

# Approximate Conditional Independence Test using Residuals

Shinsuke Uda<sup>a</sup>

Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi Higashi-ku Fukuoka 812-8582, Japan

**Keywords:** Information Theory, Conditional Mutual Information, Permutation Test, Biological Data Analysis.


**Abstract:** Conditional mutual information is a useful measure for detecting the association between variables that are also affected by other variables. Though permutation tests are used to check whether the conditional mutual information is zero to indicate mutual independence, permutations are difficult to perform because the other variables in a dataset may be associated with the variables in question. This problem is particularly acute when working with datasets of small sample size. This study aims to propose a computational method for approximating conditional mutual information based on the distribution of residuals in regression models. The proposed method can implement the permutation tests for statistical significance by translating the problem of measuring conditional independence into the problem of estimating simple independence. Additionally, a reliability of  $p$ -value in permutation test is defined to omit unreliably detected associations. We tested our proposed method's performance in inferring the network structure of an artificial gene network against comparable methods submitted to the Dream4 challenge.

## 1 INTRODUCTION

Recent progress in the field of information technology has made the analysis of high-dimensional datasets possible with novel analytic techniques. However, these techniques remain in development stage due to some problems faced to the curse of dimensionality (Hastie et al., 2001). When three variables may be dependent on each other, we may want to examine the associations between variables  $X$  and  $Y$  while ignoring the effects of  $Z$ , which apparently boosts/depresses association. The dataset is of high dimension and  $X \subset \mathcal{R}$ ,  $Y \subset \mathcal{R}$  and  $Z \subset \mathcal{R}^m$  are continuous random variables. Conditional mutual information is a statistically robust measure to quantify the association between variables while removing the effects of other variables, without assuming a particular model distribution.  $X$  and  $Y$  are conditionally independent given  $Z$  if and only if the conditional mutual information between  $X$  and  $Y$  given  $Z$  is zero, and vice versa. Conditional independence can serve as a measure to infer the network structure of associations between variables (Pearl, 2009). Perfect conditional independence between  $X$  and  $Y$  given  $Z$  indicates that no edge connects nodes  $X$  and  $Y$ . In inference networks, in most of the cases, it is assumed that the dimensions of  $X$  and  $Y$  are one, therefore, hereafter,

we will assume it. While conditional mutual information has theoretical advantages, there are two difficulties for inferring network from finite sample size datasets. First, estimating conditional mutual information is computationally difficult if the dimension of  $Z$  is high, even if  $X$  and  $Y$  are one-dimensional. Second, though conditional independence is detected by examining whether conditional mutual information is zero or not, the conditional mutual information estimated from a dataset with finite sample size is rarely zero due to statistical errors, even if the variables are conditionally independent.

To infer the associations in high-dimension networks, statistical tests must be employed to examine whether the conditional mutual information is close to zero or not. A permutation test is a powerful strategy for this analysis and involves generating the expected data distribution of the null hypothesis without making any assumptions about the distribution of the actual data. Permutations are not easy to apply for examining the conditional independence between  $X$  and  $Y$  given  $Z$ , because  $Z$  may be associated with  $X$  and/or  $Y$ , and the permutations of  $X$  and  $Y$  ignore the influence of  $Z$ . This problem can be addressed by grouping the data in resampling, but this approach requires a large sample. This article proposes a computational method for approximating conditional mutual information. The method translates the test for conditional independence with a simple independence test

<sup>a</sup>  <https://orcid.org/0000-0001-6221-3587>

using the distribution of residuals in regression models of  $X$  or  $Y$  against  $Z$ . With this translation of the problem, permutations can be used to isolate the associations between  $X$  and  $Y$ . Additionally, a reliability of  $p$ -value in permutation test is defined to omit unreliably detected associations.

The biological datasets compiled in recent years are of high dimension, because of technological developments in measurement techniques that allow the measurement of greater numbers of molecular species. The structure of associations between molecular species can be expressed as a regulatory network when seeking insight into biological phenomena. To test our proposed analytic method with a widely available dataset, the method was used to infer the artificial gene network structure in the Dream4 challenge dataset (Schaffter et al., 2011). Next, Section 2 presents related works including comparable methods of gene network inference, Section 3 introduces the proposed method for approximating conditional mutual information, Section 4 discusses the results of artificial gene network inference, and Section 5 concludes the paper.

