Supplemental Material of "Probabilistic Two-Level Anomaly Detection for Correlated Systems"

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

1 Introduction

This is a supplementary document for the paper [8], named Probabilistic Two-Level Anomaly Detection for Correlated Systems, in 21^{st} European Conference on Artificial Intelligence (ECAI 2014). This document makes more detailed discussions on the optimization of the probabilistic model, the calculation of the anomaly score, and the experiment for this paper.

2 Optimization

As shown in Eq. (9) in Section 4.1 of the paper, the posterior distribution of $\mathbf{Y}$ is defined below.

$$p(\mathbf{Y}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y}) \tag{1}$$

The MAP estimate of $\mathbf{Y}$ can be obtained by minimizing the negative logarithm of Eq. (1). By removing out terms unrelated to $\mathbf{Z}$ and $\mathbf{W}$, the objective function is rewritten as:

$$\begin{aligned} J = &\sum_{i=i}^{D}\sum_{j=1}^{N} S_{ij}(\mathbf{X}_{ij} - (\mathbf{WZ})_{ij})^2 + N \ln \text{det}(\mathbf{WW}^T) \\ &+ \text{tr}\left((\mathbf{WW}^T)^{-1}(\mathbf{WZ})\mathbf{L}(\mathbf{WZ})^T\right) \end{aligned} \tag{2}$$

where S_{ij} represents $\alpha_{\mathbf{V}_{ij}}^{-2}$. In this section, we mainly discuss the optimization of Eq. (2). Since both $\mathbf{W}$ and $\mathbf{Z}$ are regarded as variables, we can not give closed-form solutions. However, Eq. (2) can be iteratively optimized with respect to $\mathbf{Z}$ by fixing $\mathbf{W}$, and vice versa. By doing so, a local optimal can be reached since the value of Eq. (2) decreases or stays steadily even after a small number of updates.

2.1 Optimize Z for a fixed W.

To simplify the third term of Eq. (2), it is natural to assume $\mathbf{W}$ is a full rank matrix, so that $\mathbf{W}$ is invertible. We can then succinctly write the third term of Eq. (2) into the following form.

$$\text{tr}\left((\mathbf{WW}^T)^{-1}(\mathbf{WZ})\mathbf{L}(\mathbf{WZ})^T\right) = \text{tr}\left(\mathbf{ZLZ}^T\right) \tag{3}$$

Fixing $\mathbf{W}$, we then optimize the objective function with respect to $\mathbf{Z}$:

$$J_Z = \sum_{i=i}^{D}\sum_{j=1}^{N} S_{ij}(\mathbf{X}_{ij} - (\mathbf{WZ})_{ij})^2 + \text{tr}\left(\mathbf{ZLZ}^T\right) \tag{4}$$

By making the derivative with respect to $\mathbf{Z}$ (see Appendix. A), we have

$$\frac{\partial J_Z}{\partial \mathbf{Z}} = -2\mathbf{W}^T(S \circ \mathbf{X}) + 2\mathbf{W}^T(S \circ \mathbf{Y}) + 2\mathbf{ZL} \tag{5}$$

where $A \circ B$ represents Hadamard product for which $(A \circ B)_{ij} = A_{ij}B_{ij}$. We are unable to obtain a closed-form solution for $\mathbf{Z}$ by setting Eq. (5) to zero. Therefore, we attempt to utilize the gradient-based method in which $\mathbf{Z}$ is updated as follows.

$$\mathbf{Z}^{t+1} = \mathbf{Z}^t - \mu\frac{\partial J_Z}{\partial \mathbf{Z}^t} \tag{6}$$

where μ is a constant that indicates the step. $\mathbf{Z}$ is iteratively updated until the convergence condition is satisfied.

