Scalable Change Analysis and Representation Using Characteristic Function

Takaaki Tagawa[1], Yukihiro Tadokoro[2], and Takehisa Yairi[3]

[1,2] *Toyota Central R&D Labs., Inc., Aichi, Japan.*
tagawa@mosk.tytlabs.co.jp
y.tadokoro@ieee.org

[3] *Research Center for Advanced Science and Technology, University of Tokyo, Tokyo, Japan.*
yairi@ailab.t.u-tokyo.ac.jp

ABSTRACT

In this paper, we propose a novel framework to help human operators—who are domain experts but not necessarily familiar with statistics— analyze a complex system and find unknown changes and causes. Despite the prevalence, researchers have rarely tackled this problem. Our framework focuses on the representation and explanation of changes occurring between two datasets, specifically the normal data and data with the observed changes. We employ two-dimensional scatter plots which can provide comprehensive representation without requiring statistical knowledge. This helps a human operator to intuitively understand the change and the cause. An analysis to find two-attribute pairs whose scatter plots well explain the change does not require high computational complexity owing to the novel characteristic function-based approach. Although a hyper-parameter needs to be determined, our analysis introduces a novel appropriate prior distribution to determine the proper hyper-parameter automatically. The experimental results show that our method presents the change and the cause with the same accuracy as that of the state-of-the-art kernel hypothesis testing approaches, while reducing the computational costs by almost 99% at the maximum for all popular benchmark datasets. The experiment using real vehicle driving data demonstrates the practicality of our framework.

1. INTRODUCTION

One of the important roles in prognostics and health management is to ensure the correct operation of an artificial system (such as computer network, power plant, and vehicle) by monitoring its state. Thus, when unknown and/or unexpected events are observed, it is necessary to apply countermeasures as soon as possible to avoid critical scenarios. Usually, the system relies on human operators—who are the domain experts of the system—to handle such situations. They often investigate data to specify the part (position) and analyze the cause of the events. However, any analysis to uncover the cause by human operators is expensive owing to the recent progress in system technologies that have facilitated the building of large-scale systems. A similar problem is observed in several other fields, such as vehicle systems, industrial plants, satellite systems, and Internet networks(Chandola, Banerjee, & Kumar, 2009).

Most existing works (Chandola et al., 2009) focus on detecting the existence of unknown events that include anomalies, faults, frauds, and intrusions, where the detection is insufficient for operators to identify the cause of a large-scale system to take countermeasures. On the contrary, we consider a different situation herein; a change is known to have happened, but the part and the cause are unclear. One example is evident in testing new vehicle systems; in practice, test drivers often evaluate the developing system by driving the vehicle. As driving experts, they can distinguish whether changes have occurred within the system through sensory analysis (Figure 1, left). However, it is yet difficult and time consuming to specify the part and the cause of the change from a large and complex vehicle system. Thus, a method that automatically analyzes the system to quickly visualize the change and help users to identify the cause (Figure 1, right) is required. This scenario is widely applicable in cases where human experts or conventional prognostics and health monitoring systems can detect unknown changes but cannot identify their causes. Herein, we consider a situation where two datasets, namely, normal data and data with the observed changes, are given from a system. The goal here is to develop a method that identifies and visualizes the changes between the two datasets.

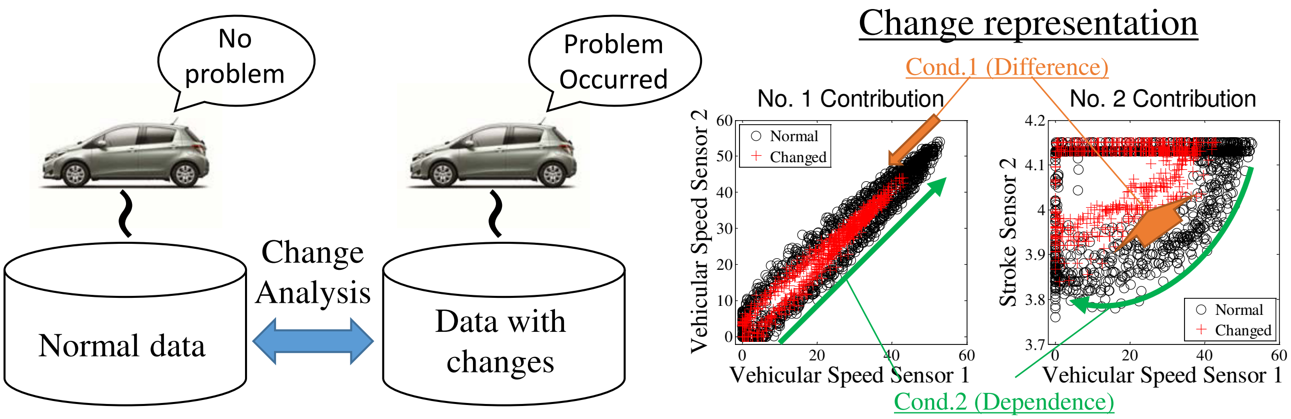

Figure 1. Extract two-variable scatter plots that pinpoint changes between datasets.

Although some existing approaches, such as that mentioned in the works of He et al., Hido et al., and Joe (He, Yang, Chen, & Zhang, 2012), (Hido, Idé, Kashima, Kubo, & Matsuzawa, 2008), and (Joe Qin, 2003), can indicate certain data attributes as causes, the approaches do not explain why the indicated attributes are the causes. Therefore, in addition to pointing out the cause of changes, its interpretable representation is also important. Conventionally, data visualization methods, such as principal component analysis, RadViz (Grinstein, Trutschl, & Cvek, 2001), parallel coordinate plot (Grinstein et al., 2001), and scatter matrix, are used to represent changes between datasets. However, if it is necessary to represent high-dimensional data, e.g., data with more than 100 features; users who are unfamiliar with statistics may find it difficult to interpret. VizRank (Leban, Zupan, Vidmar, & Bratko, 2006) tackles this issue by selecting a small subset of features that well express the changes. VizRank measures the changes by the statistic of the differences between datasets using the k-nearest neighbor (k-NN) approach. A problem is that k-NN suffers from the curse of dimensionality (Marimont & Shapiro, 1979) and requires high computational costs, as well as properly setting a hyper-parameter. In addition, VizRank relies on some heuristics when selecting a small subset of features.

The first contribution of this paper is a novel framework to analyze the cause of changes between two datasets with an interpretable representation. The changes are assumed as variations in data distributions between two datasets with a certain subset of features. Hence, the goal is to find such subsets as causes and represent the variations to operators. Two-dimensional scatter plots are employed to show the arbitrary changes in correlations using only given data attributes and samples that are understandable for operators. Our framework aims to *rank* the two-attribute pairs by how well they represent variations in data distributions between two datasets. The scatter plots of the above ranked pairs well represent the cause and give operators reliable reasons to conduct a detailed analysis of these specified parts (Figure 1).

The second contribution of this study is the proposal to establish the characteristic function-based approach to our framework. To be useful to operators in practice, our approach is the first to be able to report reliable results speedily without requiring any parameter tunings. Fast computations are achieved by using a characteristic function. As the characteristic function requires setting hyper-parameters to attain a good estimate of the change and the cause, we analyze optimal conditions for our setting and propose an appropriate prior distribution. The hyper-parameters are determined automatically with the prior distribution to achieve the same performance as that of the state-of-the-art kernel methods while reducing computational costs significantly.

Experimental results with popular benchmark datasets showed that our framework with the characteristic function can reduce the computation time by 99% at the maximum compared to state-of-the-art kernel-based methods, while maintaining its ability to analyze the changes. A practical experiment using vehicle driving data demonstrates how well our framework can support operators to analyze changes, as well as the representation power of scatter plots.

In Section 2, we introduce the framework for the change representation. The benchmark methods applicable to our framework are given in Section 3, along with our proposed approach in Section 4, with its analysis in Section 5. Experimental results follow in the next section to validate the advantage of our method and we conclude our discussion in the final section.

2. NEW FRAMEWORK FOR CHANGE REPRESENTATION

To present the intuitive representation of the change and the cause, we employ two-dimensional scatter plots. Using scatter plots, the proposed framework shows the top n two-variable pairs that express the change sufficiently well. This information is useful for operators to find the dominant factor and significantly reduce the cost of checking entire plots, especially for a high-dimensional dataset.

In the following subsections, we detail how to rank two-variable pairs correctly to extract the cause of changes.

2.1. Ranking with Two Statistics

Let $\mathbf{X} = \{X_1, \ldots, X_d\} \in \mathbb{R}^d$ and $\mathbf{Y} = \{Y_1, \ldots, Y_d\} \in \mathbb{R}^d$ be d-dimensional random variables, that represent distributions of normal data and data with the observed changes. Each variable represents a sample distribution of continuous or discrete values, such as vehicle speed sensors readings or shift positions, with which human operators are familiar. We use different variable names X and Y for convenience; however, they are generated from the same system. The problem we should solve is: How can we rank each (k, l)th two-variable pair, $k, l \in 1, \ldots, d$, with respect to how well it exhibits the change? To deal with the problem, we assume that the pair that satisfactorily displays the change should satisfy the following conditions.