## 2 RELATED WORK

Mutual information

$$I(X;Y) = \int dx dy p(x,y) \log \frac{p(x,y)}{p(x)p(y)},$$

is a statistical measure to quantify association between  $X$  and  $Y$ . Mutual information combining with permutation test is a promising method to examine statistical dependence between variables and applied to analyzing gene expression data sets (Daub et al., 2004). Conditional mutual information

$$I(X;Y|Z) = \int dx dy dz p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)},$$

is a statistical measure to quantify association between  $X$  and  $Y$  given  $Z$  removing the effect of other variables  $Z$ . Therefore, conditional mutual information would be a more suitable measure to infer direct association than mutual information. However, estimating conditional mutual information is computationally difficult if the dimension of  $Z$  is high, and permutations are not easy to apply for examining conditional independence, in general. The proposed method enables us to examine conditional independence by permutation test and applied to inference of gene network.

A number of methods to infer gene network were reported. ARACNE (Margolin et al., 2006) and

CLR (Faith et al., 2007) are network-inference methods based on information theoretic approach. Both determine the presence of an edge between nodes by calculating the mutual information of the two nodes. ARACNE employs data-processing inequalities to eliminate weakest associations in every closed triplet of nodes. The procedure would be exact if the network structure was a tree. On the contrary, CLR compares each value of mutual information between nodes to the empirically determined distribution of all the mutual information values between pairs of nodes. CLR is designed around the assumption that the empirical distribution provides background information about the absence of edges. Gene Network Inference with Ensemble of Trees (GENIE3) (Irrthum et al., 2010) uses a tree-based method combining with bootstrap method and feature selection. TIGRESS uses least angle regression (LARS) (Haury et al., 2012), which is a sparse regression method, with stability selection. NIMEFI (Ruyssinck et al., 2014) takes an ensemble approach to combine the results of regressions with feature selection such as support vector regression and the elastic net. PLSNET (Guo et al., 2016) uses PLS-based feature selection with an ensemble technique.

## 3 METHODS

### 3.1 Approximation of Conditional Mutual Information

Suppose continuous random variables  $X, Y \in \mathfrak{R}, Z \in \mathfrak{R}^{N-2}$ . We can compute the conditional mutual information  $I(X;Y|Z)$ . The conditional expectation and standard deviation are written as

$$m_{(\cdot)}(z) = \mathbb{E}[(\cdot)|Z = z], \quad \sigma_{(\cdot)}(z) = \sqrt{\text{Var}[(\cdot)|Z = z]}.$$

We write  $X$  and  $Y$  as

$$X = m_x(Z) + \sigma_x(Z)\varepsilon, \quad Y = m_y(Z) + \sigma_y(Z)\eta, \quad (1)$$

using random variables  $\varepsilon, \eta$ .

The entropy function is defined as

$$H(X) = - \int dx p(x) \log p(x).$$

The following relations hold (see Appendix).

$$H(X + f(Z)|Z) = H(X|Z), \quad (2)$$

$$H(f(Z)X|Z) = H(X|Z) + E_Z[\log f(Z)]. \quad (3)$$

$f$  is an arbitrary continuous function and  $E_Z[(\cdot)]$  indicates expectation of  $(\cdot)$  with respect to the distribution of  $Z$ . Equations (1)-(3) yield  $I(X;Y|Z) = I(\varepsilon;\eta|Z)$ .

Furthermore, we find  $I(\varepsilon; \eta|Z) = I(\varepsilon; \eta) - I(\varepsilon; \eta; Z)$ . Finally,

$$I(X; Y|Z) = I(\varepsilon; \eta) - I(\varepsilon; \eta; Z), \quad (4)$$

results. The multivariate mutual information (McGill, 1954) is defined as

$$\begin{aligned} I(\varepsilon; \eta; Z) &= H(\varepsilon, \eta) + H(\varepsilon, Z) + H(\eta, Z) \\ &- (H(\varepsilon) + H(\eta) + H(Z) + H(\varepsilon, \eta, Z)). \end{aligned}$$

One can see that  $I(\varepsilon; \eta; Z)$  corresponds to the intersection of the uncertainties of  $\varepsilon$ ,  $\eta$  and  $Z$ . Thus, if we find values of  $\varepsilon$  and  $\eta$  that give  $I(\varepsilon; \eta; Z) = 0$ , the conditional mutual information  $I(X; Y|Z)$  is equivalent to the mutual information  $I(\varepsilon; \eta)$ . This translation of conditional mutual information into mutual information avoids any integral with respect to  $Z$ . In other words, if  $I(\varepsilon; \eta; Z) = 0$  holds, we have:

$$\varepsilon \perp \eta \Leftrightarrow X \perp Y|Z.$$

The upper bound of  $I(\varepsilon; \eta; Z)$  is

$$\begin{aligned} |I(\varepsilon; \eta; Z)| &\leq \min\{I(\varepsilon; Z), I(\eta; Z), \\ &I(\varepsilon; \eta), I(\varepsilon; Z|\eta), I(\eta; Z|\varepsilon), I(\varepsilon; \eta|Z)\}. \end{aligned} \quad (5)$$

