ANOMALY DETECTION AND LOCALIZATION
BASED ON DOUBLE KERNELIZED SCORING AND MATRIX KERnels

SHUNSUKE HIROSE, TOMOTAKE KOZU AND YINGZI JIN

Deloitte Touche Tohmatsu LLC

Abstract. Anomaly detection is necessary for proper and safe operation of large-scale systems consisting of multiple devices, networks, and/or plants. Those systems are often characterized by a pair of multivariate datasets. To detect anomaly in such a system and localize element(s) associated with anomaly, one would need to estimate scores that quantify anomalousness of the entire system as well as its elements. However, it is not trivial to estimate such scores by considering changes of relationships between the elements, which strongly correlate with each other. Moreover, it is necessary to estimate the scores for the entire system and its elements from a single framework, in order to identify relationships among the scores for localizing elements associated with anomaly. Here, we developed a new method to quantify anomalousness of an entire system and its elements simultaneously.

The purpose of this paper is threefold. The first one is to propose a new anomaly detection method: Double Kernelized Scoring (DKS). DKS is a unified framework for entire-system anomaly scoring and element-wise anomaly scoring. Therefore, DKS allows for conducting simultaneously 1) anomaly detection for the entire system and 2) localization for identifying faulty elements responsible for the system anomaly. The second purpose is to propose a new kernel function: Matrix Kernel. The Matrix Kernel is defined between general matrices, which might have different dimensions, allowing for conducting anomaly detection on systems where the number of elements change over time. The third purpose is to demonstrate the effectiveness of the proposed method experimentally. We evaluated the proposed method with synthetic and real time series data. The results demonstrate that DKS is able to detect anomaly and localize the elements associated with it successfully.

1. Introduction

Stable operation in large-scale systems (e.g. factory operation) is necessary to maintain safe environment, because anomalies in such a system could result in severe losses. In case that anomaly or failure in the system emerged, it would be desirable to predict or detect anomalies as quickly as possible. To suppress such losses, it would also be necessary to localize elements associated with anomaly to fix and control the system properly.

Here, we consider performing anomaly detection and localization in an unsupervised fashion from multivariate datasets where each element correlates with each other. Anomaly detection is to decide on whether the entire system (i.e. the entire elements included in the

E-mail address: shunsuke.hirose@tohmatsu.co.jp.

A translation of “Anomaly Detection based on Doubly-Kernelized Scoring and Matrix Kernels” by Shunsuke Hirose and Tomotake Kozu (This article is written in Japanese). DOI: https://doi.org/10.1527/tjsai.AI30-D
multivariate dataset) is anomalous or not. Localization is to identify faulty elements which are responsible for the system anomaly. In order to solve these tasks simultaneously, it is necessary to conduct anomaly detection and localization in a unified framework. If anomaly detection and localization were conducted separately, it would be difficult to conduct localization as it would make the relation between the system-level anomaly and the element-wise anomaly unclear.

Here we also consider the case in which the number of elements change over time. The scope of the application would be limited if such a case was not considered because it is not unusual that the number of elements fluctuates in anomaly detection. For example, we examine the case of anomaly detection from a communications server network. The system is the entire network. Each element is a communications server. Suppose that we want to determine whether the servers in a certain area are anomalous or not. In such cases, the number of elements in the entire system could be altered because of servers newly added or deleted, resulting from malfunction, for example.

The purpose of this paper is threefold. The first one is to propose a new anomaly detection method: Double Kernelized Scoring (DKS). DKS is a unified method for performing anomaly detection and localization simultaneously in a strongly correlating system. To the best of our knowledge, this paper is the first to describe a method for detecting simultaneously entire system anomalies and elements responsible for them using a single framework. The key ideas of DKS are the following. First, we present a dataset using a kernel matrix, an element of which represents a relation between a pair of variables. Second, by combining the kernel between variables and a kernel between matrices, we construct a statistically natural measure that represents degree of change in the relation of variables, which corresponds to the anomalousness. The measure is definable between two variable groups with any number of variables. Thus, it allows us to conduct simultaneously anomaly detection by estimating the measure between a pair of datasets and localization by estimating the measure between a pair of variables.

The second one is to propose a new kernel function, which we named Matrix Kernel. The Matrix Kernel is defined using two matrices to estimate the anomalousness in DKS. In order to make DKS applicable to the aforementioned problem, we construct a kernel so that it has the following properties. First, the input matrices are general matrices, which are not restricted to those representing weighted graphs (ones with non-negative matrix elements). Second, the inputs may have different dimensions. Therefore, by using the Matrix Kernel, DKS is applicable to systems where the number of elements may change over time. Third, the Matrix Kernel is invariant under permutation of the input matrix element index. Therefore, the kernel is insensitive to non-essential changes such as the permutation.

The third one is to demonstrate the effectiveness of the proposed method by the experimental results using three datasets. Past studies have been conducted to analyze univariate time series for conducting anomaly detection [12]. However, real-world systems often consist of multivariate time series that have strong mutual correlation. In such cases, it would be difficult to detect anomalies if each of time series was monitored separately. Therefore, we applied our method to multivariate datasets for performing anomaly detection and localization simultaneously.
1.1. Related work. An element-wise anomaly scoring method has been proposed by using sparse structure learning of covariance matrices, considering correlations between time series [3]. This method detects changes of conditional probabilities of univariate time series, given the other time series as anomalies. Another anomaly detection method was proposed for a pair of elements [4]. This method detects the change of the relationship between a pair of univariate time series as anomalies. Anomaly detection methods for an entire system have also been proposed [5, 6]. These methods detect outliers of eigenvectors or eigenvalues as anomalies.

