Communications for Statistical Applications and Methods https://doi.org/10.29220/CSAM.2020.27.3.285
2020, Vol. 27, No. 3, 285–299 Print ISSN 2287-7843 / Online ISSN 2383-4757

Independence test of a continuous random variable and a discrete random variable

Jinyoung Yang[a], Mijeong Kim[1,a]

[a]Department of Statistics, Ewha Womans University, Korea

Abstract

In many cases, we are interested in identifying independence between variables. For continuous random variables, correlation coefficients are often used to describe the relationship between variables; however, correlation does not imply independence. For finite discrete random variables, we can use the Pearson chi-square test to find independency. For the mixed type of continuous and discrete random variables, we do not have a general type of independent test. In this study, we develop a independence test of a continuous random variable and a discrete random variable without assuming a specific distribution using kernel density estimation. We provide some statistical criteria to test independence under some special settings and apply the proposed independence test to Pima Indian diabetes data. Through simulations, we calculate false positive rates and true positive rates to compare the proposed test and Kolmogorov-Smirnov test.

Keywords: causation, independence test, kernel density estimation, Kolmogorov-Smirnov test

1. Introduction

We have been interested in identifying relationships among some variables in a data set. In particular, in epidemiology and social studies, researchers focus on finding a causal relationship of which variable leads to a particular result. A covariance and various kinds of correlation coefficients can be measurement tools to capture a linear relationship of two random variables, but those do not give information on causation. A regression method is mostly used to find relationships among variables; however, it implies associations and not the causal relationships between variables. Pearl *et al.* (2016) also stresses that “correlation is not causation” and suggests researchers do a different approach from finding a correlation or a regression model to identify a causal relationship among variables properly. In a causal inference study, a graph consisted of vertices and edges is used to describe causal relationship among variables effectively. Each variable is represented as a vertex and a relationship between variables is represented through an edge. If two variables are dependent then two vertices corresponding to those variables are connected through an edge. If not, those are not connected. In order to draw a graph to reflect causations among variables, Bayesian networks have been introduced. For a Bayesian network learning, an independence test is carried out in the first step of constraint-based causal structure learning such as PC algorithm (Spirtes *et al.*, 2000).

[1] Corresponding author: Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: m.kim@ewha.ac.kr

Published 31 May 2020 / journal homepage: http://csam.or.kr

For all discrete random variables, log-likelihood ratio G^2 or Pearson's X^2 can be applicable to identify independency among variables. In the case of paired continuous variables, t-test for Pearson correlation coefficients are used to validate independency among variables under the assumption of the linearity of those variables. We briefly review the test methods in Section 2. For more independence test methods, refer Table 4.1 and Table 4.2 in Scutari and Denis (2014). When continuous variables and discrete variables are mixed in data sets, R package 'deal' can be used to find the dependence structure globally through a Hill-Climbing algorithm under the assumption that the conditional continuous distribution is normal (Russell and Norvig, 2003; Scutari and Denis, 2014). The Hill-Climbing algorithm finds the structure of variables which have the highest score; however, it does not give us the local independence test results. In this paper, we propose a nonparametric method to test the independency of mixed variables of discrete and continuous variables without assuming a specific conditional distribution. The null hypothesis for the proposed test is similar to that of Kolmogorov-Smirnov test. After introducing our methods in Section 3, then we will compare our test and Kolmogorov-Smirnov test in Section 4. In Section 5, we apply our methods to real data. In Section 6, we present conclusions and directions for future research. In Appendix, some useful R functions are provided.

2. Existing independence tests

Definition 1. (Independence of random variables) *Random variables X and Y are independent if for all x and y,*

$$F_{X,Y}(x,y) = F_X(x)F_Y(y),$$

where $F_{X,Y}$, F_X, and F_Y are cumulative distribution functions.