Instead of evaluating  $I(\varepsilon; \eta; Z)$ , we evaluate sufficient condition

$$I(\varepsilon; Z) \text{ or } I(\eta; Z) = 0 \Rightarrow I(\varepsilon; \eta; Z) = 0. \quad (6)$$

Note that  $I(\varepsilon; Z)$  or  $I(\eta; Z) = 0$  is a sufficient, but not necessary, condition for  $I(\varepsilon; \eta; Z) = 0$ . In particular, if  $\varepsilon$  and  $\eta$  follow a Gaussian distribution with correlation  $\rho$  which is equivalent to partial correlation between  $X$  and  $Y$ , we have

$$I(X; Y|Z) = -\frac{1}{2} \log(1 - \rho^2).$$

### 3.2 Heteroscedastic Kernel Ridge Regression (HKRR)

We suppose that the conditional distribution of  $X$  given  $Z$  can be approximated as a Gaussian distribution,

$$p(X|Z=z) = \frac{\exp\left[-\frac{1}{2\sigma_x^2(z)}(x - m_x(z))\right]}{\sqrt{2\pi\sigma_x^2(z)}}. \quad (7)$$

Once we obtain  $m_x(z), \sigma_x(z)$  from the dataset  $\mathcal{D} \equiv \{x_i, y_i, z_i\}_{i=1}^n$  by Heteroscedastic Kernel Ridge Regression (HKRR) (Cawley et al., 2004), we have

$$\varepsilon_i = \frac{x_i - m_x(z_i)}{\sigma_x(z_i)}. \quad (8)$$

$m_x(z), \sigma_x(z)$  is inferred by solving the minimization problem of the negative log-likelihood function of (7)

$$-\log \prod_i p(x_i|z_i) = \sum_i \frac{(x_i - m_x(z_i))^2}{2\sigma_x^2(z_i)} + \log \sigma_x(z_i). \quad (9)$$

The constant term can be ignored.  $m_x(z), \log \sigma_x(z)$  are written as

$$\begin{aligned} m_x(z) &= \sum_i \alpha_i^{m_x} K^{m_x}(z, z_i) + b^{m_x}, \\ \log \sigma_x(z) &= \sum_i \alpha_i^{\sigma_x} K^{\sigma_x}(z, z_i) + b^{\sigma_x}. \end{aligned}$$

The parameters  $\alpha^{m_x}, \alpha^{\sigma_x}, b^{m_x}, b^{\sigma_x}$  and kernel matrix  $K^{(\cdot)}(z, z') = \langle \phi^{(\cdot)}(z), \phi^{(\cdot)}(z') \rangle$ , are defined by the kernel function regarding the basis function  $\phi$ . The function  $\phi$  maps the data vector  $z$  onto the high-dimensional feature space  $\mathcal{F}$ , that is,  $\phi^{(\cdot)}: \mathcal{Z} \rightarrow \mathcal{F}$ .  $\langle \cdot, \cdot \rangle$  represents the inner product between  $(\cdot)$  and  $(\cdot)$ . We can find  $\varepsilon$  by solving the minimization problem in (9) and determining the parameters  $\alpha^{m_x}, \alpha^{\sigma_x}, b^{m_x}, b^{\sigma_x}$ . The minimization problem is solved numerically with iterative solutions of subproblems that are defined as linear simultaneous equations (Cawley et al., 2004; Cawley and Talbota, 2004). In addition, the leave-one-out cross validation error of the minimization problem is obtained analytically without retraining using the Sherman-Morrison formula. Similarly,  $\eta$  is obtained by way of  $y$  instead of  $x$ .

### 3.3 Independence Measure

Using the kernel matrices,  $\{K_X\}_{ij} = K(x_i, x_j), \{K_Y\}_{ij} = K(y_i, y_j)$ , we define the measure of the mutual dependence between  $X$  and  $Y$  as

$$J(X, Y) \equiv \frac{1}{n^2} \text{Tr} \left[ K_X \left( I_n - \frac{1}{n} 1_n \right) K_Y \left( I_n - \frac{1}{n} 1_n \right) \right], \quad (10)$$

where  $I_n$  denotes the identity matrix of size  $n$  and  $1_n$  denotes the square matrix of size  $n$ , all of whose elements are 1 (Sun et al., 2007).  $J(X, Y) = 0$  means that  $X$  and  $Y$  are independent and the scale of  $J(X, Y)$  is arbitrary. Though one can use mutual information instead of  $J$  to indicate the dependence between  $X$  and  $Y$ , the calculation of  $J$  avoids the computationally costly estimation of density distributions.