These previously proposed methods have two critical restrictions. First, they cannot conduct anomaly detection of the entire system and anomaly localization (i.e. anomaly detection of a single element) simultaneously because the methods detect one of the followings: anomaly of the entire system, anomaly of an element, and anomaly of a subset of elements. Second, the previous methods are restricted to treating fixed dimensions. For treating inputs with different dimensions, graph-based anomaly detection methods have been proposed [7, 8]. These methods represent input datasets as graphs and conduct anomaly detection by comparison of the graphs. It becomes possible to treat datasets for which the numbers of elements change by using the graphs. However, this method is unable to conduct anomaly detection and localization simultaneously from a single framework. In the case described in Ref. [7], an anomalous subgraph is detected. Therefore, anomalies of the entire system cannot be detected. In the case of Ref. [8], anomalies of the entire system are detected. For that reason, localization cannot be conducted.

Kernel functions also have often been used to handle inputs with different dimensions. Previous studies proposed kernel functions defined between weighted graphs [9, 10]. They are constructed based on counting subtrees or paths. Therefore, they are applicable to graphs having different numbers of nodes. However, it is difficult to apply them to general matrices. In Ref. [11], kernels defined between matrices were constructed based on a group theoretical approach. To estimate the kernel value, phantom nodes were introduced. They are virtual and are introduced to transform the input matrices to constant dimensional square matrices. Consequently, it becomes difficult to handle general matrices when the upper limit of the input matrix dimension (the constant dimension) is not given in advance. In Ref. [13], a kernel between general matrices has been proposed, the Probability Product Kernel (PPK), which is defined as an inner product between the probability density functions of matrix elements. By definition, PPK loses the information of the matrix structure.

The remainder of this paper is organized as follows. Section 2 provides a problem setting and proposes an anomaly detection and localization method. Section 3 proposes the Matrix Kernel. Section 4 explains the experimentally obtained results. Section 5 presents concluding remarks.

2. Anomaly Detection based on Double Kernelized Scoring

In this section, we propose a method for solving the task of simultaneously performing anomaly detection and localization in multivariate time series in which the number of elements might change over time.
2.1. Problem Setting. Suppose that two multivariate datasets are given as
\[ \mathcal{D} = \{z_1, z_2, \cdots, z_d\}, \quad \mathcal{D}' = \{z'_1, z'_2, \cdots, z'_d\}, \]
where \( \mathcal{D} \) and \( z_i \) represent a dataset and a variable respectively. The numbers of variables, \( d \) and \( d' \), may be different. Hereinafter, we denote a set of all variables as \( \text{system} \).

Assume that a set of \( \text{target} \) variables is given for each dataset. In this paper, \( \text{target} \) does not represent a dependent variable used in supervised learning, but represents variables for which we want to measure anomalousness as discussed below. We denote all the variables as follows;
\[ z = (z_t, z_{\bar{t}})^T, \quad z' = (z'_{t'}, z'_{\bar{t}'})^T, \]
where \( z \) and \( z' \) represent a set of all variables in \( \mathcal{D} \) and that in \( \mathcal{D}' \) respectively. \( z \) (\( z' \)) is divided into \( z_t \) and \( z_{\bar{t}} \) (\( z'_{t'} \) and \( z'_{\bar{t}'} \)), where \( t \) and \( \bar{t} \) (\( t' \) and \( \bar{t}' \)) represent a set of target variables in dataset \( \mathcal{D} \) (\( \mathcal{D}' \)) and its complement respectively.

For a given pair of datasets and given sets of target variables, we want to solve the problem of performing anomaly detection and localization simultaneously. We define the problem as one consisting of the following two tasks. The first one is to estimate a system anomaly score, \( S(z, z') \) that represents the degree to which the system is anomalous (i.e. the higher anomaly score indicates that the system is more anomalous).

The second one is to estimate a set of \( \text{target anomaly scores} \) \( \{S_{t'}(z, z')\}_{t,t'} \), which represents how anomalous variable set \( (t, t') \) is. The higher target anomaly score of the variable set \( (t, t') \) indicates that the set is more likely to be responsible for the system anomaly. For example, if \( t = t' \) and \( t \) represents a single variable \( z \), then \( S_{t'}(z, z') \) represents how anomalous variable \( z \) is.

We propose that the anomaly scores should satisfy the following requirements. First, the system anomaly score \( S(z, z') \) and target anomaly score \( S_{t'}(z, z') \) are estimated using a single framework. If they are estimated from different ones, it becomes difficult to conduct localization because the relation between the scores is unclear. Second, the anomaly scores are statistically natural measures representing difference between the target sets, \( t \) and \( t' \).

2.2. Double Kernelized Scoring. In this section, we propose a method for solving the aforementioned problem. The method estimates the anomaly scores using kernel functions of two kinds. Therefore we designate the method as \( \text{Double Kernelized Scoring} \) (DKS). The overall flow of DKS is summarized as follows.

First, input is a pair of multivariate datasets, \( \mathcal{D} \) and \( \mathcal{D}' \). Here \( \mathcal{D} \) and \( \mathcal{D}' \) consist of \( d \) and \( d' \) variables respectively as in Section 2.1.

