 webpages over 5 labels. Each webpage is described by four views: content, inbound link, outbound link and cites. Figure 4 shows the ROC curves of all compared methods on the WebKB dataset with outlier ratio of 5% (left) and 10% (right). We can observe that clearly, our approach achieves higher AUC than its competitors, which demonstrates the strength of our Bayesian detector.

To present the qualitative analysis of Bayesian model in detecting inconsistency between users' rating behavior and movie genre, we apply the proposed model to MovieLens *small* dataset which contains 100,836 ratings over 9,742 movies by 610 users. We sample 1000 movies, and perform our model to calculate anomalous scores for them. Table 2 lists some movies high and low anomalous scores. Movie 'spirited away' is categorized into animation and fantasy genre, but it receives most of ratings from users that watch and tag action-thriller movies. In other words, it exhibits inconsistent behavior between genre view and rating view, and thus has a high anomalous score. Contrarily, low anomalous score movies, e.g. sacrifice, do not show view inconsistency.

Movie Title	Score	Movie Title	Score
Spirited Away	0.982	The Rebound	0.162
Quiz Show	0.966	Scooties	0.150
Dance of Reality	0.962	Winslow Boy	0.092
The Dark Knight	0.956	Sacrifice	0.084

Table 2: High and low anomalous score movies

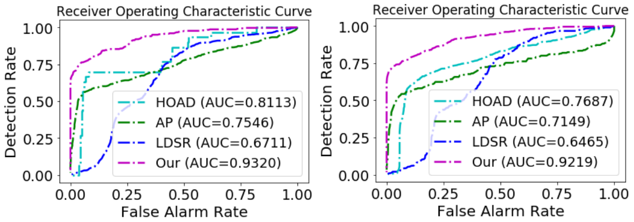


Figure 4: ROC curves of compared methods on WebKB dataset. (PLVM method misses here, because it fails to execute on high dimensional dataset due to exponent overflow of Eq. 10 in their paper.)

6 Conclusion

In this paper, we offer a novel hierarchical Bayesian model to find multi-view outliers under a semi-supervised detection scenario via inductive learning. We prove our Bayesian estimator is consistent and derive a sample complexity for the detector. In experiment, we show the proposed model consistently outperforms state-of-the-art multi-view anomaly detectors across both synthetic and real-world multi-view data.

References

[Akcay *et al.*, 2018] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian Conference on Computer Vision*, pages 622–637. Springer, 2018.

[Archambeau *et al.*, 2006] Cédric Archambeau, Nicolas Delannay, and Michel Verleysen. Robust probabilistic projections. In *ICML*, pages 33–40. ACM, 2006.

[Blanchard *et al.*, 2010] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *JMLR*, 11(Nov):2973–3009, 2010.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100, 1998.

[Chalapathy and Chawla, 2019] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *CoRR*, 2019.

[Chandola *et al.*, 2009] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[Chérief-Abdellatif, 2018] Badr-Eddine Chérief-Abdellatif. Consistency of elbo maximization for model selection. *CoRR*, 2018.

[Das *et al.*, 2016] Shubhomoy Das, Weng-Keen Wong, Thomas Dietterich, Alan Fern, and Andrew Emmott. Incorporating expert feedback into active anomaly discovery. In *ICDM*, pages 853–858. IEEE, 2016.

[Dasgupta *et al.*, 2002] Sanjoy Dasgupta, Michael L Littman, and David A McAllester. Pac generalization bounds for co-training. In *NeurIPS*, pages 375–382, 2002.

[Gai *et al.*, 2008] Jiading Gai, Yong Li, and Robert L Stevenson. Robust bayesian pca with student's t-distribution: the variational inference approach. In *ICIP*, pages 1340–1343. IEEE, 2008.

[Gao *et al.*, 2011] Jing Gao, Wei Fan, Deepak Turaga, Srinivasan Parthasarathy, and Jiawei Han. A spectral framework for detecting inconsistency across multi-source object relationships. In *ICDM*, pages 1050–1055. IEEE, 2011.

[Görnitz *et al.*, 2013] Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. Toward supervised anomaly detection. *JAIR*, 46:235–262, 2013.

[Guo, 2013] Yuhong Guo. Convex subspace representation learning from multi-view data. In *AAAI*, 2013.

[Iwata and Yamada, 2016] Tomoharu Iwata and Makoto Yamada. Multi-view anomaly detection via robust probabilistic latent variable models. In *NeurIPS*, pages 1136–1144, 2016.

[Li *et al.*, 2014] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *AAAI*, 2014.

[Li *et al.*, 2015] Sheng Li, Ming Shao, and Yun Fu. Multi-view low-rank analysis for outlier detection. In *SDM*, pages 748–756. SIAM, 2015.

[Li *et al.*, 2018a] Kai Li, Sheng Li, Zhengming Ding, Weidong Zhang, and Yun Fu. Latent discriminant subspace representations for multi-view outlier detection. In *AAAI*, 2018.

[Li *et al.*, 2018b] Sheng Li, Ming Shao, and Yun Fu. Multi-view low-rank analysis with applications to outlier detection. *TKDD*, 12(3):32, 2018.

[Liu and Lam, 2012] Alexander Y Liu and Dung N Lam. Using consensus clustering for multi-view anomaly detection. In *SPW*, pages 117–124. IEEE, 2012.

[Liu and Rubin, 1995] Chuanhai Liu and Donald B Rubin. Ml estimation of the t distribution using em and its extensions, ecm and ecme. *Statistica Sinica*, pages 19–39, 1995.

[Liu *et al.*, 2018] Si Liu, Risheek Garrepalli, Thomas G Dietterich, Alan Fern, and Dan Hendrycks. Open category detection with pac guarantees. *CoRR*, 2018.

[Marcos Alvarez *et al.*, 2013] Alejandro Marcos Alvarez, Makoto Yamada, Akisato Kimura, and Tomoharu Iwata. Clustering-based anomaly detection in multi-view data. In *CIKM*, pages 1545–1548. ACM, 2013.

[Muñoz-Marí *et al.*, 2010] Jordi Muñoz-Marí, Francesca Bovolo, Luis Gómez-Chova, Lorenzo Bruzzone, and Gustavo Camp-Valls. Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE transactions on geoscience and remote sensing*, 48(8):3188–3197, 2010.

[Neal, 2012] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[Rayana, 2016] Shebuti Rayana. ODDS library", 2016.

[Siddiqui *et al.*, 2018] Md Amran Siddiqui, Alan Fern, Thomas G Dietterich, Ryan Wright, Alec Theriault, and David W Archer. Feedback-guided anomaly discovery via online optimization. In *SIGKDD*, pages 2200–2209. ACM, 2018.

[Song *et al.*, 2017] Hongchao Song, Zhuqing Jiang, Aidong Men, and Bo Yang. A hybrid semi-supervised anomaly detection model for high-dimensional data. *Computational intelligence and neuroscience*, 2017, 2017.

[Van Erven and Harremos, 2014] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[Xu *et al.*, 2013] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *CoRR*, 2013.

[Zhao and Fu, 2015] Handong Zhao and Yun Fu. Dual-regularized multi-view outlier detection. In *IJCAI*, 2015.