### 3.4 Permutation Test

In general, tests for statistical significance are required when inferring the dependence between variables as statistical errors in the sample prevent the

measure of dependence from ever reaching zero. We employ permutation test (Edgington and Onghena, 2007) to test the hypotheses,

$H_0$  :  $X$  and  $Y$  are independent,

$H_1$  :  $X$  and  $Y$  are not independent.

In permutation tests, the expected distribution of data given the null hypothesis is generated by permutations of the dataset. When examining the degree of dependence between  $X$  and  $Y$ , the pseudo-dataset  $\{x_i, y_{I_p(i)}\}_{i=1}^n$  following the null hypothesis is generated by random permutations of  $x_i$  and  $y_i$  with the original dataset  $\{x_i, y_i\}_{i=1}^n$ .  $I_p(i)$  denotes the  $i$ -th element of the index set, which is generated by random permutation of the index set  $i \in \{1, \dots, n\}$ . We obtain the null distribution of  $J(X, Y)$  empirically by evaluating a number of  $J(X, Y)$  with iteratively generated pseudo-datasets by permutation. The percentile point  $J(X, Y)$  of the original dataset against the empirically determined null distribution is the  $p$ -value. In general, the permutation test is easy to apply when inferring the mutual dependence between  $X$  and  $Y$ . However, the permutation test is difficult to apply when trying to infer the conditional dependence between  $X$  and  $Y$  given  $Z$  when the data were compiled from a small sample. Since  $Z$  may be associated with  $X$  or  $Y$  or both, an intentional sampling technique such as grouping is needed to retain the associations between  $Z$  and  $X$  and/or  $Y$ . Even such intentional re-sampling techniques are difficult if the dataset was prepared from a small sample, and datasets used in biology tend to be compiled from small samples.

### 3.5 Reliability of $p$ -value

The standard error of independence measure  $J$  is estimated by Jackknife method

$$SE_{\text{Jack}} = \sqrt{\frac{n-1}{n} \sum_i^n (\bar{J} - J_{\setminus i})^2}, \quad (11)$$

where  $J_{\setminus i}$  is independence measure using the data set of sample size  $n-1$  removed sample  $i$ , and  $\bar{J} = \frac{1}{n} \sum_i^n J_{\setminus i}$ . We write the  $p$ -value in the permutation test for  $J = J_{(\cdot)}$  as  $p(J_{(\cdot)})$ , and define a reliability of  $p$ -value as

$$p_{\text{reli}} = p(J_{(l)}) - p(J_{(u)}), \quad (12)$$

where  $J_{(l)} = \max\{0, J_0 - SE_{\text{Jack}}\}$ ,  $J_{(u)} = J_0 + SE_{\text{Jack}}$ .  $J_0$  is the value of  $J$  calculated from the whole data set. The domain of  $p_{\text{reli}}$  is  $[0, 1]$ . The reliability of  $p$ -value decreases with an increase in  $p_{\text{reli}}$ .

---

Algorithm 1 : Let  $D$  be the  $n \times m$  matrix describing the dataset, where  $n$  is the sample size and  $m$  is the number of variables/nodes. The vector  $d_{\cdot i}$  represents the  $i$ -th column of  $D$ .

---

**for** each pair of nodes  $i$  and  $j$  **do**

$x \leftarrow d_{\cdot i}, y \leftarrow d_{\cdot j}, Z \leftarrow \{d_{\cdot k}\}_{k \in \{i, j\}}$

Infer  $m_x, \sigma_x, m_y, \sigma_y$  by HKRR for  $x$  and  $y$  on  $Z$ .

$\epsilon$  and  $\eta$  is obtained by the (8).

Compute  $J(\epsilon; \eta)$ .

Generate the null distribution of independence between  $\epsilon$  and  $\eta$  through permutations.

Obtain  $p$ -value by comparing  $J(\epsilon; \eta)$  to the null distribution.

Compute  $J(\epsilon, Z)$  and  $J(\eta, Z)$ .

Generate the null distributions of independence between  $\epsilon$  and  $Z$  and between  $\eta$  and  $Z$  through permutations.