Cond. 1 (Difference): A pair of variables, one from each dataset, exhibit a large difference;

Cond. 2 (Dependence): The two variables have a strong correlation with the normal data.

Cond. 1 is obvious, as any change must manifest itself as a difference in the datasets. This condition is measured by the statistics related to two-sample hypothesis testing (Lehmann & Romano, 2005), for instance, contingency-based approaches, Kolmogorov–Smirnov testing (Smirnov, 1948), Hotelling's T^2 (Hotelling, 1992), and kernel-based two-sample testing (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2012). Thus, let $M(k, l)$ be the difference statistic, which has a large value if the difference increases, and let $R_M(k, l)$ be the corresponding rank ordered by $M(k, l)$.

Cond. 2 is also important because variables that have strong relations in normal data tend to be a substantial part of the system as compared with the variables that have only trivial relations. The corruption of this correlation is critical, and its analysis is a priority. For example, with vehicle data, strong correlations represent some mechanical relations such as a response from an accelerator position to an engine rotation. The unknown change of such relations can be critical and should be extracted beforehand. One can also expect situations where some uncorrelated pairs will become correlated. Although such pairs are often not significant parts of the system and their priority is low, we can also extract such changes by switching the datasets to be analyzed, i.e., normal data to changed data and changed data to normal data. Cond. 2 is measured by evaluating the dependence (or independence) between two variables in independence-testing approaches (Lehmann & Romano, 2005), such as, contingency-based approaches, independence component analysis related approaches (Hyvärinen & Oja, 2000), and kernel-based independence testing (Gretton et al., 2008). Therefore, let $H(k, l)$ be the dependence statistic that has a large value if the dependence between the (k, l)th variable is strong, and let $R_H(k, l)$ be the corresponding rank ordered by $H(k, l)$.

Based on the rankings $R_M(k, l)$ and $R_H(k, l)$, we want to obtain the pair (k, l) that satisfies both Conds. 1 and 2. Thus, an average value is applied for an integrated measure, as given by

$$F(k, l) = \frac{R_M(k, l) + R_H(k, l)}{2}. \tag{1}$$

Finally, a representation power of (k, l)th two-variable pair is ranked by $F(k, l)$ to obtain the order $R_F(k, l)$. The datasets were plotted with respect to $R_F(k, l)$ to represent the effect of changes for the operators.

2.2. Representation with Scatter Plot

Given the final ranking $R_F(k, l)$, we now represent the changes using a scatter plot, which is a useful tool to depict changes between two datasets without requiring statistical knowledge. We extract the above-mentioned n variable sets (k, l) in the order determined by $R_F(k, l)$ and two datasets were plotted on each scatter plot with axis attributes, which human operators know well. If the contributions are high, certain relationships should be observed in the normal dataset (**Dependence** condition) and the relations are changed between the datasets (**Difference** condition) for each (k, l)th variable pair. Operators can easily understand where and how the changes are occurring inside the system by referring to these plots (Figure 1). In addition, since the plots are based on axes, which are not normalized nor rescaled, human operators can easily evaluate whether the change is really happening or not, i.e., false positive or otherwise, based on their domain experiences.

The accuracies of ranking with respect to the measured differences $M(k, l)$ and the measured dependencies $H(k, l)$ are essential to provide a high-quality representation. Kernel-based statistics (Gretton et al., 2012)(Gretton et al., 2008) are powerful and valid tools for our framework. The statistics are based on non-parametric approaches and a wide range of differences and dependencies are measurable. The theoretical backgrounds have been investigated intensively and are reliable. However, computational costs are high when a large dataset is involved, and hyper-parameters must be set. We consider the kernel-based statistics as a standard method; our method based on the characteristic functions overcomes these problems while retaining performance. The following section gives a summary of the kernel statistics for difference and dependence.

3. STATISTICS WITH EXISTING KERNEL BASED APPROACH

We compared our method with the kernel statistical testing approaches that deal with difference and dependence in the reproducing kernel Hilbert space (RKHS). These methods provide state-of-the-art performances for evaluating both difference and independence (Gretton et al., 2012)(Gretton et al., 2008), but their computational costs increase quadratically with the number of samples. A summary of kernel statistics is given in the following subsections.

3.1. Difference Statistic with Characteristic Kernel

Gretton et al. (Gretton et al., 2012) proposed the kernel-based two-sample statistic, which is the maximum mean discrepancy in RKHS, to evaluate the difference between two data distributions. Let $\mathbf{X}_{kl} = (X_k, X_l)$ and $\mathbf{Y}_{kl} = (Y_k, Y_l)$ be the kth and lth variables of $\mathbf{X}$ and $\mathbf{Y}$. Given RKHS $\mathcal{H}_2$, let $\psi_2 : \mathbb{R}^2 \to \mathcal{H}_2$, the feature function whose inner product is given by the kernel function $K_2 : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$. We can obtain the kernel mean embedding $\mu : \mathbb{R}^2 \to \mathcal{H}_2$ as

$$\mu_{\mathbf{X}_{kl}} := \int_{\mathbf{X}_{kl} \in \mathbb{R}^2} \psi_2(\mathbf{X}_{kl}) \mathrm{Pr}(d\mathbf{X}_{kl}).$$

The corresponding empirical estimate is obtained as follows:

$$\hat{\mu}_{\mathbf{X}_{kl}} = \frac{1}{N_{\mathbf{x}}} \sum_{m=1}^{N_{\mathbf{x}}} \psi_2(\mathbf{x}_{kl}^{(m)}), \tag{2}$$

where $\mathbf{x}_{kl}^{(m)}$ is the mth sample of $\mathbf{X}_{kl}$. $\mu_{\mathbf{Y}_{kl}}$ and $\hat{\mu}_{\mathbf{Y}_{kl}}$ are obtained in the same manner. μ represents the mean in $\mathcal{H}_2$ and contains higher-order information that characterizes the random variables. Here, we assume that the kernel is characteristic, and then μ is injective, i.e., $\mu_{\mathbf{X}_{kl}} = \mu_{\mathbf{Y}_{kl}}$ holds if $\mathrm{Pr}(\mathbf{X}_{kl}) = \mathrm{Pr}(\mathbf{Y}_{kl})$. Therefore, the difference statistic $M(k, l)$ is given by

$$\begin{aligned} M(k,l) &= \|\hat{\mu}_X - \hat{\mu}_Y\|^2 \\ &= \frac{1}{N_{\mathbf{x}}^2} \sum_{m=1,n=1}^{N_{\mathbf{x}}} K_2(\mathbf{x}_{kl}^{(m)}, \mathbf{x}_{kl}^{(n)}) + \frac{1}{N_{\mathbf{y}}^2} \sum_{m=1,n=1}^{N_{\mathbf{y}}} K_2(\mathbf{y}_{kl}^{(m)}, \mathbf{y}_{kl}^{(n)}) \\ &\quad - \frac{2}{N_{\mathbf{x}} N_{\mathbf{y}}} \sum_{m=1}^{N_{\mathbf{x}}} \sum_{n=1}^{N_{\mathbf{y}}} K_2(\mathbf{x}_{kl}^{(m)}, \mathbf{y}_{kl}^{(n)}). \end{aligned}$$

This statistic can be used to evaluate the difference between two distributions. As the characteristic kernel, we apply the Gaussian kernel given by

$$K_2(\mathbf{x}_{kl}, \mathbf{y}_{kl}) = \exp\left(-\frac{\|\mathbf{x}_{kl} - \mathbf{y}_{kl}\|^2}{2h^2}\right), \tag{3}$$

where h is the hyper-parameter and should be determined carefully to obtain a good result.