2.2 Optimize W for a fixed Z.

When $\mathbf{Z}$ is fixed, the objective function Eq. (2) with respect to $\mathbf{W}$ can be written as:

$$J_W = \sum_{i=i}^{D}\sum_{j=1}^{N} S_{ij}(\mathbf{X}_{ij} - (\mathbf{WZ})_{ij})^2 + N \ln \text{det}(\mathbf{WW}^T) \tag{7}$$

According to [7], $\ln\det$ is a concave function. Thus, we consider relaxing the $\ln\det$ term into a convex form. We relax Eq. (7) and decompose it into D separable optimization problems (see Appendix. B). That is, for $i = 1, 2, \ldots, D$, we have

$$J_W^i = \sum_{j=1}^{N} S_{ij}(\mathbf{X}_{ij} - \widehat{\mathbf{w}}_i\mathbf{z}_j)^2 + N||\widehat{\mathbf{w}}_i||_2 \tag{8}$$

where $\widehat{\mathbf{w}}_i$ represents the i-th row of $\mathbf{W}$, and $\mathbf{z}_j$ is the j-th column of $\mathbf{Z}$. We can see that Eq. (8) is a *weighted least squares* problem with a l_2 regularization for $\widehat{\mathbf{w}}_i$. We can obtain the closed-form solution for each $\widehat{\mathbf{w}}_i$ by setting $\frac{\partial J_W^i}{\partial \widehat{\mathbf{w}}_\mathbf{i}} = 0$. It can be represented as

$$\widehat{\mathbf{w}}_i = \frac{\sum\limits_{j=1}^{N} S_{ij}\mathbf{X}_{ij}\mathbf{z}_j^T}{\sum\limits_{j=1}^{N} S_{ij}\mathbf{z}_j^T\mathbf{z}_j + N} \tag{9}$$

To maintain $\mathbf{W}$ as a full rank matrix, we define a SVD decomposition for $\mathbf{W}_t$ in the t-th iteration as $\mathbf{W}_t = U\Sigma V$. The diagonal elements of Σ represent the singular values of $\mathbf{W}_t$, which are denoted as $[\gamma_1, \gamma_2, \ldots, \gamma_D]$. If $\mathbf{W}_t$ is a low rank matrix, some singular values will be zeros. To keep $\mathbf{W}_t$ as a full rank matrix, we define $\Sigma^{'}$ for which the diagonal element γ_i is replaced with a nonzero value if

[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

γ_i is zero. We then have $\mathbf{W}_{t+1}$ in the $t+1$-th iteration as $\mathbf{W}_{t+1} = U\Sigma^{'}V$. It can easily be proved that $||\mathbf{W}_{t+1} - \mathbf{W}_t||_2^2 = \sum_{i=1}^{d} \gamma_i^2$ where d denotes the number of zero singular values for $\mathbf{W}_t$. If γ_i is set to be a reasonably small value, e.g., 0.01, $||\mathbf{W}_{t+1} - \mathbf{W}_t||_2^2$ is then quite small.

3 Anomaly Score

Compared with Section 4.2 of the paper, more explanations on the reason of choosing Eq. (10) and Eq. (11) are made, when calculating the anomaly scores for instances and variables, respectively.

With respect to the abnormal scores of variables, it is natural to assume that the variables in an instance are jointly Gaussian, since $\mathbf{X}$ follows a matrix-variate normal distribution with noise. Therefore, given an instance, we can take the *conditional distribution* of a variable given other variables as the abnormal score for the variable. As mentioned before, the *generative model* tends to give high probabilities to the normal variables and low probabilities to the abnormal variables. By using negative of logarithm, the abnormal variable is then given a high anomaly score while a normal variable is given a low anomaly score.