From the datasets, we derive a pair of kernel matrices, whose element is defined between variables as follows:
\[ \mathcal{D} \rightarrow \{K(z_i, z_j)\}_{i,j=1}^d, \quad \mathcal{D}' \rightarrow \{K'(z'_i, z'_j)\}_{i,j=1}^{d'}. \]

Note that the kernel is defined between \( \text{variables} \), not between observations. For example, we can use a covariance matrix and a correlation coefficient matrix as a kernel matrix.

We now define scoring targets \( \{(t, t')\} \) (see Section 2.1) consisting of some variables for which we want to estimate anomaly score. For estimating variable-wise anomaly score, we define the targets as \( \{(t, t')\} = \{(z_1, z'_1), \cdots, (z_d, z'_d)\} \), where \( t \) and \( t' \) represent the same single variable. On the other hand, for estimating a system anomaly score, we define the target as
where \((t, t')\) = \((z, z')\), where \(t\) and \(t'\) represent a set of all variables in \(D\) and \(D'\) respectively. The numbers of variables in \(t\) and \(t'\) may be different. Corresponding to \((t, t')\), we designate the kernel matrices as

\[
K = \begin{pmatrix} K_{tt} & K_{t\tilde{t}} \\ K_{\tilde{t}t} & K_{\tilde{t}\tilde{t}} \end{pmatrix}, \quad K' = \begin{pmatrix} K'_{tt'} & K'_{t'\tilde{t}'} \\ K'_{\tilde{t}'t'} & K'_{\tilde{t}'\tilde{t}'} \end{pmatrix},
\]

(4)

where \(\tilde{t}\) and \(\tilde{t}'\) represent a complement of \(t\) and that of \(t'\) respectively.

Next, we estimate the anomaly score which corresponds to \(t\) and \(t'\) as

\[
S_{tt'}(z, z') = \left[ K_M(K, K'_{\tilde{t}'}) + K_M(K', K_{\tilde{t}}) - K_M(K, K_{\tilde{t}}) - K_M(K', K'_{\tilde{t}}) \right]
 - \left[ K_M(K_{\tilde{t}t}, K'_{\tilde{t}'\tilde{t}'}) + K_M(K'_{\tilde{t}'t}, K_{\tilde{t}'\tilde{t}'}) - K_M(K_{\tilde{t}t}, K'_{\tilde{t}'\tilde{t}'}) - K_M(K'_{\tilde{t}'t}, K_{\tilde{t}'\tilde{t}'}) \right],
\]

(5)

where \(K_M\) represents a kernel defined between matrices. As described in Section 3, we introduce the Matrix Kernel that is defined between matrices with different dimensions. Using the Matrix Kernel for \(K_M\), the anomaly score can be estimated even when \(t\) and \(t'\) have different dimensions. We derive the anomaly scores for this case in Section 2.3.

Finally, we conduct anomaly detection and localization by using the anomaly scores. If the system anomaly score is high, then we regard the system as anomalous. If the system anomaly score and target anomaly scores are high, then the variables in the targets are regarded as being responsible for system anomalies.

2.3. Derivation of Anomaly Scores. In this section, we derive the anomaly score defined in Eq. (5).

We define a system anomaly score as a distance between kernel matrices, \(D(K, K')\) as follows:

\[
S(z, z') = D(K, K').
\]

(6)

The distance between kernels represents the amount of change of relations between variables. Therefore this score is a natural measure that represents how anomalous the system is. Next, we generalize the score in Eq. (5) and define a target score as

\[
S_{tt'}(z, z') = D(K, K') - D(K_{\tilde{t}t}, K'_{\tilde{t}'\tilde{t}'}).
\]

(7)

The system anomaly score in Eq. (6) is included in the definition of a target anomaly score in Eq. (7) because \(S(z, z') = S_{tt'}(z, z')|_{t=\tilde{t}=\phi} = D(K, K')\) holds. Here, \(\phi\) represents an empty set.

Using the scores in Eq. (7) and Eq. (8), we can estimate both a system anomaly score and target anomaly scores that includes variable-wise anomaly scores from a single framework. That is, we can estimate an anomaly score for a set with any number of variables, from a single variable to all variables, by using the score in Eq. (7).

As a distance function defined between kernels, we use the symmetrized Burg divergence [12]:

\[
D(K, K') = D_B(K||K') + D_B(K'||K),
\]

(8)

\[
D_B(X||Y) = \text{tr}[XY^{-1}] - \log|XY^{-1}| + m,
\]

(9)
where $D_B$ and $m$ represent Burg divergence and the dimension of matrix $X$ respectively. Then the target anomaly score becomes

$$S_{tw}(z, z') = \text{tr}[KK'^{-1}] + \text{tr}[K'K^{-1}] - \text{tr}[KK'^{-1}] - \text{tr}[K'K^{-1}]$$

$$- \text{tr}[K_{\bar{t}}K_{\bar{t}}'] \text{tr}[K_{\bar{t}}'K_{\bar{t}}'^{-1}] + \text{tr}[K_{\bar{t}}K_{\bar{t}}'^{-1}] + \text{tr}[K_{\bar{t}}'K_{\bar{t}}'^{-1}]$$

(10)

For deriving Eq. (10), we used $m = \text{tr}[XX^{-1}]$ and $\log[XY^{-1}] = \log[YX^{-1}] = 0$. $K_{\bar{t}}^{-1}$ represents an inverse of submatrix $K_{\bar{t}}$. A system anomaly score is derived using $S(z, z') = S_{tw}(z, z')|_{t=\bar{t}=\phi} = D(K, K')$.