3.2. Dependence Statistic with Characteristic Kernel

Gretton et al. (Gretton et al., 2008) proposed a method using the Hilbert–Schmidt independence criterion (HSIC) as a statistic to measure the independence between two variables. Given RKHS $\mathcal{H}_1$, let $\psi_1 : \mathbb{R} \to \mathcal{H}_1$ with the corresponding kernel $K_1 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. HSIC is derived from the covariance operator $\Sigma_{\mathbf{X}_{lk}} : \mathcal{H}_1 \otimes \mathcal{H}_1$, where $\otimes$ is the tensor product, given by

$$\Sigma_{\mathbf{X}_{lk}} = \mu_{X_k, X_l} - \mu_{X_k} \mu_{X_l}, \tag{4}$$

where μ_{X_k, X_l} is the kernel mean embedding of (X_k, X_l) on $\mathcal{H}_1 \otimes \mathcal{H}_1$. $\Sigma_{\mathbf{X}_{lk}}$ evaluates the dependence in RKHS and contains higher-order information. If the kernels are characteristic, $X_k \perp X_l$ holds when $\Sigma_{\mathbf{X}_{lk}}$ is zero. Therefore, the independent statistic $H(k, l)$ is given by

$$H(k,l) = \|\Sigma_{\mathbf{X}_{lk}}\|_{HS}^2 \approx \frac{1}{N_{\mathbf{x}}^2} \mathrm{Tr}[\mathbf{K}_{x_k} \mathbf{Q} \mathbf{K}_{x_l} \mathbf{Q}], \tag{5}$$

where Tr[·] is the trace, $\|\cdot\|_{HS}$ is the Hilbert–Schmidt operator, $\mathbf{K}_{X_k}$ and $\mathbf{K}_{X_l}$ are $N_{\mathbf{x}} \times N_{\mathbf{x}}$ matrices with (m, n)th entries such that $K_1(x_k^{(m)}, x_k^{(n)})$ and $K_1(x_l^{(m)}, x_l^{(n)})$, $\mathbf{Q} = \mathbf{I}_{N_{\mathbf{x}}} - \frac{1}{N_{\mathbf{x}}} \mathbf{J}_{N_{\mathbf{x}}}$, $\mathbf{I}_{N_{\mathbf{x}}}$ is an $N_{\mathbf{x}} \times N_{\mathbf{x}}$ identical matrix and $\mathbf{J}_{N_{\mathbf{x}}}$ is an $N_{\mathbf{x}} \times N_{\mathbf{x}}$ matrix where all entries are 1. This statistic can be used to evaluate the dependence between two variables. We adopt the Gaussian kernel as the characteristic kernel given by

$$K_1(x_k, x_l) = \exp\left(-\frac{\|x_k - x_l\|^2}{2h^2}\right), \tag{6}$$

where h is the hyper-parameter and should be determined carefully to obtain a good result.

3.3. Computational Costs

The kernel-based statistics require $\mathcal{O}(N^2)$ computation for each (k, l) set (Gretton et al., 2012)(Gretton et al., 2008). Therefore, it is time-consuming to compute the statistics with a large number of data samples and is impractical.

4. STATISTICS WITH CHARACTERISTIC FUNCTION

This section shows our proposed ranking measures with characteristic functions (Bisgaard & Sasvári, 2000). We choose the characteristic function to overcome the issue on kernel-based approaches introduced in the previous section. This enables us to compute the statistics significantly faster than the kernel-based approaches. While the kernel-based methods need careful hyper-parameter tuning to keep its performance high, we analyze the characteristic function and propose a method to find a good hyper-parameter, which achieves approximately the same level of performance compared to the kernel-based methods.

We denote a characteristic function as $\phi : \mathbb{R} \to \mathbb{C}$, whose empirical estimates for uni- and bi-variate cases can be given by

$$\hat{\phi}_{X_k}(\omega_k) = \frac{1}{N_{\mathbf{x}}} \sum_{j=1}^{N_{\mathbf{x}}} \exp(i\omega_k x_k^{(j)}), \tag{7}$$

$$\hat{\phi}_{\mathbf{X}_{kl}}(\boldsymbol{\omega}_{kl}) = \frac{1}{N_{\mathbf{x}}} \sum_{j=1}^{N_{\mathbf{x}}} \exp(i\boldsymbol{\omega}_{kl}\mathbf{x}_{kl}{}^T), \tag{8}$$

where $\boldsymbol{\omega}_{kl} = \{\omega_k, \omega_l\} \in \mathbb{R}^2$. We use these functions to evaluate the difference and the dependence.

4.1. Difference Statistic with Characteristic Function

As characteristic functions are injective (Bisgaard & Sasvári, 2000), $\phi_{\mathbf{X}_{kl}}(\boldsymbol{\omega}_{kl}) = \phi_{\mathbf{Y}_{kl}}(\boldsymbol{\omega}_{kl})$ holds for all $\boldsymbol{\omega}_{kl}$ if $\Pr(\mathbf{X}_{kl}) = \Pr(\mathbf{Y}_{kl})$. Therefore, the distance is given by

$$m(\boldsymbol{\omega}_{kl}) = \|\phi_{\mathbf{X}_{kl}}(\boldsymbol{\omega}_{kl}) - \phi_{\mathbf{Y}_{kl}}(\boldsymbol{\omega}_{kl})\|, \tag{9}$$

where $||\cdot||$ denotes the complex norm. This measure can be used to evaluate the difference between two distributions. We rank each two-variable pair with respect to this measure, where if $m(\boldsymbol{\omega}_{kl}) = 0$ for all $\boldsymbol{\omega}_{kl}$, $\mathbf{X}_{kl}$ and $\mathbf{Y}_{kl}$ have the same distribution but otherwise they are different to some extent.

4.2. Dependence Statistic with Characteristic Function

If X_k and X_l are independent of each other, the following equation holds for all $\boldsymbol{\omega}_{kl}$(Bisgaard & Sasvári, 2000).

$$\phi_{\mathbf{X}_{kl}}(\boldsymbol{\omega}_{kl}) = \phi_{X_k}(\omega_k)\phi_{X_l}(\omega_l). \tag{10}$$

Therefore,

$$h(\boldsymbol{\omega}_{kl}) = \|\phi_{\mathbf{X}_{kl}}(\boldsymbol{\omega}_{kl}) - \phi_{X_k}(\omega_k)\phi_{X_l}(\omega_l)\| \tag{11}$$

can be used to evaluate dependence. If $h(\boldsymbol{\omega}_{kl}) = 0$, the k, lth variables are independent of each other but otherwise are dependent on each other to some extent. We rank each two-variable pair with this measure.

4.3. Computational Costs

The characteristic-function-based approach requires $\mathcal{O}(N)$ computation for each (k, l) pair as it requires only the summation of (7) and (8). Thus, computational cost is still moderate with a large number of data samples.

5. APPROPRIATE PRIOR FOR RANKING

Characteristic functions can be used to ensure that there is no difference and/or dependence within the (k, l)th variables pair. However, these approaches must ensure $m(\boldsymbol{\omega}_{kl}) = 0$ and/or $h(\boldsymbol{\omega}_{kl}) = 0$ for all $\boldsymbol{\omega}_{kl}$. This process is both analytically and computationally difficult (Bisgaard & Sasvári, 2000). Our purpose is to *rank* two-variable pairs, and so our approach only requires a comparison of difference and dependence between each two-variable pair. Our intuition is that, for ranking, we do not need to consider all $\boldsymbol{\omega}_{kl}$, but we do need to consider the specific values desirable for ranking.

In this section, we analyze the desirable properties of dependence and difference measures for ranking to determine an appropriate prior distribution to select ω. Given the prior distribution $\Pr(\omega)$, we can use the expectation as an alternative measure to obtain

$$\begin{aligned} M(k,l) &= \int_{\boldsymbol{\omega}_{kl}\in\mathbb{R}^2} m(\boldsymbol{\omega}_{kl})\Pr(\boldsymbol{\omega}_{kl})d\boldsymbol{\omega}_{kl}, & (12) \\ H(k,l) &= \int_{\boldsymbol{\omega}_{kl}\in\mathbb{R}^2} h(\boldsymbol{\omega}_{kl})\Pr(\boldsymbol{\omega}_{kl})d\boldsymbol{\omega}_{kl},. & (13) \end{aligned}$$

We adopt these two measures for ranking instead of (9) and (11). A considerable advantage over the conventional kernel-based approaches is that we can avoid tuning the hyper-parameter.

5.1. Desirable Properties

To order each two-variable pair properly, the following two properties are required:

Prop. 1 (Monotonicity): The measures must have a *monotonic* increase as the size of the difference or the strength of the dependence grows. This property is fundamental for ranking;

Prop. 2 (Sensitivity): This property is desirable when we need to compare between some tiny changes. If the measures are not sensitive enough, it is difficult to distinguish and order them correctly.

The priority of Prop. 1 is higher than Prop. 2. This is because if Prop. 1 is not well satisfied, the measures are not consistent with the difference or dependence changes and cannot rank variables correctly.

5.2. Optimum Appropriate Prior

Based on the analysis of the optimal conditions of ω for dependence and difference measures (see the section 6 for details), we find that the following conditions are expected to satisfy Props. 1 and 2.

- ω needs to be a value near 0 but $\omega \neq 0$ to satisfy Prop. 1;
- ω that takes maximum values for (9) or (11) with each (k, l) pair are desired to satisfy Prop. 2.

