Probabilistic Two-Level Anomaly Detection for Correlated Systems

Bin Tong[1] and **Tetsuro Morimura**[2] and **Einoshin Suzuki**[3] and **Tsuyoshi Idé**[4]

Abstract. We propose a novel probabilistic semi-supervised anomaly detection framework for multi-dimensional systems with high correlation among variables. Our method is able to identify both abnormal instances and abnormal variables of an instance.

1 Introduction

Anomaly detection is one of the most practical artificial intelligent problems. It aims at recognizing unusual patterns within normal behaviors. Unlike traditional anomaly detection whose task is to identify the anomalous samples, we invent an anomaly detection framework that is capable of detecting the abnormal at both the *variable* and *instance* levels.[5]

One of pioneering studies on variable-level anomaly detection is presented in [2], in which a sparse Graphical Gaussian Model (GGM) [5] was shown to be effective. However, we found that GGM may fail to achieve a fair performance for high correlated data. Most anomaly detection methods based on Principal Component Analysis (PCA) implicitly assume that abnormal patterns rarely span the *normal subspace*, which is generally referred to as a subspace with main variances [3]. However, this kind of assumption does not always hold for the high correlated data, since most of abnormal patterns are wrapped by normal patterns that lie along the direction of the main variance.

In this paper, by clarifying the relationship between GGM and Probabilistic PCA (PPCA), we propose a probabilistic model for anomaly detection at both the *variable* and *instance* levels. We calculate anomaly scores for both the variables and the instances.

2 Problem Setting

We are given N observed samples represented by a centered matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$. Each sample $\mathbf{x}_i$ $(i = 1, 2, \dots, N)$ is denoted by a D-dimensional vector $[x_1, x_2, \dots, x_D]^T$. A label vector for the samples is defined as $\mathbf{p} = [p_1, p_2, \dots, p_N]$, where p_i is the label of $\mathbf{x}_i$. In the practical setting of anomaly detection, $p_i \in \{1, 2, 3\}$ in which 1, 2 and 3 represent normal label, abnormal label, and unknown labels, respectively. For the variables, we also define a label matrix $\mathbf{V}$ the same size as $\mathbf{X}$, in which the ij-th entry is set to be 1, 2 or 3, if the corresponding variable is normal, abnormal, or in an unknown state, respectively. Our task is to identify which variables and which samples are in abnormal states.

3 Relationship between GGM and PPCA

In GGM, D-dimensional random variables are modeled by a Gaussian distribution, which is associated with a graph with D nodes (variables) and a set of edges. Two variables without an edge indicates the two variables are conditionally independent given the other variables. The edge connections among nodes can be represented by a *precision matrix*. The logarithm of likelihood for the Gaussian distribution is written as:

$$J_{GGM}(\mathbf{\Lambda}) = \ln\det\mathbf{\Lambda} - \text{tr}(\mathbf{S}\mathbf{\Lambda}) + \text{const.} \qquad (1)$$

where $\mathbf{\Lambda}$ denotes the *precision matrix*, $\mathbf{S}$ represents the empirical estimate of covariance matrix, which is calculated as $\mathbf{S} = N^{-1}\mathbf{X}\mathbf{X}^T$, tr denotes the trace operator, and det represents the determinant of a matrix. In PPCA [4], a linear mapping for each observed data $\mathbf{x}_n$ that is corrupted by noise is defined as:

$$\mathbf{x}_n = \mathbf{W}\mathbf{z}_n + \boldsymbol{\eta}_n \qquad (2)$$

where the mapping matrix is $\mathbf{W} \in \mathbb{R}^{D\times D}$ if all dimensions are kept, $\mathbf{z}_n \in \mathbb{R}^D$ is a latent vector having a Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ where $\mathbf{I}$ is an identity matrix, $\boldsymbol{\eta}_n$ is a noise vector having a Gaussian distribution $\mathcal{N}(0, \beta^2\mathbf{I})$ where β^2 is a variance. By imposing a Gaussian prior for the latent data, the logarithm of likelihood of $\mathbf{W}$ after marginalizing over $\mathbf{z}_n$ is written as:

$$J_{PPCA}(\mathbf{W}) = -\ln\det\mathbf{C} - \text{tr}(\mathbf{C}^{-1}\mathbf{S}) + \text{const.} \qquad (3)$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \beta^2\mathbf{I}$. Through the equality $\ln\det\mathbf{C} = -\ln\det\mathbf{C}^{-1}$, we see that the precision matrix $\mathbf{\Lambda}$ in Eq. (1) corresponds to $\mathbf{C}^{-1}$ in Eq. (3).

From the viewpoint of optimizing $\mathbf{C}$, PPCA can be considered as a parameterized version of GGM, since $\mathbf{C}$ is parameterized into the form $\mathbf{W}\mathbf{W}^T + \beta^2\mathbf{I}$. The relationship between GGM and PPCA provides a novel perspective on the transformation matrix $\mathbf{W}$ to understand the precision matrix.