We show that the anomaly score in Eq. (10) is a natural measure that represents the change amount between $t$ and $t'$. We denote a feature vector in a kernel space as $\psi$ and denote its $\alpha$-th component as $\psi_\alpha$: $K(z_i, z_j) = \sum_\alpha \psi_\alpha(z_i)\psi_\alpha(z_j)$. By using a vector $w$ that follows a standard normal distribution, we construct variables $\{y_i\}$ and $\{y'_j\}$ from $\{z_i\}$ and $\{z'_j\}$ as

$$w \sim \mathcal{N}(w|\mu=0, \Sigma=I),$$

$$y_i = \sum_\alpha w_\alpha \psi_\alpha(z_i), \quad y'_j = \sum_\alpha w_\alpha \psi_\alpha(z'_j).$$

(11)

(12)

Vectors $y$ and $y'$, of which the $i$-th components are $y_i$ and $y'_j$ respectively, follow multivariate normal distributions as

$$y \sim p(y) = \mathcal{N}(y|0, K), \quad y' \sim p'(y') = \mathcal{N}(y'|0, K').$$

(13)

If one assumes that $t = t'$ and $\bar{t} = \bar{\bar{t}}$ hold, then the following relationship among the distributions of $y$ and $y'$, and the anomaly scores hold:

$$S_{tw}(z, z') = 2E_{p(y_i)} \left[ D_{KL}(p(y_i|y)|p'(y_i|y)) \right] + 2E_{p'(y_i)} \left[ D_{KL}(p'(y_i|y)|p(y_i|y)) \right]$$

$$= 2 \int dy_i p(y_i) \int dy_i p(y_i|y_i) \log \frac{p(y_i|y_i)}{p'(y_i|y_i)} + 2 \int dy_i p'(y_i) \int dy_i p'(y_i|y_i) \log \frac{p'(y_i|y_i)}{p(y_i|y_i)},$$

(14)

where $D_{KL}(p||p')$ represents a KL divergence between probability density functions (pdfs) $p$ and $p'$. The corresponding representation $z = (z_i, z_i)$, we designated $y$ as $y = (y_t, y_t)$. We omit the proof of Eq. (14) because of space limitations, but it would be easy to derive it as the KL divergence in Eq. (10) is analytically tractable for Gaussians. From Eq. (14), the anomaly score of $(t, t')$, Eq. (10), represents the expected KL divergence between conditional probabilities $p(y_i|y_i)$ and $p'(y_i|y_i)$, integrated over the distributions $p(y_i)$ or $p'(y_i)$. Thus, the anomaly score in Eq. (10) is a natural measure that represents the change amount between $t$ and $t'$ because the score has a clear interpretation that it represents the change amount between pdfs in the space of $y$.

The anomaly score in Eq. (10) is definable only when $\text{dim}(t) = \text{dim}(t')$ and $\text{dim}(\bar{t}) = \text{dim}(\bar{\bar{t}})$ hold, where $\text{dim}(t)$ represents a number of variables in a variable set $t$. Below, we eliminate this limitation and generalize the score. The anomaly score is represented as a sum of traces of matrix products. A trace of a matrix product, $\text{tr}[X^TY]$, is equal to an inner product of vectorized matrices: $\text{tr}[X^TY] = \sum_{ij} X^T_{ij} Y_{ji} = \text{vec}(X) \cdot \text{vec}(Y)$. A Kernel function computes a generalized inner product. Therefore, we generalize the score by introducing the following replacement, where $K_M$ is a kernel defined between matrices: $\text{tr}[XY] \rightarrow K_M(X, Y)$. By the replacement, the anomaly score defined in Eq. (5), $S_{tw}(z, z')$, is derived. Using the
kernel introduced in the next section for $K_M$, we can define anomaly scores even when $t = t'$ or $\bar{t} = \bar{t}'$ does not hold. If we use a dot product as $K_M$

$$K_M(X, Y) = \text{vec}(X) \cdot \text{vec}(Y) = \text{tr}[XY],$$

then the anomaly score in Eq. (5) becomes that in Eq. (10) inversely.

3. Matrix Kernel

In this section, we propose Matrix Kernel as $K_M$ included in the anomaly score proposed in Section 2 (the anomaly score is defined as Eq. (5)). The Matrix Kernel is a kernel defined between matrices.

The reason that we introduce the Matrix Kernel is to make DKS robust and applicable to multivariate time series in which the number of elements might change over time.

3.1. Problem Setting. We assume that two real matrices $A$ and $A'$ are given and that they are the $d \times d$ matrix and $d' \times d'$ matrix respectively. We aim to derive a kernel $K_M$ for which inputs are these matrices. Under the problem setting in Section 2, the inputs are restricted to kernel matrices. Therefore, we can consider $A$ and $A'$ as positive semidefinite.

We require that the Matrix Kernel satisfies the following two conditions. The first condition is that input matrices may have different dimensions (i.e. $d \neq d'$ is permitted). This condition is necessary to consider fluctuation of the number of elements (i.e. the number of variables) as mentioned in Section 1. In this case, either $K$ and $K'$ have different dimensions, or $K_{\bar{t}}$ and $K'_{\bar{t}'}$ have different dimensions. Therefore, to estimate the anomaly scores in Eq. (5), it is necessary for $K_M$ to satisfy the first condition.