If X and Y are both discrete, The above equation is equivalent to

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \qquad \text{for all } x \text{ and } y,$$

where $p_{X,Y}$, p_X, and p_Y are probability mass functions.

If X and Y are both continuous, The above equation is equivalent to

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \qquad \text{for all } x \text{ and } y,$$

where $f_{X,Y}$, f_X, and f_Y are probability distribution functions.

2.1. Independence test for finite discrete variables

In the following contingency table of X and Y, let π_{ij} be the joint probability for $X = x_i$ and $Y = y_j$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

		Y				
		y_1	y_2	$\cdots$	y_J	
X	x_1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
	x_2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}
		n_{+1}	n_{+2}	$\cdots$	n_{+J}	n

Then the expected frequency μ_{ij} can be estimated as $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$.

If $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$, then X and Y are independent. In order to test independency of X and Y, we can set the following hypothesis test.

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for all } i \text{ and } j, \quad H_1 : \text{Not } H_0.$$

For the independence test, Pearson (Pearson, 1900) and likelihood ratio (Neyman and Pearson, 1933) statistics are proposed as

$$X^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \quad G^2 = 2\sum n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right)$$

respectively and they follow approximately chi-squared distribution with degree of freedom $df = (I-1)(J-1)$.

2.2. Independence test for continuous variables

In general, Pearson correlation ρ does not represent the independence or dependence, but a linear relationship between two random variables. If we assume that paired two random variables are either independent or just linearly related, then Pearson correlation ρ can be used to measure independency.

$$H_0 : \rho = 0, \quad H_1 : \text{Not } H_0.$$

Let n be the number of observations.

- Exact t test for Pearson's correlation coefficient ρ

$$t = \rho\sqrt{\frac{n-2}{1-\rho^2}}$$

  is approximately distributed as $t(n-2)$ under H_0.

- Fisher's Z test

$$z = \frac{\sqrt{n-3}}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$$

  is approximately distributed as $N(0,1)$ under H_0.

3. Independence test of a continuous and a discrete random variable

Let X be a discrete random variable and Y be a continuous random variable. X and Y are independent if and only if

$$f(y|x) = f(y),$$

for all $X = x$ and $Y = y$. Although X can be either finite or infinite, we have finite number of x's in a given data set. Thus, for fixed x, we can estimate $\widehat{f(y|x)}$ using kernel density estimation if the number of y is enough for fixed x. Because x's are finite in the data set, we can estimate $\widehat{f(y|x)}$ for all x's belong to the data set. Practically, we can compare the conditional probability distribution $\widehat{f(y|x)}$ and the marginal distribution $\widehat{f(y)}$ estimated from kernel density estimation for finite x's. We generate random numbers from the distribution of X and Y independently and estimate $f(y|x)$ and $f(y)$, repeatedly. By numerically integrating the overlapped region of $\widehat{f(y|x)}$ and $\widehat{f(y)}$, we obtain the empirical distribution of overlapped region. Next, we can test the following hypothesis based on a given significant level such as $\alpha = 0.01$ and $\alpha = 0.05$.

$$H_0 : f(y|x) = f(y) \quad \text{for all } x \text{ and } y, \quad H_1 : \text{Not } H_0.$$

We use the following procedure.

1. We estimate marginal distributions for X and Y separately if we assume a parametric distribution for variables.

2. We generate random numbers of X and Y independently.

   - If we assume parametric distribution for X and Y, then we generate random numbers of X or Y based on the estimated parameters.
   - If a parametric assumption is not feasible, we can generate random numbers of X or Y using Bootstrap.

3. Estimate $\widehat{f(y|x)}$ at fixed x in the data set and $\widehat{f(y)}$.

4. Calculate the overlapped area of $\widehat{f(y|x)}$ and $\widehat{f(y)}$ from the simulated data.

5. Repeat 1–4 many times. Then we obtain the empirical distribution of the overlapped area.

6. Calculate the overlapped area of $\widehat{f(y|x)}$ and $\widehat{f(y)}$ of the data set and check whether the value is statistically significant based on the empirical distribution obtained in 5.