Considering the conditions, we introduce a Gaussian distribution as a prior to ω as follows.

$$\Pr(\omega) = \mathcal{N}(\omega|\mu = 0, \sigma = \omega_E), \omega \neq 0, \tag{14}$$

where μ is the mean and σ is the standard deviation. Recall that Prop. 1 has a higher priority than Prop. 2, so we can set a high probability around $\omega = 0$ using a Gaussian distribution as long as we avoid $\omega = 0$. To also satisfy Prop. 2 well with this condition, we use the local maximum nearest to $\omega = 0$

Algorithm 1 Rapid change analysis algorithm

```
Require: $D_{\mathbf{x}}, D_{\mathbf{y}}, N_\omega$
  (1) Sample $\omega$ with the appropriate prior.
  for $k = 1$ to $d - 1$ do
      for $l = k + 1$ to $d$ do
          $\omega_{h,kl} \leftarrow$ Local maximum of (11) nearest to $\omega = 0$.
          $\omega_{m,kl} \leftarrow$ Local maximum of (9) nearest to $\omega = 0$.
      end for
  end for
  $\omega_{E_h}, \omega_{E_m} \leftarrow$ the average of $\omega_{m,kl}, \omega_{h,kl}$ for all $(k, l)$.
  $D_{h,\omega}, D_{m,\omega} \leftarrow N_\omega$ sample from (14) with $\sigma = \omega_{E_h}, \omega_{E_m}$.

  (2) Calculate the contribution rank.
  for $k = 1$ to $d - 1$ do
      for $l = k + 1$ to $d$ do
          Compute $\hat{M}(k, l)$ and $\hat{H}(k, l)$.
      end for
  end for
  $R_M(k, l), R_H(k, l) \leftarrow$ Ranking by $\hat{M}(k, l)$ and $\hat{H}(k, l)$.
  $F(k, l) \leftarrow (R_M(k, l) + R_H(k, l))/2$.
  return $R_F(k, l) \leftarrow$ Ranking by $F(k, l)$.
```

as the variance ω_E. Note that (9) and (11) have extreme points because of periodicity in $\exp(i\omega_k X_k)$ of the characteristic functions. ω_E is estimated simply as follows. For every (k, l)th variable pair, we search from $\omega = 0$ in the positive direction to find the first local maximum $\omega_{E_{kl}}$ using the empirical estimates (7) and (8), and then obtain ω_E as an expectation of $\omega_{E_{kl}}$ with respect to all (k, l)th pairs for both difference and dependence measures. We use ω_E to satisfy Prop. 2 while simultaneously preserving property Prop. 1 by applying the Gaussian distribution. Consequently, N_ω samples $D_{m,\omega} = \{\boldsymbol{\omega}_m^{(1)}, \dots, \boldsymbol{\omega}_m^{(N_\omega)}\}$ and $D_{h,\omega} = \{\boldsymbol{\omega}_h^{(1)}, \dots, \boldsymbol{\omega}_h^{(N_\omega)}\}$, where $\boldsymbol{\omega}^{(j)} = \{\omega_1^{(j)}, \omega_2^{(j)}\}$, are sampled from (14) to obtain the sample mean of (12) and (13) as follows.

$$\hat{M}(k, l) = \frac{1}{N_\omega} \sum_{j=1}^{N_\omega} m(\boldsymbol{\omega}_m^{(j)}), \tag{15}$$

$$\hat{H}(k, l) = \frac{1}{N_\omega} \sum_{j=1}^{N_\omega} h(\boldsymbol{\omega}_h^{(j)}). \tag{16}$$

$m(\boldsymbol{\omega}_m^{(j)})$ and $h(\boldsymbol{\omega}_h^{(j)})$ are calculated with empirical estimates (7) and (8).

5.3. Entire Algorithm

Algorithm 1 shows the entire pseudo-algorithm of our proposed method with the appropriate prior. Given $D_{\mathbf{x}}$ and $D_{\mathbf{y}}$, and the $N_{\mathbf{x}}$ and $N_{\mathbf{y}}$ observations of the d-dimensional variables $\mathbf{X}$ and $\mathbf{Y}$, the algorithm first estimates the hyper-parameter ω_E of (14) for both measures to sample ω. Then, the empirical estimates $\hat{M}(k, l)$ and $\hat{H}(k, l)$ are computed to obtain the integrated measurement $F(k, l)$.

6. ANALYSIS OF APPROPRIATE PRIOR

This section analyzes the optimal conditions of ω for dependence and difference measures (9) and (11). Based on the parameterization using a kernel density estimator, the Gaussian prior distribution (14) is expected to have the best performances for the change analysis.

6.1. Parameterization

In the following, we assume $\omega_k = \omega_l = \omega$ then $\boldsymbol{\omega}_{kl} = \boldsymbol{\omega} = (\omega, \omega)$ for analytical simplicity. Let the probability density function of two random variables $\mathbf{Z} = (Z_1, Z_2) \in \mathbb{R}^2$ be $\Pr(\mathbf{Z}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter that determines the shape of distribution $\mathbf{Z}$. The corresponding characteristic function is

$$\phi_{\mathbf{Z}}(\boldsymbol{\omega}|\boldsymbol{\theta}) = \int_{\mathbf{Z} \in \mathbb{R}^2} \exp(i\boldsymbol{\omega}\mathbf{Z}^T) \Pr(\mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}, \tag{17}$$

As discussed in equations (9) and (11), we obtain

$$\begin{aligned} m(\boldsymbol{\omega}|\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) &= \|\phi(\boldsymbol{\omega}|\boldsymbol{\theta}) - \phi(\boldsymbol{\omega}|\tilde{\boldsymbol{\theta}})\|, & (18) \\ h(\boldsymbol{\omega}|\boldsymbol{\theta}) &= \|\phi(\boldsymbol{\omega}|\boldsymbol{\theta}) - \phi(\omega_1|\boldsymbol{\theta}_1)\phi(\omega_2|\boldsymbol{\theta}_2)\|, & (19) \end{aligned}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the parameters of Z_1 and Z_2, respectively. $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ represent the parameters of different distributions. Here, we assume that all two-variable pairs have their own distribution parameters, $\boldsymbol{\theta}$. Thus, we want to analyze the relationship between the change in $\boldsymbol{\theta}$ and the changes of the two measures, (18) and (19), to rank them correctly.

6.2. Kernel Density Estimator Modeling

The kernel density estimator is a general model to estimate a probability distribution (Silverman, 1986). This model is used to analyze two measures, (18) and (19). Let $D_{z_1} = \{z_1^{(1)}, \dots, z_1^{(N)}\}$ be the N observations of the random variable Z_1. We model the data distribution by a univariate kernel density estimator with the Gaussian kernel written as

$$\Pr(Z_1|h, D_{z_1}) = \frac{1}{Nh} \sum_{j=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(Z_1 - z_1^{(j)})^2}{2h^2}\right), \tag{20}$$

where $\boldsymbol{\theta} = \{h, D_{z_1}\}$ and h is the parameter of the Gaussian kernel. Thus, the corresponding characteristic function is

$$\begin{aligned} \phi_{Z_1}(\omega|h, D_{z_1}) &= \int_{-\infty}^{\infty} \exp(i\omega Z_1) \Pr(Z_1|h, D_{z_1}) dZ_1 \\ &= \frac{1}{Nh\sqrt{2\pi}} \sum_{j=1}^{N} \int_{-\infty}^{\infty} \exp\Bigg(i\omega Z_1 - \\ &\qquad \frac{(Z_1 - z_1^{(j)})^2}{2h^2}\Bigg) dZ_1 \\ &= \frac{1}{N} \sum_{j=1}^{N} \exp\left(i\omega z_1^{(j)} - \frac{(h\omega)^2}{2}\right). \end{aligned} \tag{21}$$

For the bivariate case $\mathbf{Z} = (Z_1, Z_2)$, given N samples $D_{\mathbf{z}} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}\}$, the bivariate kernel density estimator is given as

$$\mathrm{Pr}(\mathbf{Z}|\mathbf{H}, D_{\mathbf{z}}) = \frac{1}{N|\mathbf{H}|^{-1/2}} \times \sum_{j=1}^{N} \frac{1}{2\pi} \exp\left(-(\mathbf{Z} - \mathbf{z}^{(j)})^T \mathbf{H}^{-1} (\mathbf{Z} - \mathbf{z}^{(j)})\right). \quad (22)$$

For simplicity, we set

$$\mathbf{H} = \begin{bmatrix} h^2 & 0 \\ 0 & h^2 \end{bmatrix} \quad (23)$$

Therefore, the corresponding characteristic function is

$$\phi_{\mathbf{Z}}(\boldsymbol{\omega}|\mathbf{H}, D_{\mathbf{x}}) = \int_{-\infty}^{\infty} \exp(i\boldsymbol{\omega}\mathbf{Z}^T)\mathrm{Pr}(\mathbf{Z}|\mathbf{H}, D_{\mathbf{x}})d\mathbf{Z} = \frac{1}{N}\sum_{j=1}^{N} \exp\left(i\omega\left(z_1^{(j)} + z_2^{(j)}\right) - (h\omega)^2\right). \quad (24)$$

In the following sections, we use these models to analyze the optimal conditions of ω to rank difference and independence.