The second condition is that the Matrix Kernel has permutation invariance. Permutation invariance means that the output of the kernel does not change if we permute the index of the matrix elements of $A$ and $A'$ separately. Assume that $A$ and $A'$ are respectively transformed as $A_p$ and $A'_{p'}$ by matrix element index permutation. The condition requires $K_M(A, A') = K_M(A_p, A'_{p'})$ holds for any pair of such permutations. Here $p$ and $p'$ may be different. The second condition is necessary to make DKS robust. If $K_M$ in Eq. (5) satisfies this condition, whereas the anomaly scores are invariant under non-essential changes because of the permutation (i.e. non-topological changes of kernel matrices because they can be transformed by the permutation), they are sensitive to topological changes of kernel matrices.

Hereinafter we designate these two conditions as matrix kernel conditions.

3.2. Representation of Matrix Kernel. Suppose that the input matrices are decomposed as

$$A = \sum_{k=1}^{d} u_k \lambda_k u_k^T, \quad A' = \sum_{l=1}^{d'} u'_l \lambda'_l u'_l^T,$$

(16)

For example, $A$ is transformed as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \rightarrow A_p = \begin{pmatrix} A_{22} & A_{21} \\ A_{12} & A_{11} \end{pmatrix},$$

where $p$ represents a permutation of elements “1” and “2”.

\[1\text{For example, } A \text{ is transformed as} \]

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \rightarrow A_p = \begin{pmatrix} A_{22} & A_{21} \\ A_{12} & A_{11} \end{pmatrix},$$

where $p$ represents a permutation of elements “1” and “2”.

\[1\text{For example, } A \text{ is transformed as} \]
where \( \lambda_k \) and \( u_k \) (\( \lambda'_k \) and \( u'_k \)) represent the \( k \)-th eigenvalue and the \( k \)-th eigenvector of the input matrix \( A \) (\( A' \)) respectively. We assume that the eigenvalue decompositions in Eq. (16) become unique if we require additional conditions such as “\( u_k^T u_k = 1 \)”, “\( \sum_{i=1}^d (u_k)_i \geq 0 \)”, and “If \( \lambda_k = \lambda_{k+1} \), then we replace \( u_k \) with \( v_k = c_k u_k + c_{k+1} u_{k+1} \), where \( v_k \) maximizes \( \sum_{i=1}^d |(v_k)_i| \)”.

Here \( (u_k)_i \) represents the \( i \)-th component of \( u_k \).

We define the Matrix Kernel by Eq. (19) below. For experimental results in Section 4, we used the following Matrix Kernel derived for the univariate normal distribution:

\[
K_M(A, A') = \sum_{k=1}^d \sum_{l=1}^{d'} \lambda_k \lambda'_l \left( \frac{2\sigma(u_k)\sigma(u'_l)}{\sigma(u_k)^2 + \sigma(u'_l)^2} \right) \exp \left[ -\frac{(\mu(u_k) - \mu(u'_l))^2}{2(\sigma(u_k)^2 + \sigma(u'_l)^2)} \right],
\]

where \( \sigma(x) \) and \( \mu(x) \) represent the standard deviation of vector \( x \’s \) components and the mean of the components respectively.

### 3.3. Derivation of Matrix Kernel.

In this section, we derive the kernel in Eq. (17).

If the dimensions of \( A \) and \( A' \) (\( d \) and \( d' \) in Eq. (16)) are the same, then we can define the following kernel function represented as an inner product:

\[
K_I(A, A') = \sum_{i,j=1}^d A_{ij} \cdot A'_{ij} = \sum_{k,l=1}^d \lambda_k \lambda'_l \left[ \left( u_k^T u'_l \right)^2 \right],
\]

where \( (u_k)_i \) represents the \( i \)-th component of \( u_k \). \( K_I \) in Eq. (18) is a kernel function defined between matrices. However, \( K_I \) does not satisfy the matrix kernel conditions described in Section 3.1 because \( K_I \) is not permutation invariant and is definable only when \( d = d' \) holds.

By generalizing the inner product in Eq. (18), we define a Matrix Kernel as

\[
K_M(A, A') = \sum_{k=1}^d \sum_{l=1}^{d'} K_s(\lambda_k, \lambda'_l) \left[ K_v(u_k, u'_l) \right]^2.
\]

Eq. (19) is derived from Eq. (18) by replacing a scalar product with a kernel \( K_s \) defined between scalars and a vector inner product with a kernel \( K_v \) defined between vectors. Note that \( K_M \) in Eq. (19) is a kernel defined between matrices because (1) its inputs are matrices and (2) it is represented as a sum of kernel products with positive coefficients (\( \geq 1 \)). \( K_M \) satisfies the matrix kernel conditions if both \( K_s \) and \( K_v \) satisfy the conditions.