We carry out simulations in a few special settings. Simply, we assume that X is distributed as Bernoulli with the probability $p = P(X = 1)$ and $Y \sim N(0, 1)$. We set $p = 0.2, 0.4, 0.6, 0.8$. We generate random numbers of X and Y independently and estimate the conditional density $f(y|x)$ for $x = 0, 1$, and the marginal density $f(y)$. For kernel density estimation for $f(y)$,

$$\widehat{f(y)} = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right),$$

where K and h are called the kernel and bandwidth, respectively. In R, we used default for K and h in the function "density". It defaults a standard normal density for K and Silverman's 'rule of thumb' (Silverman, 1986) for h. For conditional density $f(y|x)$ for $x = 0, 1$, we also use the default in R function "density". According to probability theory, $f(y|x) = f(y)$ for all $x = 0, 1$ if X and Y are independent. Thus, we can expect overlapped area of $\widehat{f(y|x)}$ and $\widehat{f(y)}$ is close to 1 if the number of observations n is sufficiently large. In Figure 1, it appears that most areas of $\widehat{f(y|x)}$ and $\widehat{f(y)}$ are overlapped when $p = 0.4$ and $n = 500$. Under the setting of $X \sim$ Bernoulli($p = 0.4$), the number of 0 is more than that of 1, so that the

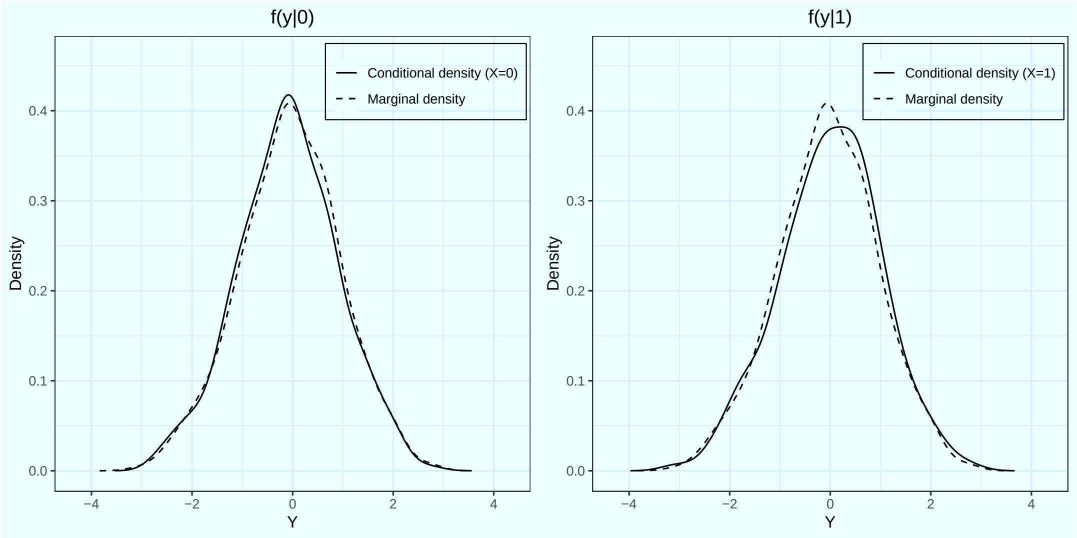

Figure 1: *Y is randomly generated from $N(0,1)$, and X is randomly generate from Bernoulli distribution with $p = 0.4$ independently with X. From obtained $n = 500$ observations, we compare estimated conditional distributions and estimated marginal distribution. In the left figure, estimated conditional density on $X = 0$ (solid line) and estimated marginal density (dotted line) are represented. In the right figure, estimated conditional density on $X = 1$ (solid line) and estimated marginal density (dotted line) are represented.*