We obtain the optimal $\mathbf{W}$ from Eq. (2). 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 (see Appendix. C) as:

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

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

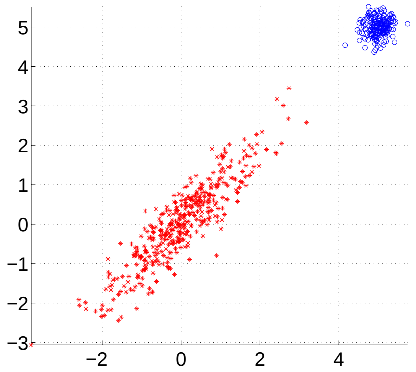


Figure 1: Toy example

In the calculation of the anomaly scores for the instances, we first normalize $\mathbf{s}$, which is denoted by $\mathbf{b} = [b_1, b_2, \ldots, b_D]$. A larger value of b_i represents a higher probability of an abnormal variable, and the abnormal variables with high probabilities are also expected to contribute more to a high anomaly score of an instance than the other normal variables with low probabilities. Rényi entropy [6] of order λ satisfies this requirement, since a larger value of the order λ gives an entropy value that is increasingly determined by the higher probability events. This is also the reason that we do not use other entropies, such as Shannon entropy, etc. Given an instance $\mathbf{x}$, its abnormal score derived from Rényi entropy of order λ is defined as

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

We empirically set λ to be 10 in the experiments.

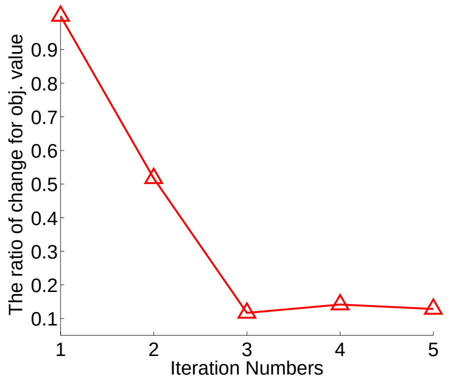


Figure 2: Convergence in toy example

4 Experiments

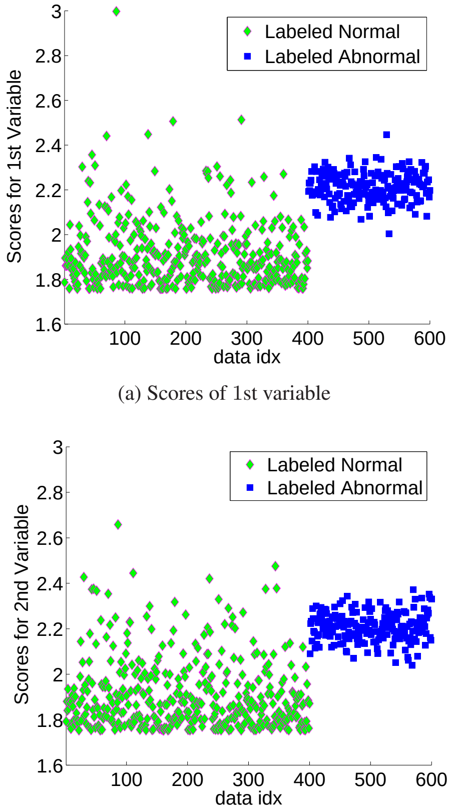
We first generate a toy example to analyze the behaviors of methods before we do experiments on the Train Sensor data for a case study.

4.1 Toy Example

The toy data set was generated as follows. We have two classes data drawn from Gaussian distributions, as shown in Figure 1. For the blue circle points, the mean and covariance matrix are $[5, 5]$ and $\text{diag}(0.05, 0.05)$, respectively. For the red star points, its mean and covariance matrix are $[0, 0]$ and $[1, 0.9; 0.9, 1]$, respectively. We regard the blue circle points as the abnormal data and the red star points are the normal data. We can see that the whole data are of high correlation and the correlation matrix is $[1, 0.98; 0.98, 1]$.