As a kernel \( K_s \), we use the following representation:

\[
K_s(\lambda_k, \lambda'_l) = \lambda_k \lambda'_l.
\]

\( K_s \) in Eq. (20) satisfies the matrix kernel conditions for the following reasons. First, \( \lambda_k \lambda'_l \) can be defined independently of the range of indices \( k \) and \( l \). Second, \( K_s \) has the permutation invariance because eigenvalues are permutation invariant.2

---

2The eigenvalues of matrix \( A \) are invariant under an orthogonal transformation \( U \), as we demonstrate below.

\[
Au_k = \lambda_k u_k
\]

(\( UA^TU^T \))(\( Uu_k \)) = \( \lambda_k (Uu_k) \)

Eq. (22) is derived by multiplying \( U \) from the left of the eigenequation of \( A \) in Eq. (21). Therefore, they are equal. It is apparent that matrix \( A \) and eigenvector \( u_k \) are transformed as \( A \rightarrow UA^TU^T \) and \( u_k \rightarrow Uu_k \) under
Next, we construct $K_v$ such that it satisfies the matrix kernel conditions. We denote a pdf of $u$'s ($u'$'s) component as $p(x|u)$ ($p(x|u')$), where $x$ represents a scalar. Such a pdf has the following two properties. First, the pdf $p(x|u)$ can be regarded as an infinite dimensional vector independently of the dimension of $u$ because $p(x|u)$ is a univariate function and because $x$ can be regarded as a vector component index. Second, $p(x|u)$ is invariant under a permutation of vector component index of $u$ because a pdf of $u$'s component is independent of the order of the components.

Using these properties, we construct $K_v$ as

$$ K_v(u, u') = \int_{-\infty}^{\infty} dx \sqrt{p(x|u)p(x|u')} . \tag{23} $$

Eq. (23) is derived from $K_v = u^T u' = \sum_i u_i u'_i$ by replacing $u_i, u'_i$ and $\sum_i$ respectively with $\sqrt{p(x|u)}, \sqrt{p(x|u')}$ and $\int_{-\infty}^{\infty} dx$. $K_v$ in Eq. (23) has the following properties. First, $K_v$ is a kernel function because Eq. (23) is an inner product of infinite dimensional vectors. Second, $K_v$ satisfies the matrix kernel conditions because 1) $K_v$ is definable between vectors even when $u$ and $u'$ have different dimensions, and 2) $K_v$ is permutation invariant because $p(x|u)$ and $p(x|u')$ are permutation invariant. As a pdf, we use a univariate normal distribution $\mathcal{N}$:

$$ p(x|u) = \frac{1}{\sqrt{2\pi \sigma(u)}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu(u)}{\sigma(u)} \right)^2 \right] , \tag{24} $$

where $\sigma(u)$ and $\mu(u)$ represent the standard deviation of vector $u$'s components and the mean of the components respectively.

By substituting Eqs. (20), (23) and (24) to Eq. (19), the representation of the Matrix Kernel in Eq. (17) is derived. The integration in Eq. (23) is analytically tractable because the pdfs in Eq. (23) are Gaussians. Therefore Eq. (17) is derived analytically.

4. Experiments

Here, we examined the utility of DKS using three multivariate datasets. We show that DKS can detect 1) element anomalies more accurately than an existing method (Section 4.1), 2) changes of the numbers of elements as anomalies using the Matrix Kernel (Section 4.2), and 3) system anomalies and elements responsible for the anomalies simultaneously (Section 4.3). No standard method exists for detecting system anomalies and element anomalies simultaneously from a single framework. Therefore, in Section 4.3 we did not compare DKS with any other method.

4.1. Single Variable Anomaly Scoring.
4.1.1. **Experimental Setting.** In this experiment, we conducted anomaly scoring for a single variable included in a multivariate time series. Then, we compared the results obtained with DKS with those obtained from an existing method.

We used *Synthetic Control Chart Time Series dataset* \[15, 16\]. We used 60 “normal” mode time series, each of which consisted of 100 time steps, and replaced a set of their \( t = 51, \ldots, 100 \) data points with that of “cyclic” mode time series randomly with probability 1/3. We considered \( t = 1, \ldots, 50 \) data points (60 time series \( \times \) 50 time steps) as \( D \) in Eq. (1) and \( t = 51, \ldots, 100 \) data points as \( D' \) in Eq. (1), respectively. Then we estimated the anomaly scores of the individual time series. We considered the replacement (change) as the anomaly to be detected.

As a kernel defined between variables, we used a covariance matrix and a diffusion kernel \[14\] described below. We defined a diffusion kernel \( K \) as follows: \( C_{ij} = (\text{correlation between } z_i \text{ and } z_j) \), \( L_{ij} = \| C_{ik} \| \delta_{ij} - |C_{ij}|, K = \exp \{-\lambda L\} \). Here \( \delta_{ij} \) represents Kronecker’s delta. It becomes 1 if \( i = j \) holds, and 0 otherwise. Matrices \( C, L, \) and \( K \) represent a correlation matrix, a graph Laplacian, and a diffusion kernel matrix respectively. We fixed parameter \( \lambda \) as \( \lambda = 1.0 \).

As a kernel defined between matrices, we used the Matrix Kernel in Eq. (17), and a dot product in Eq. (15).

We compared the results obtained with DKS with those obtained using Sparse Structure Learning (SSL)\[3\]. SSL (1) estimates sparse precision matrices (inverse of covariance matrices) of a pair of multivariate time series, and (2) estimates the anomaly scores of the individual time series using the sparse (and therefore robust) structure. SSL is not a kernel method. However, SSL estimates covariance matrices and uses a dot product (trace) of them. Therefore, it can be said that SSL corresponds to a DKS where a covariance and a dot product are used as a kernel between variables and that between matrices respectively. We set the free parameter of SSL as \( \rho = 0.7 \), which represented a ratio of graphical lasso\[17\]’s penalty and derived the best experimental results in a study reported in Ref. [3].