Obtain  $p$ -values by comparing  $J(\epsilon; Z)$  and  $J(\eta; Z)$  against each of the null distributions.

Compute the reliability of  $p$ -value by the (12) and omit the low reliable  $p$ -values.

**end for**

---

## 4 ARTIFICIAL GENE NETWORK INFERENCE

The datasets from the Dream4 challenge (Schaffter et al., 2011) were used to evaluate the proposed method. The datasets are generated by simulations based on ordinary differential equations that imitate the biochemical reactions that make up a gene regulation network. We employed multifactorial datasets comprising five networks that include 100 nodes. The multifactorial dataset corresponds to steady-state measurements of all network nodes after a multifactorial perturbation. The multifactorial perturbations are given by slightly increasing or decreasing the basal activation of all genes of the network simultaneously by different random amounts. Thus, the data matrix  $D$  for each network consists of 100 rows that correspond to 100 measurements under perturbation of a single node and 100 columns that correspond to all 100 nodes of the network. We infer the presence of an edge between nodes  $i$  and  $j$  by calculating the approximating conditional mutual information between  $i$  and  $j$  given all other nodes except  $i$  and  $j$ . The data matrix is divided into three parts,  $x$ ,  $y$  and  $Z$ , where  $x$  and  $y$  are vectors consisting of the  $i$ -th and  $j$ -th columns of data matrix  $D$ , respectively, and  $Z$  is the redefined data matrix that consists of  $D$  excluding the  $i$ -th and  $j$ -th columns. The parameters  $m_x(z)$  and  $\sigma_x(z)$  are inferred by HKRR, setting  $x$  as output and  $Z$  as in-

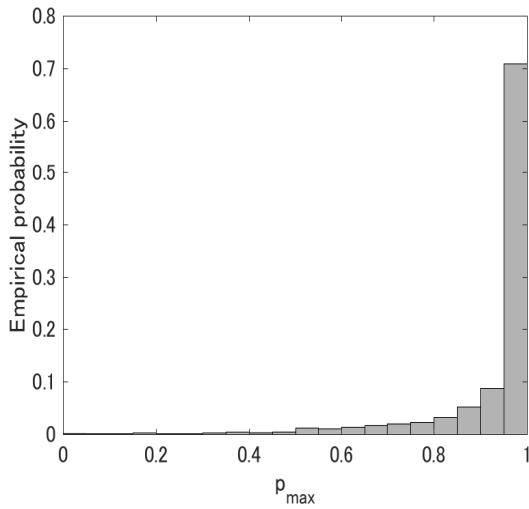


Figure 1: Empirical probability distribution of  $p_{\max}$  for all inferred edges of five networks. A total of 24750 were inferred.

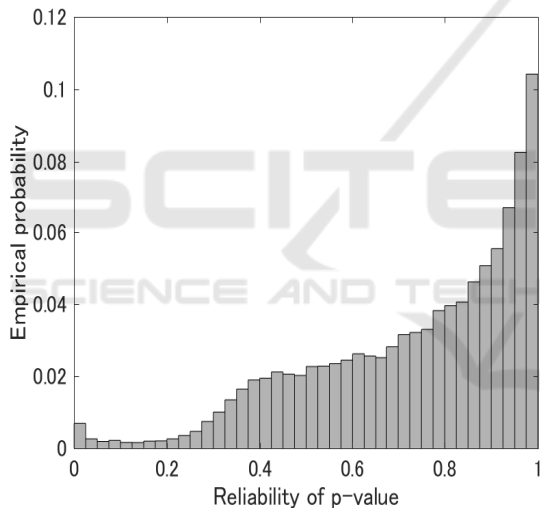


Figure 2: Empirical probability distribution of reliability of  $p$ -value for all inferred edges of five networks. A total of 24750 were inferred.

put. Similarly, by setting  $y$  as output, the parameters  $m_y(z)$  and  $\sigma_y(z)$  are inferred.  $\varepsilon$  and  $\eta$  are determined with (8). Leave-one-out training of sample  $\mu$  is used to infer  $m_x(z)$  and  $m_y(z)$ , from which  $\varepsilon_\mu$  and  $\eta_\mu$  follow analytically. The Gaussian kernel function

$$K^{(\cdot)}(x, x') = \exp(-\beta_{(\cdot)} \|x - x'\|_2^2), \quad (13)$$

is used for kernel functions  $K^{m_x}$  and  $K^{\sigma_x}$ . The same hyper parameters of HKRR are set for the  $x$  and  $y$  input models, and are determined so that the leave-one-out cross validation error, which is analytically estimated without retraining (Cawley et al., 2004), is minimized with a grid search. In the grid