6.3. Difference Measure Analysis

Given another N sample dataset $D_{\tilde{\mathbf{z}}} = \{\tilde{\mathbf{z}}^{(1)}, \ldots, \tilde{\mathbf{z}}^{(N)}\}$, based on equations (18), (20) and (22), $\boldsymbol{\theta} = \{\mathbf{H}, D_{\mathbf{z}}\}$ and $\tilde{\boldsymbol{\theta}} = \{\mathbf{H}, D_{\tilde{\mathbf{z}}}\}$, we then obtain

$$m(\boldsymbol{\omega}|\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \frac{1}{N}\exp\left(-(h\omega)^2\right)\left\|\sum_{j=1}^{N}\left\{\exp\left(i\omega(z_1^{(j)} + z_2^{(j)})\right) - \exp\left(i\omega(\tilde{z}_1^{(j)} + \tilde{z}_2^{(j)})\right)\right\}\right\|. \quad (25)$$

Here, we assume $\mathbf{H}$ is fixed and the change in the difference is that of $D_{\mathbf{z}}$ and $D_{\tilde{\mathbf{z}}}$. Therefore, if $\omega = 0$, $m(\boldsymbol{\omega}|\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = 0$ for any $D_{\mathbf{z}}$ and $D_{\tilde{\mathbf{z}}}$, so that $\omega \neq 0$. We also find that (25) has periodicity with $\exp(i\omega(z_1^{(j)} + z_2^{(j)}))$ and $\exp(i\omega(\tilde{z}_1^{(j)} + \tilde{z}_2^{(j)}))$ inside the complex norm. Indeed, the following inequality holds for any $D_{\mathbf{z}}$ and $D_{\tilde{\mathbf{z}}}$ because of the periodicity.

$$\begin{aligned} 0 &\leq \left\|\sum_{j=1}^{N}\left\{\exp\left(i\omega(z_1^{(j)} + z_2^{(j)})\right) - \exp\left(i\omega(\tilde{z}_1^{(j)} + \tilde{z}_2^{(j)})\right)\right\}\right\| \\ &\leq \sum_{j=1}^{N}\left\|\exp\left(i\omega(z_1^{(j)} + z_2^{(j)})\right) - \exp\left(i\omega(\tilde{z}_1^{(j)} + \tilde{z}_2^{(j)})\right)\right\| \\ &\leq 2N, \end{aligned} \quad (26)$$

where

$$0 \leq \left\|\exp\left(i\omega(z_1^{(j)} + z_2^{(j)})\right) - \exp\left(i\omega(\tilde{z}_1^{(j)} + \tilde{z}_2^{(j)})\right)\right\| \leq 2. \quad (27)$$

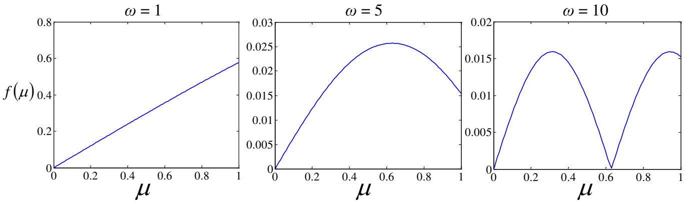

Figure 2. Empirical study of $f(\mu)$ with different $\omega = 1, 5, 10$.

Because of these restricted ranges, if ω is large, (27) is sensitive to the variance of $D_{\mathbf{z}}$ and $D_{\tilde{\mathbf{z}}}$ and the periodicity is dominant, such that monotonicity does not hold. Figure 2 shows this fact empirically. It shows $f(\mu) = \| \sum_{j=1}^{N}\{\exp(i\omega z^{(j)}) - \exp(i\omega\tilde{z}^{(j)})\}\|$ by using $N = 10000$ samples of $z \sim \mathcal{N}(0, 1)$ and $\tilde{z} \sim \mathcal{N}(\mu, 1)$ with μ varying from 0 to 1. With the same range of μ, $f(\mu)$ shows monotonicity with small ω and the periodicity increases with large ω. This implies that we need to select ω near 0 to avoid periodicity and instead ensure monotonicity with a wide range of $D_{\mathbf{z}}$ and $D_{\tilde{\mathbf{z}}}$. In addition, the equation (25) can be bound as follows according to (26).

$$0 \leq m(\boldsymbol{\omega}|\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \leq 2\exp\left(-(h\omega)^2\right). \quad (28)$$

The upper bound is maximized when $\omega = 0$ and monotonically decreases with ω. This fact means the size of $m(\boldsymbol{\omega}|\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ becomes smaller with larger ω, and the sensitivity with the change is expected to be lower for any $D_{\mathbf{z}}$ and $D_{\tilde{\mathbf{z}}}$.

In consideration of the above analysis, we need ω to have a value near 0 but to satisfy Prop. 1 $\omega \neq 0$. To satisfy also Prop. 2 with this condition, we use the local maximum nearest to $\omega = 0$. Note that (25) has extreme points because of periodicity. These facts are ensured by the empirical analysis shown in Section 7.3.

6.4. Dependence Measure Analysis

Based on equations (19), (20) and (22), $\boldsymbol{\theta} = \{\mathbf{H}, D_{\mathbf{z}}\}$ and we obtain

$$h(\boldsymbol{\omega}|\boldsymbol{\theta}) = \frac{1}{N}\exp\left(-(h\omega)^2\right)\left\|\left\{\sum_{j=1}^{N}\exp\left(i\omega(z_1^{(j)} + z_2^{(j)})\right)\right\} - \frac{1}{N}\left\{\sum_{j=1}^{N}\sum_{k=1}^{N}\exp\left(\omega(z_1^{(j)} + z_2^{(k)})\right)\right\}\right\|. \quad (29)$$

Here, we assume a fixed $\mathbf{H}$ and the change in dependence as that of $D_{\mathbf{z}}$. If $\omega = 0$, $h(\boldsymbol{\omega}|\boldsymbol{\theta}) = 0$ for any $D_{\mathbf{z}}$, so that $\omega \neq 0$. Each $\exp(i\omega(z_1^{(j)} + z_2^{(k)}))$ inside the complex norm also has periodicity. Therefore, we need to select ω near 0 to preserve monotonicity with a wide range of $D_{\mathbf{z}}$. In addition, equation (29) has bound

$$0 \leq h(\boldsymbol{\omega}|\theta) \leq 2 \exp\left(-(h\omega)^2\right). \tag{30}$$

The upper bound is maximized when $\omega = 0$ and monotonically decreases with ω. Therefore, the range of $h(\boldsymbol{\omega}|\theta)$ is smaller with larger ω; thus, the sensitivity with the change is expected to be lower.

As a result, the same conditions described in Section 6.3 are required. We need ω to have a value near 0 but $\omega \neq 0$ to satisfy property Prop. 1. To also satisfy Prop. 2 with this condition, we use the local maximum nearest to $\omega = 0$. Note that (29) has extreme points because of periodicity.

7. EXPERIMENTAL RESULTS

This section presents the experimental results that validate the proposed method. First, we compared our method with related approaches to show that the computational cost has been decreased markedly, in addition to maintaining good performance in the kernel-based method. Second, real driving data were used to evaluate how well our method can represent the changes occurring between the data. We used a Windows 7 (64-bit) computational environment with Intel$^{(\mathrm{R})}$ Core$^{(\mathrm{TM})}$ i7-3970X CPU @ 3.50 GHz, 3.50 GHz and 64.0 GB memory. All implementations were performed by MATLAB R2012b.

7.1. Evaluation with Benchmark Datasets

We evaluated our proposed method using several benchmark datasets from the UCI Machine Learning Repository (Dua & Graff, 2017), as shown in Table 1. Given that some datasets of Covtype, Pamap2, and Statlog have more than two classes, the largest class in the dataset was selected as the normal data, while all other classes were chosen to be changed data. As preprocessing, every row that had a missing value and every constant or binary attribute were eliminated. Samples of Covtype and Pamap2 were restricted to 10000 samples to manage the computational overhead for the kernel methods. The data were also normalized such that each attribute of the normal data had mean 0 with unit variance.

Table 1. Benchmark datasets.