As an estimation measure, we used the area under the curve (AUC) of the ROC curve. AUC becomes high when the time series with the replacement have higher scores and those without the replacement have lower scores. The random replacement was iterated 100 times. We then evaluated the two methods, SSL and DKS, using AUC.

4.1.2. **Experimental Results.** The experimentally obtained results are presented in Table 1. From the table, it is readily apparent that DKS is effective for anomaly detection for a single variable included in a multivariate time series for the following reasons.

First, DKS with a diffusion kernel or a Matrix Kernel outperformed SSL. It is one of DKS’s own features that we can select kernels between variables and those between matrices, although we cannot in the case of SSL. Therefore it was not possible to adjust SSL to outperform DKS. For real-world applications, it is important to select optimal kernels for DKS. Selection methods include a cross validation. However, to construct a kernel selection method is beyond the scope of this paper.

Second, DKS with a covariance matrix and a dot product was comparable to SSL. Here, comparable means that their AUCs were inferred as the same based on the results of the \( t \)-test with 95% confidence. This result indicates that the anomaly score defined in Eq. (5) is valid for anomaly detection even in a linear space, which SSL considers.
Table 1. Experimental results of Section 4.1, single variable anomaly scoring. “kernel(variables)” and “kernel(matrices)” represent a kernel defined between variables and a kernel defined between matrices, which were used in the method respectively. “Covariance”, “Diffusion”, “DP”, “PPK”, and “MATRIX” represent the covariance matrix, diffusion kernel, dot product, probability product kernel, and the Matrix Kernel respectively. AUC is represented as (mean)±(standard deviation). The best result is presented in bold typeface.

| method | kernel(variables) | kernel(matrices) | AUC     |
|--------|------------------|------------------|---------|
| SSL    | Covariance       | DP               | 0.685±0.120 |
| DKS    | Covariance       | DP               | 0.659±0.113 |
| DKS    | Covariance       | MATRIX           | 0.583±0.111 |
| DKS    | Diffusion        | DP               | 0.865±0.079 |
| DKS    | Diffusion        | MATRIX           | 0.938±0.064 |

4.2. Change Detection of Constituent Variables.

4.2.1. Experimental Setting. In this experiment, we used DKS for detecting anomalies where the numbers of variables change, to compare the Matrix Kernel with an existing kernel.

Two multivariate datasets were generated randomly in which each of the variables followed a standard normal distribution and consisted of 200 observations. We designated the datasets as \( D \) in Eq. (1) and \( D' \) in Eq. (1). \( D \) and \( D' \) consisted of 9 and 10 variables respectively.

Under two settings presented in Table 2, we used DKS to estimate anomaly scores for variable groups (“Group”s in Table 2). These settings had the same datasets (\( D \) with 9 variables and \( D' \) with 10 variables) and different group assignments. This led to change in the number of variables in Group 9 of Setting 1 and that in Group 10 of Setting 2 changed.

For example, the anomaly score for Group 9 in the case of Setting 1 (see Table 2) was estimated by substituting \( t = \{z_9\} \) and \( t' = \{z_9, z_{10}\} \) to Eq. (3).

Table 2. Left: Setting 1 of Section 4.2, change detection of constituent variables, where variable \( z_{10} \) was newly generated in variable group 9. Right: Setting 2 of Section 4.2, where variable group 10 was newly generated.

| Group | \( D \)  | \( D' \) |
|-------|---------|---------|
| 1     | \( z_1 \) | \( z_1 \) |
| \vdots | \vdots | \vdots |
| 8     | \( z_8 \) | \( z_8 \) |
| 9     | \( z_9 \) | \( z_9 \) |
| 10    | None    | \( z_{10} \) |

As a kernel between variables, we used a covariance matrix. To compare a kernel matrix with an existing kernel between matrices, we used the Matrix Kernel and the Probability Product Kernel (PPK) as kernels between kernels (i.e. kernels between matrices). As
mentioned in Section 4.2, PPK can take different dimensional matrices as inputs. We iterated the random data generation 100 times to estimate the anomaly scores.

Figure 1. Experimental results of Section 4.2 change detection of constituent variables. Horizontal and vertical lines represent the variable group and the anomaly score of the variable group respectively. A shown point and an error bar represent the mean and the standard deviation (±1σ) respectively.

4.2.2. Experimental Results. The experimentally obtained results are presented in Figure 1. As the figures show, it is readily apparent that DKS is effective for systems in which the numbers of variables change for the following reasons.

As shown in the figure, we were able to estimate anomaly scores for datasets with different numbers of variables, using DKS. DKS with the Matrix Kernel was able to detect the changes of assignments as anomalies successfully. The anomaly scores for Group 9 in Setting 1 and Group 10 in Setting 2 had sufficiently higher than those for the other groups. The results also indicate that the Matrix Kernel is more effective than the PPK. This is because DKS with the Matrix Kernel detected changes of the assignments as anomalies whereas DKS with the PPK failed to detect such changes.

4.3. Anomaly Detection and Localization in Economic Time Series.

4.3.1. Experimental Setting. DKS was applied for economic time series in this experiment. Namely, we conducted anomaly scoring for a single variable in the time series and the entire system simultaneously by using DKS.