We call our method Probabilistic Two-Level Anomaly Detection (PTLAD). For the comparison, we consider three methods. The first one is an extension of Glasso [4], called EGlasso. Although Glasso was proposed to detect the abnormal in *variable* level for two graphs, it can easily be extended to detect the abnormal at the *variable* level for instances by using Eq. (10), since the precision matrix is known. The second one is an supervised extension of GLasso, called SEGlasso, that involves three steps. (1) Remove the high correlation information upon the first k ($k = 1, \ldots, D$) directions with main variances. Note that $k = 0$ indicates keeping all information; (2) Derive the data in a discriminative subspace by using supervised information; and (3) Estimate the empirical covariance matrix and perform EGlasso. The third one is JSPCA [5]. Since the first step of SEGlasso ($k \geq 1$) has a similar effect with the methods [5, 3], we infer that EGlasso ($k \geq 1$) and JSPCA would fail when the abnormal data locate in the same directions with the main variances.

Figure 3 shows the abnormal scores for the 1st variable and the 2nd variable, in which the anomaly scores for the abnormal instances



(a) Scores of 1st variable

(b) Scores of 2nd variable

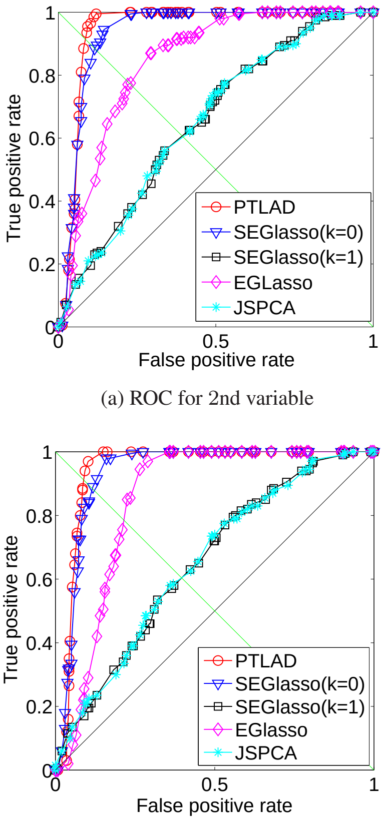
Figure 3: Abnormal scores for variables

are obviously higher than those for most of the normal instances, as shown in Figure 3a and Figure 3b. We may infer that Eq. (10) is effective in calculating the abnormal scores for the variables. Figure 4a and Figure 4b show ROCs for the various methods of anomaly detection for the 2nd variable and for the instances, respectively. PTLAD is able to achieve the best performances in the two cases. It is worth noting that SEGlasso ($k = 0$) outperforms SEGlasso ($k = 1$) and JSPCA. The reason is probably that SEGlasso ($k = 1$) and JSPCA remove the information on the direction with the maximal variance such that the abnormal and normal instances are highly blended in the remanding subspace.

Figure 2 shows PTLAD empirically converges fast. Note that we re-scale the values into the range $(0, 1)$ by simply using $val./max.$ for a better illustration, where $val.$ and $max.$ represent the value of each term and the maximum value during the iterations, respectively. We can see that the value of the objective function tends to be stable after only three iterations.

4.2 Experiments on Case Study

In this experiment, we use the train sensor data set which is reproduced from a real train sensor data. The real data is collected by 40 temperature sensors set in different gear boxes of a train. To eliminate the side-effect of air temperature, we subtract all the temperature values by air temperature. In the real data set, we have 7 normal labeled data and 1 abnormal labeled data with 7-th abnormal variable, others are unlabeled data. We created abnormal instances and variables without loss of generality. We defined a random variable following a Gaussian distribution $\varphi \sim \mathcal{N}(5, 0.5)$. We created 365 artificial abnormal data by randomly adding φ to one or two variables, such that the variable(s) are regarded as abnormal one(s). In statistics, each variable is generally labeled as abnormal one in 10 samples. This means that each variable has 10 chances labeled as abnormal after the artificial processing. According to *experts' experiences*, the unlabeled data is more likely to be normal data. Therefore, for the setting of similarity matrix $\mathbf{G}$, θ and δ are set to be 0.5 and 0.8, respectively.