shape of estimated conditional density on $x = 0$ is more similar to the estimated marginal distribution than that of $x = 1$. In Figure 2, most cases of overlapped area of $\widehat{f(y|x)}$ and $\widehat{f(y)}$ is very close to 1 for $x = 0, 1$ when $p = 0.4$ and $n = 500$ in 1,000 simulations. In Table 1, we represent the 1^{st} and 5^{th} percentile of overlapped area when $n = 50, 100, 500$. For smaller n, the overlapped area of $\widehat{f(y|x)}$ and $\widehat{f(y)}$ decreases. For example, in the case of $p = 0.2$, the 1^{st} percentile of overlapped area of $\widehat{f(y|1)}$ and $\widehat{f(y)}$ is 0.491 for $n = 50$ while the 1^{st} percentile of overlapped area of $\widehat{f(y|1)}$ and $\widehat{f(y)}$ is 0.877 for $n = 500$. It implies that the estimated density may not fit well if the data is sparse. Moreover, we cannot use kernel density for the number of observation is very small. Even if the total number of observations is enough, we may encounter the subset is sparse when data is filtered given $X = x$. In this respect, this method may not be applied when the number of observations of Y given x is small.

4. Comparison to Kolmogorov-Smirnov test

It is possible that the Kolmogorov-Smirnov test can be applied. The Kolmogorov-Smirnov test (Chakravart *et al.*, 1967) can be used to determine if a sample has a specific distribution. The Kolmogorov-Smirnov test works well when comparing an estimated distribution to the fully specified one; however, the Kolmogorov-Smirnov test often makes wrong decision for two unequally distributed distributions if the two distributions are both estimated. In this section, we compare rejection rate of the Kolmogorov-Smirnov test and the proposed test through simulations when the null is true and when the null is false through simulations for

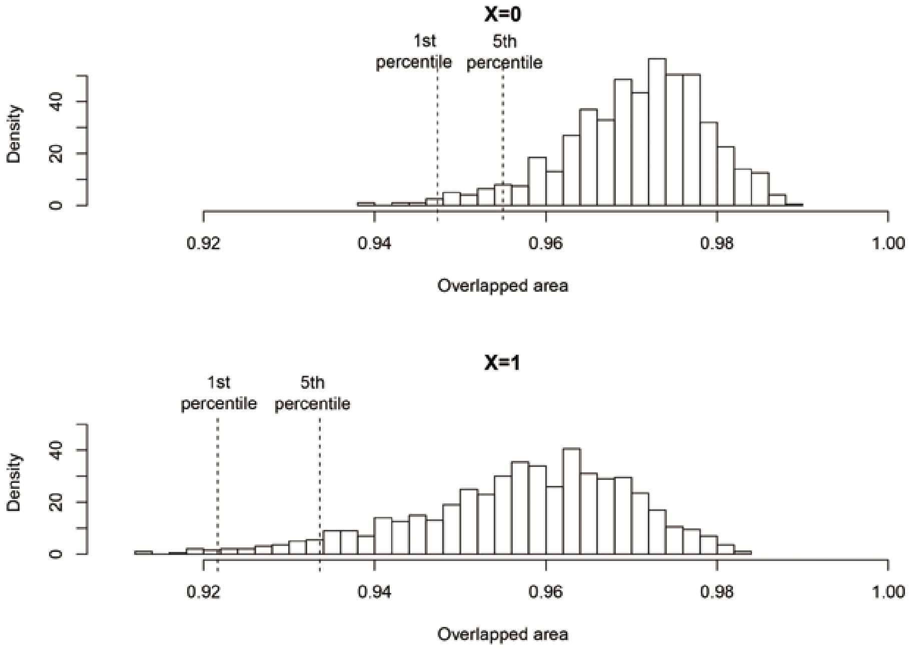

Figure 2: *Y is randomly generated from $N(0,1)$, and X is randomly generate from Bernoulli distribution with $p = 0.4$ independently with Y. From 1,000 simulations with $n = 500$ observations, we draw histograms of the overlapped area of estimated conditional density $\widehat{f(y|x)}$ and $\widehat{f(y)}$, when $X = 0$ (top) and $X = 1$ (bottom).*