We used twenty economic time series consisting of nine FX (foreign exchange) time series and eleven stock index time series as input. These FX time series include USDAUD, USDBRL, USDCAD, USDEUR, USDGBP, USDKRW, USDJPY, USDRUB. The FX time series took values per 1 USD, whereas the stock index time series consist of AORD, BVSP, CAC40, DAX, DJI, FTSE100, Hang Seng, KOSPI composite, N225, RTSI, and TSX composite. The time series were observed once a day from 1st January 2004 to 31st December 2008.
We decided to use the economic dataset for the following reasons. First, it includes the great depression starting in September 2008, which was triggered by the bankruptcy of Lehman Brothers. We suspected that DKS would detect this economic disorder as the anomaly. Second, it is expected that anomalies appear in response to changes in the relationships between some variables (some time series), as currencies and stocks strongly correlate with each other.

We applied DKS to the dataset with the following settings. We set time window width as 50 days, where the $n$-th window consisted of the time series from $t = n - 49$ [day] to $t = n$ [day]. By comparing the $n$-th and the $(n - 1)$-th windows, we estimated the anomaly scores of the system and the individual variables. We designated the scores as “scores at $t = n$”. We used a correlation matrix and a Matrix Kernel respectively as a kernel between variables and that between matrices.

4.3.2. Experimental Results. There were several peaks found in the system anomaly score time series (Figure 2). One of the peaks appeared on 8th September 2008, which was a week before Lehman Brothers announced its bankruptcy. The system anomaly score remained relatively high from November 2007 to October 2008 when the subprime mortgage crisis were ongoing from 2007. These results suggest that DKS successfully detected changes in the relationship among economic time series caused by the economic disorders as anomalies.

The variable-wise anomaly scores on 8th September 2008 are shown in Figure 3. The anomaly score of USDRUB was the highest among the variable-wise scores. USDRUB continued to decrease stably around the day of the bankruptcy while the other stocks and the currencies fluctuated significantly. This difference in trend could produce the high anomaly score of USDRUB.

Here, we observed the behavior of DKS only qualitatively because we could not know what the anomalies to be detected were. However, the experimental results suggest that DKS is simultaneously applicable to change detection (anomaly detection in a sense) by using the system anomaly score, and localization by using the variable-wise anomaly scores.
Figure 3. Experimental results of Section 4.3 on anomaly localization from the economic time series. Anomaly scores of the individual time series on 8th September 2008, a week before the bankruptcy of Lehman Brothers. The horizontal line represents variable-wise anomaly score.

5. Summary

We have developed the new anomaly detection method, Double Kernelized Scoring (DKS). This is a unified method to perform anomaly detection and localization simultaneously in a strongly correlating system with a changing number of elements. For comparing matrices with different dimensions, we have proposed a new kernel function, Matrix Kernel. The Matrix Kernel is defined between square matrices that might have different dimensions. We have demonstrated the effectiveness of DKS and Matrix Kernel through the experimental results using three datasets.

Acknowledgements

We would like to thank Dr. Makoto Fukushima for fruitful discussions and comments on an earlier version of the manuscript.

References

[1] Yamanishi, K. and Takeuchi, J. A unifying framework for detecting outliers and change points from non-stationary time series data. In Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002).

[2] Siffer, A., Fouque, P. A., Termier, A., and Largouet, C. Anomaly detection in streams with extreme value theory. In Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017).

[3] Idé, T., Lozano, A. C., Abe, N., and Liu, Y. Proximity-based anomaly detection using sparse structure learning. In Proc. of the SIAM International Conference on Data Mining (2009).

[4] Montazpour, M., Zhang, J., Rahman, S., Sharma, R., and Ramakrishnan, N. Analyzing Invariants in Cyber-Physical Systems using Latent Factor Regression. In Proc. of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015).
[5] Idé, T., and Kashima, H. Eigenspace-based anomaly detection in computer systems. In Proc. of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004).

[6] Hirose, S., Yamanishi, K., Nakata, T., and Fujimaki, R. Network anomaly detection based on eigenequation compression. In Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2009).

[7] Hara, S., Morimura, T., Takahashi, T., Yanagisawa, H., and Suzuki, T. A Consistent Method for Graph Based Anomaly Localization. In Proc. of the 18th International Conference on Artificial Intelligence and Statistics (2015).

[8] Manzoor, E., Milajerdi, S. and Akoglu L. Fast memory-efficient anomaly detection in streaming heterogeneous graphs. In Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).

[9] Haussler, D. Convolution kernels on discrete structures. Technical Report UCSCRL-99-10, University of California at Santa Cruz (1999).

[10] Kashima, H., Tsuda, K., and Inokuchi, A. Marginalized kernels between labeled graphs. In Proc. of the 20th International Conference on Machine Learning (2003).

[11] Kondor, R., Shervashidze, N., and Borgwardt, K. M. The graphlet spectrum. In Proc. of the 26th International Conference on Machine Learning (2009).

[12] Kulis, B., Sustik, M. A., and Dhillon, I. S. Learning low-rank kernel matrices. In Proc. of the 23rd International Conference on Machine Learning (2006).

[13] Jebara, T., Kondor, R., and Howard, A. Probability product kernels. The Journal of Machine Learning Research, 5, pp. 819–844 (2004).

[14] Kondor, R., and Lafferty, J. Diffusion kernels on graphs and other discrete structures. In Proc. of the 19th International Conference on Machine Learning (2002).

[15] Pham, D. T., and Chan, A. B. Control chart pattern recognition using a new type of self-organizing neural network. In Proc. of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 212(2), pp. 115–127 (1998).

[16] Bache, K. and Lichman, M. UCI Machine Learning Repository (2013). URL http://archive.ics.uci.edu/ml

[17] Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3), pp. 432–441 (2008).