(a) ROC for 2nd variable

(b) ROC for instances

Figure 4: ROC Performance

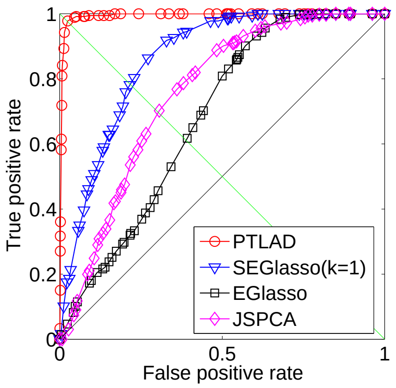


Figure 5: Train: ROC Curve

The parameter settings for PTLAD are defined as: $\beta = 1$, $\alpha_1 = 0.1$, $\alpha_2 = 10$, and $\alpha_3 = 1$. We can see from Figure 5 that the Area Under Curve (AUC) of PTLAD is obviously larger than those of other methods. Since the performances of SEGlasso (k=1) and SEGlasso (k=0) are very similar, we do not show the performance of SEGlasso (k=0) in this experiment. Figure 7a and Figure 7b show the abnormal scores for the 1st variable and the 7-th variable, respectively. We can see that the score of 'labeled abnormal' data in the 7-th variable is much higher than those of other normal data, while its score in the 1st variable is similar to those of other normal data. This is consistent with the fact that only the 7-th variable of the 'labeled abnormal' data is abnormal. We can also see that the scores of the 'artificial abnormal' data in both the 1st and 7-th variables are relatively higher than those of the other normal data. Since the unlabeled data is roughly regarded as normal data, the scores of the 'unlabeled normal' data are similar to those of 'labeled normal' and 'artificial normal'. The abnormal scores for the other 38 variables are not shown, since their score distributions are similar to that of the 1st variable. We employ Signal to Noise Ratio (SNR) to evaluate the differences between the anomaly scores for normal and abnormal data. The higher value of SNR has, the larger the differences on scores between the normal data and abnormal data. Table 1 presents SNRs for the instances and the average values over 40 variables. We can see that SNRs for PTLAD are larger than those of the other methods, which shows the effectiveness of PTLAD for anomaly detection at both the *instance* and *variable* levels. Figure 6 also shows that PTLAD empirically converges within a limited number of iterations.

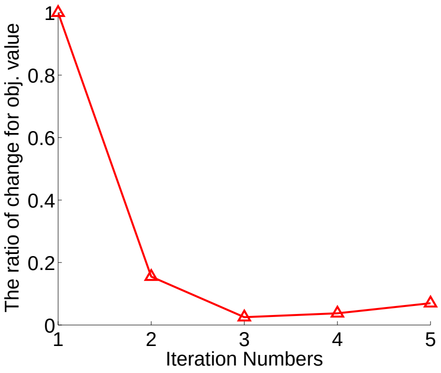
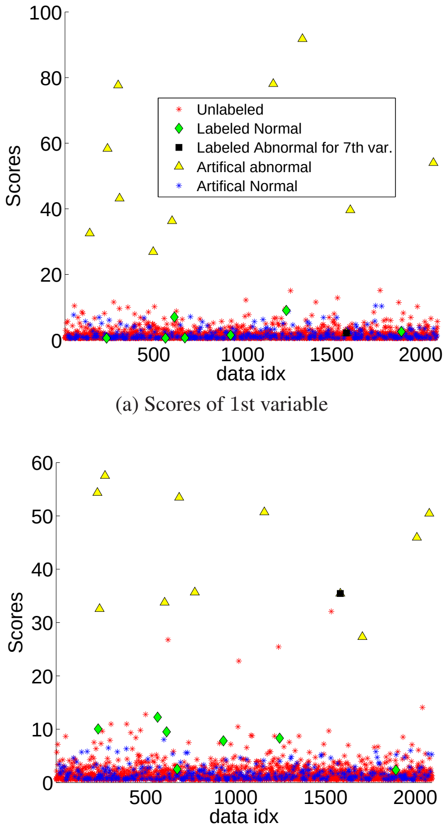