search, we set the hyper parameters as  $\beta_m = [1.00e-4, 1.78e-4, 3.16e-4, 5.62e-4, 1.00e-3]$ ,  $\beta_\sigma = [1.00e-6, 3.16e-6, 1.00e-5]$ ,  $\lambda_m = [5.00e-4, 7.50e-4, 1.00e-3, 1.30e-3, 1.50e-3]$ , and  $\lambda_\sigma = [0.1, 0.55, 1]$ .  $\lambda_m$  and  $\lambda_\sigma$  are weight parameters of the  $\ell_2$  norm-regularization terms for  $m_x(z)$  and  $\sigma_x(z)$ , respectively. The Gaussian kernel function is also included in the independence measure  $J$ , and the scaling parameter is set as  $\beta_J = 0.01$ .

Permutation tests in 50000 trials are applied to check for the independence of  $\varepsilon$  and  $\eta$ . Instead of imposing the condition that  $I(\varepsilon; \eta; Z) = 0$  for  $\varepsilon$  and  $\eta$ , the sufficient condition, which is  $x$  and  $Z$  or  $y$  and  $Z$  are independent, is examined in post-processing. The term  $p_{\max}$  is defined as  $\max[p_{xZ}, p_{yZ}]$ , where  $p_{xZ}$  and  $p_{yZ}$  are the  $p$ -values of the permutation tests for the independence of  $x$  and  $Z$ , and of  $y$  and  $Z$ , respectively. The ratio of  $p_{\max} < 0.05$  to all results is almost 0.0019(Fig.1). Thus, we do not discriminate between edges that failed to fulfill the sufficient condition below. The result  $p_{\max} < 0.05$  is not equivalent to  $I(\varepsilon; \eta; Z) \neq 0$ , as the latter is a sufficient, but not necessary condition. Even if  $p_{\max} < 0.05$ ,  $I(\varepsilon; \eta; Z) = 0$  may hold. In general, there no guarantee that  $I(\varepsilon; \eta; Z) = 0$  when inferring  $\varepsilon$  and  $\eta$  using nonlinear regression models. However, we consider that the measure is sufficiently reliable if the generalization error in the data set is small enough, since most regression models assume that noise is added to the output independently. We plan to clarify this hypothesis in the future. We employed top 10% reliable  $p$ -values for network inference. Thus, the  $p$ -values which have larger  $p_{\text{reli}}$  than the value of 10 percentile point of  $p_{\text{reli}}$  were omitted. The each empirical probability distribution of  $p_{\text{reli}}$  is similar among five inferred networks (Fig.2), and that of 10 percentile point of  $p_{\text{reli}}$  was around 0.4. The point can be seen as the edge of plateau of distribution, and empirical probability rapidly decreases with an decrease in  $p_{\text{reli}}$ . The proposed method for network inference is summarized as pseudo-code in algorithm1.

We compared our proposed method against GENIE3(Irrthum et al., 2010), TIGRESS(Hauray et al., 2012), CLR(Faith et al., 2007), ARACNE(Margolin et al., 2006), NIMEFI(Ruyssinck et al., 2014), PLSNET(Guo et al., 2016). The proposed method performs better than the other methods in terms of both the average area under the receiver operating characteristic curve (AUROC) and the average area under precision recall curve (AUPR) (Fig.3), although the low reliable  $p$ -values are omitted in the performance evaluation of proposed method. This result indicates that approximate conditional independence and omitting low reliable  $p$ -value is effective for in-