Table 1: Y is randomly generated from $N(0,1)$, and X is randomly generate from Bernoulli distribution with $p = 0.2, 0.4, 0.6, 0.8$ independently with Y. We find empirical 100α percentile of overlapped area of $\widehat{f(y|x)}$ and $\widehat{f(y)}$ for $X = 0, 1$ and $\alpha = 0.01, 0.05$.

α	n	$p = 0.2$		$p = 0.4$		$p = 0.6$		$p = 0.8$	
		$x = 0$	$x = 1$	$x = 0$	$x = 1$	$x = 0$	$x = 1$	$x = 0$	$x = 1$
0.01	50	0.895	0.491	0.836	0.723	0.731	0.825	0.504	0.900
	100	0.928	0.705	0.884	0.833	0.827	0.883	0.706	0.931
	500	0.967	0.877	0.949	0.924	0.920	0.948	0.876	0.967
0.05	50	0.917	0.649	0.865	0.800	0.794	0.866	0.646	0.919
	100	0.944	0.766	0.906	0.866	0.859	0.908	0.767	0.944
	500	0.972	0.896	0.955	0.934	0.935	0.956	0.898	0.972

the following hypothesis.

$$H_0 : f(y|x) = f(y) \quad \text{for all } x \text{ and } y, \quad H_1 : \text{Not } H_0.$$

4.1. False positive rate

The rejection of a correct null hypothesis is called a Type I error or false positive in a statistical hypothesis test. A significant level is the theoretical probability of making a Type I error. That is, a significant level $\alpha = 0.05$ implies that among 100 equally distributed distribution

Table 2: False positive rates are presented for 1,000 simulations based on significant level $\alpha = 0.05$

(1) $X \sim \text{Ber}(p)$, $Y \sim N(0,1)$									
n	$p = 0.3$			$p = 0.4$			$p = 0.5$		
	cor	KS	prop	cor	KS	prop	cor	KS	prop
50	0.000	0.002	0.001	0.001	0.000	0.001	−0.001	0.000	0.000
100	0.001	0.001	0.002	0.000	0.001	0.004	0.003	0.000	0.002
200	0.000	0.001	0.007	0.001	0.000	0.011	0.002	0.000	0.012
500	0.000	0.001	0.021	0.000	0.000	0.025	0.001	0.000	0.022
(2) $X \sim \text{Ber}(p)$, $Y \sim 0.5N(0,1) + 0.5N(5,1.5)$									
n	$p = 0.3$			$p = 0.4$			$p = 0.5$		
	cor	KS	prop	cor	KS	prop	cor	KS	prop
50	0.000	0.004	0.001	0.004	0.001	0.001	0.011	0.000	0.000
100	−0.004	0.001	0.002	−0.001	0.002	0.003	0.003	0.000	0.008
200	0.000	0.001	0.008	0.001	0.000	0.014	0.000	0.000	0.010
500	0.000	0.001	0.030	−0.001	0.000	0.036	−0.001	0.001	0.039
(3) $X \sim \text{Ber}(p)$, $Y \sim SN(0,1,1)$									
n	$p = 0.3$			$p = 0.4$			$p = 0.5$		
	cor	KS	prop	cor	KS	prop	cor	KS	prop
50	−0.001	0.000	0.000	−0.002	0.000	0.001	−0.002	0.000	0.000
100	−0.002	0.002	0.003	0.004	0.000	0.009	−0.007	0.000	0.003
200	0.001	0.002	0.007	0.001	0.002	0.006	−0.003	0.000	0.007
500	0.000	0.004	0.031	0.000	0.000	0.032	0.000	0.000	0.023

For each simulation, we report the average of the correlation coefficients (cor), the rejection rate of Kolmogorov-Smirnov test (KS) and proposed test (prop).

sets, only five sets can be determined as having unequal distributions. In the simulation, we generate X and Y independently ; in addition, we conduct a Kolmogorov-Smirnov test and the proposed test under the significant level $\alpha = 0.05$ that is repeated 1,000 times in the following settings.