4 Probabilistic Two-Level Anomaly Detection

In order to integrate the supervised information on variables and instances, we naturally extend PPCA to a matrix-variate linear model and derive anomaly scores by using the relation between GGM and PPCA.

4.1 Probabilistic Model

Starting from a linear model, we can write Eq. (2) into a matrix form:

$$\mathbf{X} = \mathbf{Y} + \mathbf{\Psi} \qquad (4)$$

[1] Central Research Laboratory, Hitachi, email: bin.tong.hh@hitachi.com

[2] IBM Research - Tokyo, email: tetsuro@jp.ibm.com

[3] Kyushu University, email: suzuki@inf.kyushu-u.ac.jp

[4] IBM T.J. Watson Research Center, email: tide@us.ibm.com

[5] This work was mainly done during Bin Tong's internship at IBM Research - Tokyo.

where $\mathbf{Y} = \mathbf{W}\mathbf{Z}$, $\mathbf{X}$ is the data matrix, $\mathbf{W} \in \mathbb{R}^{D\times D}$ is the mapping matrix, $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N] \in \mathbb{R}^{D\times N}$ is the latent data matrix for $\mathbf{X}$ where each $\mathbf{z}_i$ $(i = 1, 2, \ldots, N)$ is a D-dimensional vector $[z_1, z_2, \ldots, z_D]^T$, and $\mathbf{\Psi}$ is a noise matrix. We assume that $\mathbf{Z}$ is drawn from a matrix-variate normal distribution [1], a matrix extension of Gaussian for vectors, $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D, \mathbf{K}_N)$, where $\mathbf{I}_D \in \mathbb{R}^{D\times D}$ represents the row covariance matrix which encodes the relationships among the variables, and $\mathbf{K}_N \in \mathbb{R}^{N\times N}$ denotes the column covariance matrix which describes the relationships among the instances. The distribution of $\mathbf{Z}$ is with the mean of a zero matrix, each row independent of each other, and each column correlated with $\mathbf{K}_N$.

We start to build a generative model. In anomaly detection, we attempt to learn a generative model for generating the normal data with high probabilities and the abnormal data with low probabilities. In addition, it is often the case that, even if an instance is labeled as abnormal, some variables of the instance may be normal. Inspired by this observation, we define a vector $\mathbf{a} = [\alpha_1^2, \alpha_2^2, \alpha_3^2]^T$ where α_1^2, α_2^2, and α_3^2 represent the variances of the Gaussian noise for normal variables, abnormal variables, and variables with unknown labels, respectively. In a general setting, it follows $\alpha_1^2 \leq \alpha_3^2 \leq \alpha_2^2$, since a generative model tends to generate the data with high probabilities, which have low degrees of noise. Therefore, using the label matrix $\mathbf{V}$ for the variables, the conditional probability of the observed data $\mathbf{X}$ can be defined as:

$$p(\mathbf{X}|\mathbf{Y}, \mathbf{V}, \mathbf{a}) = \prod_{i=1}^{D}\prod_{j=1}^{N} \mathcal{N}(\mathbf{X}_{ij}|(\mathbf{Y})_{ij}, \alpha^2_{\mathbf{V}_{ij}}). \tag{5}$$

In order to integrate the supervised information on the instances, Gaussian Random Field (GRF) is utilized. Following the idea in [6], we can write the distribution of $\mathbf{Z}$ as follows.

$$p(\mathbf{Z}) = \frac{1}{F'}\exp\left\{-\frac{\tau}{2}\mathrm{tr}\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^T\right)\right\} \tag{6}$$

where F' denotes a constant, τ is a scale parameter, and $\mathbf{L}$ is seen as a Laplacian matrix for a similarity matrix $\mathbf{G}$ that encodes the supervised information on the instances. The entries of $\mathbf{G}$ are defined as:

$$\mathbf{G}_{ij} = \begin{cases} 1, & \mathbf{x}_i \text{ and } \mathbf{x}_\mathrm{j} \text{ (i} \neq \text{j)} \in \text{ normal class} \\ \theta, & \mathbf{x}_i \text{ and } \mathbf{x}_\mathrm{j} \text{ (i} \neq \text{j)} \in \text{ unlabeled class} \\ \delta, & \mathbf{x}_i \in \text{ normal class}, \mathbf{x}_\mathrm{j} \in \text{ unlabeled class} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

such that $\mathbf{L} = \mathbf{D} - \mathbf{G}$ where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{G}_{ij}$, θ and $\delta \in [0, 1]$. From the definition of $\mathbf{G}$, we can see that the larger the value of $\mathbf{G}_{ij}$ is, the closer $\mathbf{x}_i$ and $\mathbf{x}_j$ are to each other. In Eq. (6), we interpret that $\mathbf{Z}$ follows a Gaussian distribution with precision matrix $\mathbf{L}$. Compared with the definition for $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D, \mathbf{K}_N)$, we have $\mathbf{L} = \mathbf{K}_N^{-1}$. According to the Lemma on pp. 64 of [1], we can define a prior for $\mathbf{Y}$, which is a matrix variate normal distribution on $\mathbf{W}\mathbf{Z}$, as follows:

$$\mathbf{Y} = \mathbf{W}\mathbf{Z} \sim \mathcal{N}_{D,N}(\mathbf{0}, \mathbf{W}\mathbf{W}^T, \mathbf{K}_N). \tag{8}$$