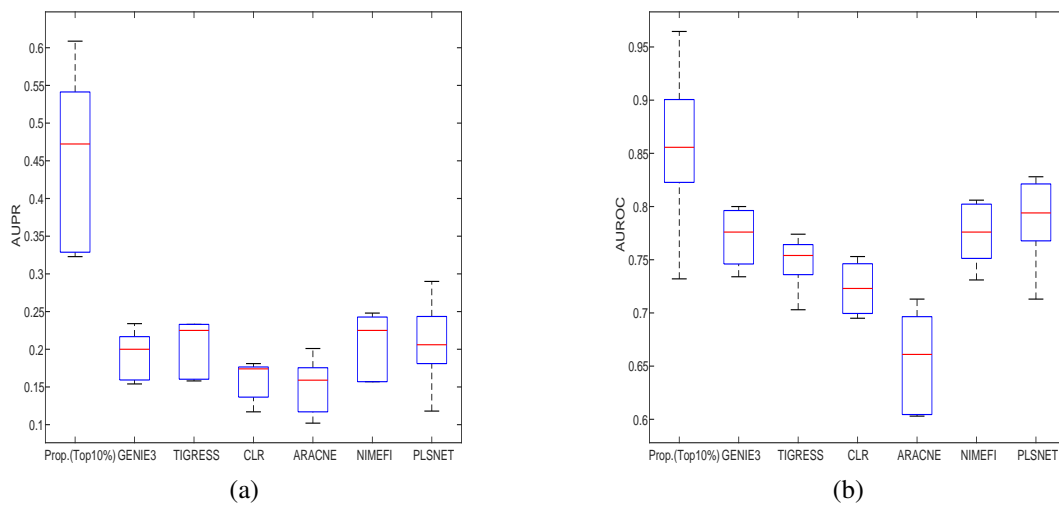


Figure 3: Boxplot of (a)AUPR and (b) AUROC of the results for five inferred networks. The results except proposed method were taken from (Guo et al., 2016)

ferring the associations of a gene network. Inferring whole network is not always needed. One of the goals of network inference for biologists is to find unknown associations in biological knowledge. Therefore, there is an advantage to infer a part of network more accurately rather than whole of that, although it is depending on the intended use. Furthermore, merging parts of network of high reliable  $p$ -values inferred by various estimators would be a promising strategy to identify whole network structure.

The AUPR almost monotonically decreases with an increase in unreliable  $p$ -values, whereas the AUROC is skewed bell-shaped curve (Fig.4, 5). In general, the precision recall (PR) curve is a useful measure, when dealing with highly skewed datasets in the class distribution (Davis and Goadrich, 2006). In other words, the PR curve offers insight about the quality of inferred association networks that are not dense, as biological networks tend to be. Given that AUPR is put more weight than AUROC, the characteristics of monotonic decrease of AUPR is useful, and the threshold of  $p_{\text{reli}}$  is determined by the degree to which accuracy is required.

The advantage of permutation testing is that it returns a  $p$ -value. AUROC and AUPR reveal potential performances of network inference, but the network structure affects how the threshold for judging the presence of an edge is determined in practice. The  $p$ -value can be considered as a normalized score defined in  $[0, 1]$ , so the scale of the measure is always constant. Additionally, the score can be easily interpreted because the probability was under a given statistical assumption that when conditional or ordinary independence is true, the conditional or ordinal

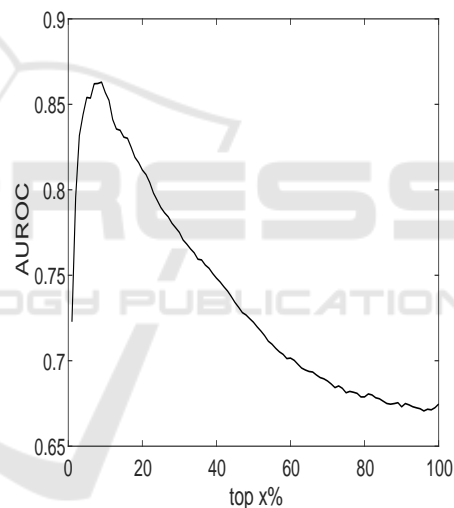


Figure 4: Average AUROC for inferred networks of reliable top  $x\%$  edges.

mutual information would be greater than or equal to observed result.

## 5 CONCLUSION

This article has proposed a novel computational method for approximating conditional mutual information based on the distribution of residuals in regression models of the data. The proposed method translates the problem of conditional independence to that of determining independence, to enable the use of permutation testing. The translation of the problem is an essential feature of the proposed method. The pro-

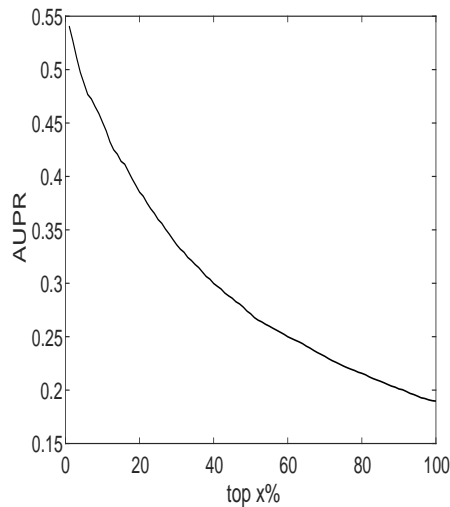


Figure 5: Average AUPR for inferred networks of reliable top  $x\%$  edges.

posed method can translate a problem to use other estimators of residuals  $\varepsilon, \eta$  and independence measures instead of HKRR and  $J$ . The score of  $p$ -value offered by the permutation test can be interpreted as the significance level, and this threshold is easier to determine than other scores such as the conditional mutual information. Additionally, the reliability of  $p$ -value is defined.