Datasets	Normal Samples	Changed Samples	Attributes
Arrhythmia	245	207	174
Covtype	10000	10000	10
Mfeat	1000	1000	648
Optdigits	3822	1796	60
Pamap2	10000	10000	51
Spambase	2788	1813	57
Statlog	1533	4902	36

Table 2. Comparison of approaches

Approaches	Parameters	Difference	Dependence
CharG(ω_E)	$N_\omega = 100$, $\sigma = \omega_E$	✓	✓
CharG(1)	$N_\omega = 100$, $\sigma = 1$	✓	✓
CharU	$N_\omega = 100$, $[-\omega_E, \omega_E]$	✓	✓
Hist	$N_b = 8$	✓	✓
KS	-	✓	-
T^2	-	✓	-
PD(1)	$N_b = 8$, $\lambda = 1$	-	✓
PD(2)	$N_b = 8$, $\lambda = 2$	-	✓
KernG	h	✓	✓

We compared our approach with several related methods, as shown in Table 2, i.e., three kinds of priors, namely, Gaussian with $\sigma = \omega_E$ (CharG(ω_E)), Gaussian with $\sigma = 1$ (CharG(1)), and the uniform distribution with the range $[-\omega_E, \omega_E]$ (CharU). The purpose of comparing different priors is to confirm the benefits in using $\sigma = \omega_E$ rather than simply $\sigma = 1$, as well as using a Gaussian distribution rather than a uniform distribution, as discussed in Section 5. Our method, denoted by CharG(ω_E), is also compared with some traditional approaches. Hist is a contingency-based approach that divides a data region by $N_b \times N_b$ bins to approximate the density of each (m, n)th bin by the proportion of data falling in. The following measures were used to evaluate difference and dependence based on the probabilistic identity $\Pr(X_{kl}) = \Pr(Y_{kl})$ and independence $\Pr(X_{kl}) = \Pr(X_k)\Pr(X_l)$ for the (k, l)th variable pair.

$$M(k,l) = \sum_{m,n=1}^{N_b} \|\hat{P}_{X_{kl}}^{mn} - \hat{P}_{Y_{kl}}^{mn}\|_2, \tag{31}$$

$$H(k,l) = \sum_{m,n=1}^{N_b} \|\hat{P}_{X_{kl}}^{mn} - \hat{P}_{X_k}^{m}\hat{P}_{X_l}^{n}\|_2, \tag{32}$$

where $\|\cdot\|_2$ is the L_2 norm, $\hat{P}_{X_{kl}}^{mn}$ denotes the proportion of data falling into the (m, n)th bin with samples from X_{kl} and $\hat{P}_{X_k}^{m} = \sum_{n=1}^{N_b} \hat{P}_{X_{kl}}^{mn}$. KS is the two-dimensional version of the Kolmogorov–Smirnov statistic, which is a popular approach for two-sample testing (Smirnov, 1948)(Franceschini & Fasano, 1987). T^2 is Hotelling's T^2 statistic (Hotelling, 1992), which is the multivariate extension of the univariate t-test. PD(1) and PD(2) are the power divergences (Chen-Jen & Terrence, 2005), which are contingency-based approaches to measure dependence; we set N_b to be the same as for the Hist. The hyper-parameter λ was set to $\lambda = 1$ and 2, respectively. Finally, KernG denoted the kernel statistics.A bandwidth parameter h was set to a medium distance between all data points, i.e., $h = \mathrm{medium}\{\|\mathbf{x}^{(j)} - \mathbf{x}^{(k)}\|\}$, where $\mathbf{x}^{(j)}$ was the jth sample, according to (Gretton et al., 2012)(Gretton et

Table 3. Rank correlation with KernG (difference).

	Arrhythmia	Covtype	Mfeat	Optdigits	Pamap2	Spambase	Statlog	Average
CharG(ω_E)	0.9660	**0.9924**	**0.9723**	**0.9759**	**0.9761**	**0.8807**	**0.9785**	**0.9629**
CharG(1)	**0.9697**	0.9916	0.9691	0.9740	0.9461	0.8689	0.9650	0.9427
CharU	0.8377	0.9301	0.7703	0.8443	0.7587	0.8237	-0.5502	0.6645
Hist	0.7309	0.9036	0.8338	0.0690	0.6124	0.5242	0.9100	0.6907
KS	-0.1077	0.9195	0.1282	-0.3341	0.8514	-0.2816	0.9244	0.3837
T^2	0.5617	0.9191	0.8183	0.6654	0.4345	0.6072	-0.5189	0.5565
PD(1)	-	-	-	-	-	-	-	-
PD(2)	-	-	-	-	-	-	-	-

Table 4. Rank correlation with KernG (dependence).

	Arrhythmia	Covtype	Mfeat	Optdigits	Pamap2	Spambase	Statlog	Average
CharG(ω_E)	**0.9654**	**0.9907**	**0.9760**	**0.9931**	**0.9865**	**0.9481**	**0.9947**	**0.9792**
CharG(1)	0.9594	0.9849	0.9621	0.9895	0.9655	0.9278	0.9936	0.9706
CharU	0.8895	0.9301	0.9240	0.9671	0.9198	0.8974	0.9439	0.9249
Hist	0.8086	0.9036	0.8790	0.9273	0.6238	0.4081	0.9479	0.7719
KS	-	-	-	-	-	-	-	-
T^2	-	-	-	-	-	-	-	-
PD(1)	0.6277	0.7799	0.7185	0.7829	0.5557	0.9324	0.9041	0.7693
PD(2)	0.4004	0.5762	0.6523	0.5411	0.4666	0.8584	0.8578	0.6340

Table 5. Total computational time [s] (ratio to KernG [%]).

	Arrhythmia	Covtype	Mfeat	Optdigits	Pamap2	Spambase	Statlog
CharG(ω_E)	54.11(38)	5.330(0.2)	2532(4.4)	71.98(0.7)	151.6(0.2)	50.42(1.3)	17.37(1.4)
CharG(1)	26.70(19)	3.843(0.1)	1501(2.6)	46.82(0.5)	108.6(0.1)	32.06(0.8)	13.20(1.1)
CharU	54.89(39)	5.198(0.2)	2462(4.2)	72.46(0.7)	158.5(0.2)	50.09(1.3)	17.09(1.4)
Hist	12.17(8.6)	0.1112(0.003)	199.1(0.3)	2.250(0.02)	2.840(0.003)	1.773(0.05)	0.7675(0.06)
KS	189.5(134)	553.6(16)	27364(47)	1591(16)	15300(16)	701.2(18.2)	593.6(48)
T^2	1.754(1.2)	0.0188(0.0005)	28.80(0.05)	0.3687(0.004)	0.5306(0.006)	0.3111(0.008)	0.1395(0.01)
PD(1)	8.676(6.1)	0.0714(0.002)	138.0(0.2)	1.691(0.02)	1.935(0.002)	1.263(0.03)	0.4525(0.04)
PD(2)	8.808(6.2)	0.0709(0.002)	138.1(0.2)	1.669(0.02)	1.920(0.002)	1.264(0.03)	0.4507(0.04)
KernG	141.1(100)	3434(100)	58051(100)	10000(100)	96140(100)	3855(100)	1247(100)

al., 2008).

Tables 3–5 show the results of Spearman's rank correlation between KernG and others, as well as their total calculation times. The calculation time includes the evaluations of all two-variable pairs. The results are only for the difference with the KS and T^2 methods and the dependence with PD(1) and PD(2) because these methods cannot evaluate both. The total computational time of CharG(ω_E) and CharU includes the time taken to determine the hyper-parameter ω of Algorithm 1 (1).

According to the results, CharG(ω_E), our proposed method reported the best correlation with KernG for almost all datasets and had a correlation coefficient consistently close to unity, whereas the other methods that were used for comparison failed to obtain a consistently high correlation to KernG. Thus, the performance of our method is only comparable to the state-of-the-art kernel method. The total computational time is also notable in that our method yielded a significant reduction compared with KernG. This demonstrates the advantages of our approach in regard to computational cost and performance.

Among the proposed methods with different priors between CharG(ω_E), CharG(1) and CharU, CharG(ω_E) reported the best performances for almost all datasets. Therefore, by comparing CharG(ω_E) with CharG(1), it is better to set $\sigma = \omega_E$ to satisfy property Prop. 2 rather than ignoring ω_E. Comparing CharG(ω_E) with CharU, we can say that sampling around $\omega = 0$ to satisfy Prop. 1 should have a higher priority than sampling around $\omega = \omega_E$ to satisfy Prop. 2, as CharG(ω_E) used a Gaussian distribution to sample more around $\omega = 0$. Nevertheless, CharG(1) reported the fastest computational time among the three methods because it used the preset hyper-parameter $\omega_E = 1$. Despite that, there is no justification for the performance, and it is better to assume some cost to choose a proper hyper-parameter, as CharG(ω_E) did.

The other methods that were compared, i.e., Hist, KS, T^2, and PD, reported fast computational times but often poor rank correlations. Hist and PD are contingency-based methods, hence the number of bins should be set properly for each dataset to achieve good performances; however, it is usually

Table 6. List of real driving data attributes.