With the prior in Eq. (8) and the likelihood in Eq. (5), the posterior distribution of $\mathbf{Y}$ is defined below.

$$p(\mathbf{Y}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y}) \tag{9}$$

The MAP estimate of $\mathbf{Y}$ can be obtained by minimizing the negative logarithm of Eq. (9). For details of the optimization on both $\mathbf{W}$ and $\mathbf{Z}$, refer to Section 2 of the supplementary document [6].

4.2 Anomaly Score

After obtaining $\mathbf{W}$ through the optimization, the precision matrix $\mathbf{\Lambda}$ for the distribution on $\mathbf{X}$ is calculated as $(\mathbf{W}\mathbf{W}^T + \beta^2\mathbf{I})^{-1}$. Given an instance $\mathbf{x}$, the abnormal scores $\mathbf{s} = [s_1, s_2, \ldots, s_D]$ for all variables are calculated as:

$$\mathbf{s} \equiv \mathbf{s}_0 + \frac{1}{2}\mathrm{diag}(\mathbf{\Lambda}\mathbf{x}\mathbf{x}^T\mathbf{\Lambda}\mathbf{P}^{-1}) \tag{10}$$

where $\mathrm{diag}(\cdot)$ represents a vector in which the elements correspond to the diagonal elements of a matrix. The matrix $\mathbf{P} = \mathrm{diag}^2(\mathbf{\Lambda})$ where $\mathrm{diag}^2(\cdot)$ denotes a matrix with the diagonal elements of a matrix and zero off-diagonal elements. The vector $\mathbf{s}_0$ is defined so that $(\mathbf{s}_0)_i = \frac{1}{2}\ln\frac{2\pi}{\mathbf{\Lambda}_{i,i}}$.

With respect to the anomaly scores for the instances, we first normalize $\mathbf{s}$, which is denoted by $\mathbf{b} = [b_1, b_2, \ldots, b_D]$. Given an instance $\mathbf{x}$, its abnormal score, which is derived from Rényi entropy of order λ, is defined as:

$$\mathbf{t} \equiv \frac{1}{\lambda - 1}\ln\left(\sum_{i=1}^{D} b_i^\lambda\right). \tag{11}$$

For the detailed discussion on anomaly score, refer to Section 3 of the supplementary document [6].

5 Experiment

As a case study, we made an experiment on the high correlated data from a train sensor system. We denote our method Probabilistic Two-Level Anomaly Detection as PTLAD, an extension of Glasso [2] as EGlasso, an supervised extension of GLasso as SEGlasso (k), where the first k $(k = 1, \ldots, D)$ directions of main variances of data are removed. We utilize Signal to Noise Ratio (SNR) to evaluate the differences between the anomaly scores for normal and abnormal data. Table 1 presents SNRs for the instances and the average values over variables, showing that PTLAD outperforms the other methods. For the detailed discussion on the experiment, refer to Section 4 of the supplementary document [6].

Table 1: SNRs for variables and instances

Type	PTLAD	SEGlasso (k=1)	EGlasso	JSPCA[3]
Ave. Variable	25.58	3.22	3.22	9.30
Instance	4.14	1.49	0.39	0.60

6 Conclusion

We clarified the relationship between GGM and PPCA, and proposed a novel anomaly detection framework at both the *variable* and *instance* levels for high correlated data.

REFERENCES

[1] A. K. Gupta, *Matrix Variate Distributions*, Chapman & Hall/CRC, October 1999.

[2] T. Idé, A. C. Lozano, N. Abe, and Y. Liu, ‘Proximity-Based Anomaly Detection Using Sparse Structure Learning’, in *SDM*, pp. 97–108, (2009).

[3] R. Jiang, H. Fei, and J. Huan, ‘Anomaly Localization for Network Data Streams with Graph Joint Sparse PCA’, in *KDD*, pp. 886–894, (2011).

[4] M. E. Tipping and C. M. Bishop, ‘Probabilistic Principal Component Analysis’, *Journal of the Royal Statistical Society*, **61**, 611–622, (1999).

[5] N. Meinshausen, P. Bhlmann, and E. Zrich, ‘High Dimensional Graphs and Variable Selection with the Lasso’, *Annals of Statistics*, **34**, 1436–1462, (2006).

[6] G. Zhong, W. Li, D. Yeung, X. Hou, and C.-L. Liu, ‘Gaussian Process Latent Random Field’, in *AAAI*, pp. 679–684, (2010).

[6] `http://ide-research.net/papers/ecai14_doc.pdf`