The proposed method was applied to inferring artificial gene networks from the Dream4 challenge datasets, and had the better performance in terms of AUROC and AUPR. Although the proposed method infers an part of network by omitting low reliable  $p$ -values, there an advantage to find unknown associations. Furthermore, merging parts of network of high reliable  $p$ -values inferred by various estimators would be a promising strategy to identify whole network structure.

The proposed method would be basically extended to the inference of associative networks in case that the dataset is many-dimensional, that is,  $X$  and  $Y$  can be multidimensional. This method may prove useful, for instance, in the examination of associations between layers or pathways in transomics datasets. The HKRR method is employed to infer  $\varepsilon$  and  $\eta$  in the method proposed above. However, if  $p(x|Z)$  or  $p(y|Z)$  can have a multi-modal distribution for some fixed  $Z$ , HKRR is no longer suitable. The appropriate method for inferring  $\varepsilon$  and  $\eta$  depends on the distribution of the data in question, and demands further study.

## ACKNOWLEDGEMENTS

This work was supported by the Creating information utilization platform by integrating mathematical and information sciences, and development to society, CREST (JPMJCR1912) from the Japan Science and Technology Agency (JST) and by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number (JP18H04801, JP18H02431) and Kayamori Foundation of Informational Science Advancement.

## REFERENCES

- Cawley, G. and Talbota, N. (2004). Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks*, 17:1467–1475.
- Cawley, G. C., Talbota, N., Foxalla, R. J., Dorlingb, S. R., and Mandic, D. P. (2004). Heteroscedastic kernel ridge regression. *Neurocomputing*, 57:105–124.
- Cover, T. M. and Thomas, J. (2006). *Elements of Information Theory*. Wiley-Interscience New York, NY, USA, 2nd edition.
- Daub, C. O., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using b-spline functions—an improved similarity measure for analysing gene expression data. *BMC bioinformatics*, 5(1):118.
- Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23th international conference on Machine learning (ICML'06)*, pages 233–240.
- Edgington, E. and Onghena, P. (2007). *Randomization Tests*. Chapman and Hall/CRC, 4th edition.
- Faith, J., Hayete, B., Thaden, J., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J., and Gardner, T. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, 5(1):e8.
- Guo, S., Jiang, Q., Chen, L., and Guo, D. (2016). Gene regulatory network inference using pls-based methods. *BMC Bioinformatics*, 17(545).
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer New York Inc, NY, USA.
- Hauray, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012). Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):145.
- Irrthum, A., Wehenkel, L., Geurts, P., et al. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776.
- Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., and Califano, A. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(57).

- McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, 19(2):97–116.
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press New York, NY, USA.
- Ruyssinck, J., Geurts, P., Dhaene, T., Demeester, P., Saeys, Y., et al. (2014). Nimefi: gene regulatory network inference using multiple ensemble feature importance algorithms. *PLoS One*, 9(3):e92709.
- Schaffter, T., Marbach, D., and Floreano, D. (2011). Genetweaver: in silico benchmark generation and performance profiling of network inference methods. *BIOINFORMATICS*, 27(16):2263–2270.
- Sun, X., Janzing, D., Schölkopf, B., and Fukumizu, K. (2007). A kernel-based causal learning algorithm. In *Proceedings of the 24th international conference on Machine learning (ICML'07)*, pages 855–862.

## APPENDIX

The formula for equation (2) is widely known (Cover and Thomas, 2006).

**Theorem 1.** When  $X = f(Z)\epsilon$ ,  $H(X|Z) = H(\epsilon|Z) + E_Z[\log f(Z)]$ .

*proof.* we have  $p_{X|Z}(x|z) = \frac{1}{f(z)} p_{\epsilon|Z}\left(\frac{x}{f(z)}|z\right)$ , then,  $H(X|Z = z) = H(\epsilon|Z = z) + \log f(z)$ . Thus, we obtain

$$H(X|Z) = E_Z[H(X|Z = z)] = H(\epsilon|Z) + E_Z[\log f(Z)].$$

Similarly, we obtain  $H(X, Y|Z) = H(\epsilon|Z) + E_Z[\log f(Z)g(Z)]$ .