Figure 6: Train: Convergence

Table 1: SNRs for variables and instances

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

We also analyze the parameter setting of the noise degrees for normal data, abnormal data, and unlabeled data. Each α_i $(i = 1, 2, 3)$ changes its value in a searching grid $[100, 10, 1, 0.1, 0.01]$. For a better illustration, the axises are plotted as $1/(1 + \alpha_i)$. Figure 8a presents the changes in the average SNRs over 40 variables when the noise degrees for the normal and unlabeled data, say α_1 and α_3, vary and the noise degree for the abnormal data is fixed at 10. We can see that, when α_1 approaches 0.01, the average SNRs over 40 variables remain the highest. We also examine the situation in which the noise degree for the normal data is fixed at 0.01, as shown in Figure 8b. This shows that the changes for the noise degree of the abnormal and unlabeled data do not distinctly influence the performance. Therefore, we can infer that the noise degree for the normal data is the most important parameter compared with the other two. As mentioned before, this is also consistent with our expectation that the *generative* model tends to give a bias to the normal data.



(a) Scores of 1st variable

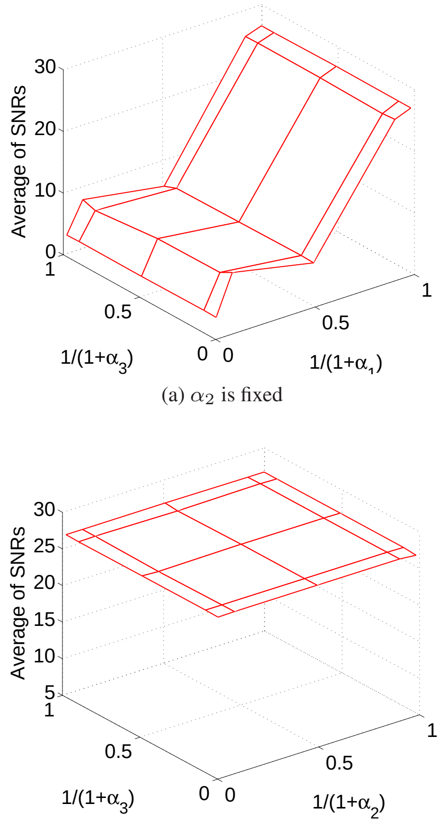
(b) Scores of 7th variable

Figure 7: Abnormal scores for the variables

Appendix: A

Since $\mathbf{Y} = \mathbf{WZ}$, $\mathbf{Y}_{ij}$ can be represented by $\mathbf{W}_{i,*}\mathbf{Z}_{*,j}$, where $\mathbf{W}_{i,*}$ denotes the i-th row of $\mathbf{W}$ and $\mathbf{Z}_{*,j}$ denotes the j-th column of $\mathbf{Z}$. The derivative for the first term of Eq. (4) with respect to $\mathbf{Z}_{ij}$ can be written as follows.

$$\begin{aligned} \frac{\partial P}{\partial \mathbf{Z}_{ij}} &= \frac{\partial}{\partial \mathbf{Z}_{ij}} \sum_{d=1}^{D}\sum_{n=1}^{N} S_{dn}(\mathbf{X}_{dn} - \mathbf{W}_{d,*}\mathbf{Z}_{*,n})^2 \qquad (12) \\ &= -2\sum_{d=1}^{D} S_{dj}\mathbf{X}_{dj}\mathbf{W}_{di} + 2\sum_{d=1}^{D} S_{dj}\mathbf{W}_{d,*}\mathbf{Z}_{*,j}\mathbf{W}_{di} \\ &= -2\sum_{d=1}^{D}(S \circ \mathbf{X})_{dj}\mathbf{W}_{di} + 2\sum_{d=1}^{D}(S \circ \mathbf{Y})_{dj}\mathbf{W}_{di} \\ &= -2\mathbf{W}^T(S \circ \mathbf{X}) + 2\mathbf{W}^T(S \circ \mathbf{Y}) \end{aligned}$$