No.	Attributes	No.	Attributes
1	Shift Position	24	Objective Air/Fuel Ratio
2	Parking Brake	25	Air/Fuel Ratio
3	Sports Mode Switch	26	Purge Rate
4	Engine Stop Request	27	O_2 Sensor Voltage
5	Idle Control	28	Ignition Timing
6	P Range Racing	29	Objective Exhaust Gas-Recirculation Valve Position
7	Warming Request		
8	Electric W/P Motor Rotation	30	Stroke Sensor 1
9	Vehicular Speed Sensor 1	31	Stroke Sensor 2
10	Vehicular Speed Sensor 2	32	Accumulator Pressure
11	Accelerator Position	33	Front Rear G Sensor
12	Intake Air Volume	34	Regenerative Cooperation Brake
13	Required Throttle Position	35	Executed Regenerative Torque
14	Throttle Position (Sensor Value)	36	Required Regenerative Torque
15	Throttle Position (Directed Voltage)	37	Yaw Rate Sensor 1
16	Throttle Position	38	Yaw Rate Sensor 2
17	Engine Speed	39	Steering Angle Sensor
18	Required Engine Output	40	Lateral G Sensor
19	Objective Engine Speed	41	Yaw Rate Value
20	Real Engine Torque	42	Steering Angle Value
21	Idle Speed Control Flow	43	Zero-Point Corrected-Steering Angle Sensor
22	Idle Speed Control Position		
23	Idle Speed Control Flow (Learned Value)		

difficult, and they reported poor results for some datasets. KS uses the maximum difference between the cumulative distributions of two samples. However, only using the maximum difference as a statistic is insufficient to distinguish any kind of differences to rank two-variable pairs correctly. This is because KS is built only to test if two samples have a difference. T^2 is the linear statistic so it was unable to evaluate non-linear correlations.

Among the datasets, CharG(ω_E), our proposed method, reported a lower reduction ratio with the Arrhythmia dataset (38 %) than the other datasets. This is because the Arrhythmia dataset has a significantly smaller number of samples, i.e. 245 and 207 samples (Table 1). With the sample size, as mentioned earlier, the computation scales quadratically with kernel methods (Section 3.3) and linearly with our method (section 4.3). Then, if the number of samples is small, there will be a small difference in computation. Nevertheless, it is inconsequential because the absolute computational time is short with a small amount of samples.

7.2. Evaluation with Real Driving Data

We conducted the practical experiment using real vehicle driving data to demonstrate how well our framework can support operators to analyze changes, as well as the representation power of scatter plots.

7.2.1. Datasets

We used the driving data to evaluate the performance of the representation. The problem considered here is to analyze changes between two driving environments. The fault diagnosis equipment recorded information about the vehicle, and the data contained 43 attributes with sampling rates of 0.5 s. The attributes are listed in Table 6. The data were recorded under the following two conditions. First, under flat-road conditions, the vehicle ran on the same flat road on several occasions with constant acceleration and deceleration to obtain 1450 samples as the normal dataset. Second, we used the same flat-road condition with slower acceleration and deceleration to obtain 737 samples as the changed dataset. For preprocessing, every constant or binary attribute were eliminated. The data are also normalized such that each attribute of the normal data has mean of 0 with unit variance. Between the two datasets, driver inputs, specifically Stroke Sensor (representing deceleration) and Accelerator Position, had been changed. Thus, the causes of the change are Stroke Sensor (representing deceleration) and Accelerator Position. The goal of this experiment was to evaluate how well our framework identifies and represents the causes of changes for operators so that they can well interpret why the change occurred.

7.2.2. Results

Table 7 shows Spearman's rank correlation between CharG(ω_E) and KernG with their total calculation time. All results are the average of ten trials. The result is consistent with that of Section **??** where CharG(ω_E) reported the best correlation with the KernG while reducing the computational time significantly.

The change representation results based on the KernG are presented in Figure 3 and that for CharG(ω_E) in Figure 4.

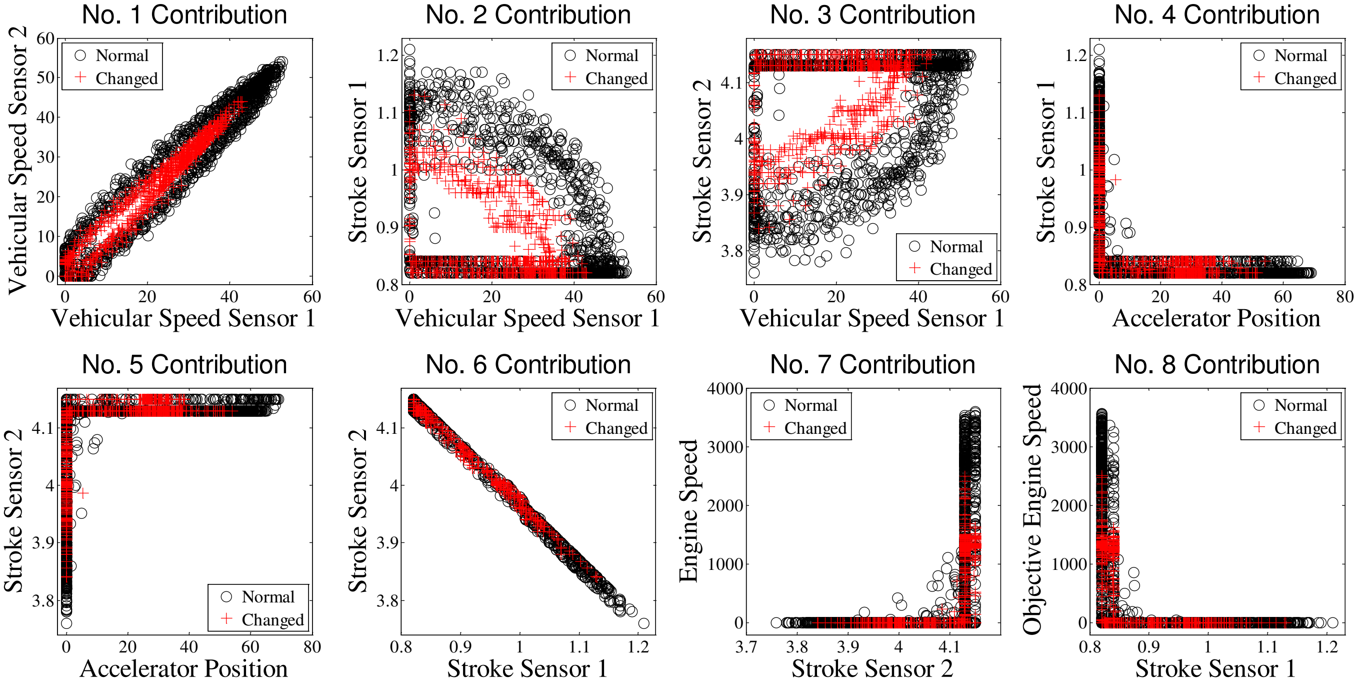

Figure 3. Eight contributing combinations determined by the characteristic kernel methods, which identify changes, i.e., Stroke Sensor (representing deceleration) and Accelerator Position, and the changes in their relationships with the Vehicular Speed Sensor.

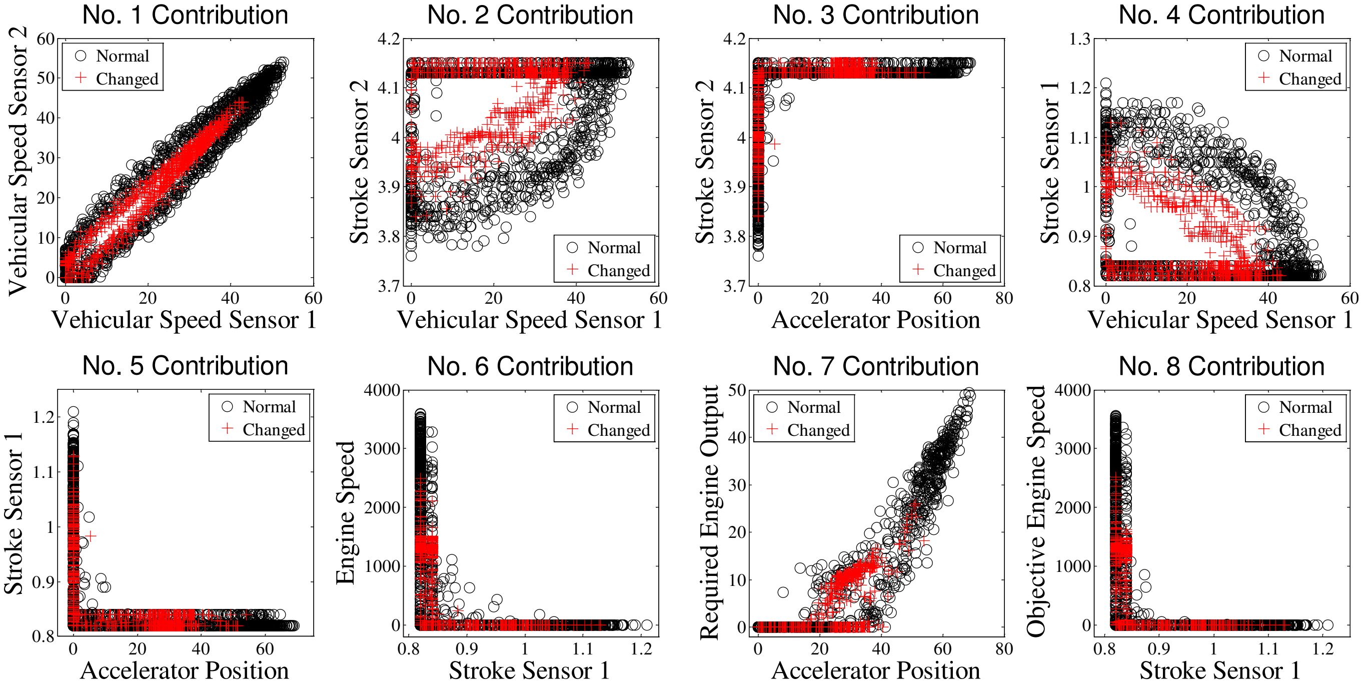

Figure 4. Eight contributing combinations determined by our proposed method, which identify the changes. These results are almost the same as for the characteristic kernel methods of Figure 3.