(a) α_2 is fixed

(b) α_1 is fixed

Figure 8: Analysis for the noise degrees

Appendix: B

Since $\mathbf{W}$ is a full rank matrix, $\mathbf{W}\mathbf{W}^T$ would be a symmetric positive definite matrix. By using the equality $\text{tr}(\ln(A)) = \ln(\det(A))$ [1] for a symmetric matrix A, we have

$$\text{tr}\left(\ln(\mathbf{W}\mathbf{W}^T)\right) = \ln\det(\mathbf{W}\mathbf{W}^T) \tag{13}$$

$\text{tr}\left(\ln(\mathbf{W}\mathbf{W}^T)\right)$ can be written as $\sum_{i=1}^{D}\ln(\widehat{\mathbf{w}}_i\widehat{\mathbf{w}}_i^T)$ where $\widehat{\mathbf{w}}_i\widehat{\mathbf{w}}_i^T = ||\widehat{\mathbf{w}}_i||_2$. By using the inequality $\ln(x) \leq x + 1$, we can relax Eq. (7) and decompose it into D separable optimization problems as shown in Eq. (8).

Appendix: C

Given the precision matrix $\mathbf{\Lambda}$ and an instance $\mathbf{x}$, the probability for $\mathbf{x}$ is calculated as

$$p(\mathbf{x}|\mathbf{0},\,\mathbf{\Lambda}) = \frac{(\det\mathbf{\Lambda})^{\frac{1}{2}}}{(2\pi)^{\frac{D}{2}}}\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{\Lambda}\mathbf{x})$$

We use the conditional distribution for $p(x_i)$ as an example. According to the marginal and conditional distribution given Gaussian distribution [2], we have

$$p(x_1|x_2,\ldots,x_D,\,\mathbf{\Lambda}) = \mathcal{N}(x_1|-\frac{1}{\mathbf{\Lambda}_{1,1}}\sum_{i=2}^{D}\mathbf{\Lambda}_{1,i}x_i,\frac{1}{\mathbf{\Lambda}_{1,1}})$$

By taking a negative logarithm, we obtain

$$-\ln p(x_1|x_2,\ldots,x_D,\,\mathbf{\Lambda}) = \frac{1}{2}\ln\frac{2\pi}{\mathbf{\Lambda}_{1,1}} + \frac{1}{2\mathbf{\Lambda}_{1,1}}\left(\sum_{i=1}^{D}\mathbf{\Lambda}_{1,i}x_i\right)^2$$

Similarly, we can easily derive the conditional distributions for $p(x_i)$ $(x = 2, 3, \ldots, D)$.

REFERENCES

[1] Z. Bai and G. H. Golub, ‘Bounds for the Trace of the Inverse and the Determinant of Symmetric Positive Definite Matrices’, *Annals of Numerical Mathematics*, **4**, 29–38, (1997).

[2] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 1st ed. 2006. corr. 2nd printing edn., October 2007.

[3] L. Huang, X. L. Nguyen, M. N. Garofalakis, M. I. Jordan, A. D. Joseph, and N. Taft, ‘In-Network PCA and Anomaly Detection’, in *NIPS*, pp. 617–624, (2006).

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

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

[6] A. Renyi, ‘On Measures of Entropy and Information’, in *Berkeley Symposium Mathematics, Statistics, and Probability*, pp. 547–561. University of California Press, (1960).

[7] B. Stephen and V. Lieven, *Convex Optimization*, Cambridge University Press, March 2004.

[8] B. Tong, T. Morimura, E. Suzuki, and T. Idé, ‘Probabilistic Two-Level Anomaly Detection for Correlated Systems’, in *ECAI*, (2014).