Table 7. Rank correlation and computational time.

	Difference	Dependence	Time [s]
CharG(ω_E)	**0.8876**	**0.9284**	14.32
CharG(1)	0.8808	0.9265	8.600
CharU	0.7654	0.8914	13.36
Hist	0.4904	0.7553	0.5947
KS	0.0079	-	60.75
T^2	0.4309	-	0.1191
PD(1)	-	0.8163	0.3854
PD(2)	-	0.6890	0.3837
KernG	1.0000	1.0000	283.9

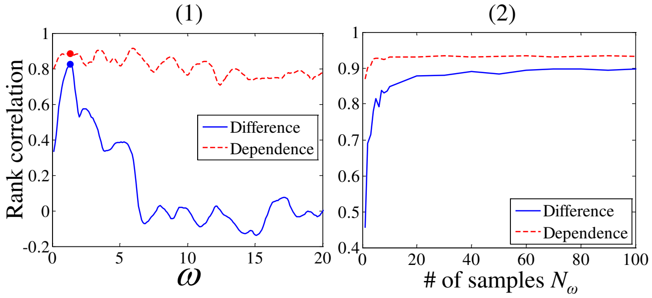

Figure 5. Relation to (1) ω and (2) N_ω.

Each show eight main pairs for the contributing attributes, and these are plotted for both the normal data (Normal) and the changed data (Changed).

Among the results, the same six out of eight pairs were extracted. This result is consistent because the rank correlation is high. The representation power is also valid (Figure 4) in that it successfully extracts the Stroke Sensor (representing deceleration) and Accelerator Position, the cause of changes, combined with the proper attributes, which explicitly reflect the difference between the datasets. Indeed, the operators can see that the acceleration and deceleration power are lower in the changed dataset than in the normal dataset as well as the vehicular speed. Therefore, it suggests that the driver has changed driving patterns.

7.3. Empirical Analysis of ω and N_ω

This section shows the effect of ω and N_ω on performance when applying our method to the real driving datasets used in Section 7.2. We calculated the average from ten trials to get the following results.

Figure 5 (1) shows the Spearman rank correlation between KernG and our method with ω varying from 0 to 20 instead of using any priors. We used only a single ω to evaluate every pair. The dots represent the computed values $\omega_E = 1.3191$ for the difference measure and $\omega_E = 1.3089$ for the dependence measure. This result supports the analysis in Section 5: the higher rank correlation was observed with the expected local maximum of ω_E near $\omega = 0$, and the performance declines with larger ω values.

Figure 5 (2) shows the rank correlation between KernG and CharG(ω_E), but with the number of samples, N_ω, varying from 1 to 100, it can be seen that the performance saturates quickly with fewer samples and obtains a high rank correlation with KernG. Therefore, our method is robust and does not require operators to set the hyper-parameters carefully.

8. CONCLUSION

In this paper, we proposed a new approach for change analysis by using the characteristic function with the appropriate prior. Our method follows a novel framework that uses the statistics of difference and dependence to compare a normal data with a data containing unknown changes. Based on statistics, each two-variable pair of the given datasets is ranked to represent the change well. Given the order, the representation is based on a bivariate scatter plot, which can be easily understood by operators not familiar with statistics. The main contribution of this study is the use of the characteristic-function-based ranking strategy with its hyper-parameter characterized by the proposed prior distribution, which is optimum for the ranking purpose. Using the prior enables us to reduce the computational cost by marginalizing the hyper-parameter while maintaining performance to the same accuracy to the state-of-the-art kernel-based methods. The experimental results based on popular benchmark datasets validated the advantage of our strategy. The practical experiment using real vehicle driving data demonstrated how well our framework can support operators to analyze changes, as well as the representation power of scatter plots.

REFERENCES

Bisgaard, T., & Sasvári, Z. (2000). *Characteristic functions and moment sequences: Positive definiteness in probability*. Nova Science Publishers.

Chandola, V., Banerjee, A., & Kumar, V. (2009). /em anomaly detection: A survey. *ACM Comput. Surv.*, *41*(3), 15:1–15:58.

Chen-Jen, K., & Terrence, L. (2005, May). Testing for stochastic independence: application to blind source separation. *IEEE Transactions on Signal Processing*, *53*(5), 1815-1826.

Dua, D., & Graff, C. (2017). *UCI machine learning repository.* Retrieved from `http://archive.ics.uci.edu/ml`

Franceschini, A., & Fasano, G. (1987, 03). A multidimensional version of the Kolmogorov–Smirnov test. *Monthly Notices of the Royal Astronomical Society*, *225*(1), 155-170.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012, March). A kernel two-sample test.

J. Mach. Learn. Res., *13*(1), 723–773.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2008). A kernel statistical test of independence. In *Advances in neural information processing systems 20* (pp. 585–592). Curran Associates, Inc.

Grinstein, G., Trutschl, M., & Cvek, U. (2001). High-dimensional visualizations. In *Proceedings of the data mining conference (kdd).*

He, B., Yang, X., Chen, T., & Zhang, J. (2012, 08). Reconstruction-based multivariate contribution analysis for fault isolation: A branch and bound approach. *Journal of Process Control*, *22*, 1228-1236.

Hido, S., Idé, T., Kashima, H., Kubo, H., & Matsuzawa, H. (2008). Unsupervised change analysis using supervised learning. In T. Washio, E. Suzuki, K. M. Ting, & A. Inokuchi (Eds.), *Advances in knowledge discovery and data mining* (pp. 148–159). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hotelling, H. (1992). The generalization of student's ratio. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Foundations and basic theory* (pp. 54–65). New York, NY: Springer New York.

Hyvärinen, A., & Oja, E. (2000, May). Independent component analysis: Algorithms and applications. *Neural Netw.*, *13*(4-5), 411–430.

Joe Qin, S. (2003). Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, *17*(8-9), 480-502.

Leban, G., Zupan, B., Vidmar, G., & Bratko, I. (2006, Sep 01). Vizrank: Data visualization guided by machine learning. *Data Mining and Knowledge Discovery*, *13*(2), 119–136.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (Third ed.). New York: Springer.

Marimont, R. B., & Shapiro, M. B. (1979, 08). Nearest Neighbour Searches and the Curse of Dimensionality. *IMA Journal of Applied Mathematics*, *24*(1), 59-70.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Taylor & Francis.

Smirnov, N. (1948, 06). Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, *19*(2), 279–281.

BIOGRAPHIES

Takaaki Tagawa received the BE in mechanical engineering from Waseda University, Tokyo, Japan, in 2010, and the ME in aeronautics and astronautics engineering from the University of Tokyo, Japan, in 2012. Since 2012 he has been with Toyota Central R&D Labs., Inc. His current research interests include data mining and machine learning for anomaly detection and their applications.

Yukihiro Tadokoro received the BE, ME, and PhD degrees in information electronics engineering from Nagoya University, Aichi, Japan, in 2000, 2002, and 2005, respectively. Since 2006 he has been with Toyota Central R&D Labs., Inc. In 2011 and 2012 he worked as a Research Scholar in the Department of Physics and Astronomy, Michigan State University, USA, to study nonlinear phenomena for future application in the signal and information processing fields. His current research interests include data mining and machine learning, in addition to noise-related phenomena in nonlinear systems and their applications. He is a member of the Institute of Electronic, Information and Communication Engineers (IEICE), Japan and IEEE.

Takehisa Yairi received the MSc, and PhD degrees in aerospace engineering from the University of Tokyo, Japan in 1996 and 1999, respectively. He is currently a Professor with the Research Center for Advanced Science and Technology, the University of Tokyo. His research interests include machine learning theory and its application.