1. $X \sim \text{Ber}(p)$, $Y \sim N(0,1)$.

2. $X \sim \text{Ber}(p)$, $Y \sim 0.5N(0,1) + 0.5N(5,1.5)$.

3. $X \sim \text{Ber}(p)$, $Y \sim SN(0,1,1)$ (Each parameter represents location, scale, shape).

We set $p = 0.3, 0.4, 0.5$ and sample size $n = 50, 100, 200, 500$. In Table 2, we present the result. In most cases, the false positive rates of Kolmogorov-Smirnov test are less than those of the proposed test. However, we can conclude that all test results are acceptable because all false positive rates in the proposed test are less than the significant level $\alpha = 0.05$.

4.2. True positive rate

When the null hypothesis is false but the test result do not reject the null, it is called Type II error or false negative. In contrast, the test result rejects the null, it is called power. In order to compare power of the Kolmogorov-Smirnov test and the proposed test, we generate X and Y dependently and then conduct the tests under the significant level $\alpha = 0.05$ and repeat 1,000 times in the following settings.

1. $X \sim 0.5\text{Ber}(p)$, $Y \sim N(X,1)$

2. $X \sim \text{Ber}(p)$, $Y \sim N(X,1)$

Table 3: True positive rates are represented for 1,000 simulations based on significant level $\alpha = 0.05$

(1) $X \sim \text{Ber}(p)$, $Y \sim 0.5N(X,1)$									
n	$p = 0.3$			$p = 0.4$			$p = 0.5$		
	cor	KS	prop	cor	KS	prop	cor	KS	prop
50	0.225	**0.028**	**0.018**	0.243	**0.011**	**0.009**	0.239	0.005	0.006
100	0.220	0.084	0.087	0.237	0.047	0.129	0.241	0.028	0.140
200	0.222	0.327	0.430	0.240	0.215	0.539	0.244	0.121	0.562
500	0.224	0.894	0.979	0.236	0.816	0.975	0.244	0.720	0.990
(2) $X \sim \text{Ber}(p)$, $Y \sim N(X,1)$									
n	$p = 0.3$			$p = 0.4$			$p = 0.5$		
	cor	KS	prop	cor	KS	prop	cor	KS	prop
50	0.419	**0.250**	**0.068**	0.437	**0.150**	**0.105**	0.450	0.083	0.106
100	0.417	**0.727**	**0.709**	0.442	0.652	0.811	0.447	0.517	0.830
200	0.416	0.992	0.998	0.442	0.991	1.000	0.447	0.966	1.000
500	0.414	1.000	1.000	0.438	1.000	1.000	0.447	1.000	1.000
(3) $X \sim \text{Ber}(p)$, if $X = 0$, $Y \sim N(X,1)$, else $Y \sim N(5,1.5)$									
n	$p = 0.3$			$p = 0.4$			$p = 0.5$		
	cor	KS	prop	cor	KS	prop	cor	KS	prop
50	0.892	**0.800**	**0.776**	0.897	**1.000**	**0.996**	0.893	**1.000**	**0.999**
100	0.890	1.000	1.000	0.895	1.000	1.000	0.891	1.000	1.000
200	0.891	1.000	1.000	0.895	1.000	1.000	0.891	1.000	1.000
500	0.890	1.000	1.000	0.895	1.000	1.000	0.891	1.000	1.000

For each simulation, we report the average of the correlation coefficients (cor), the rejection rate of Kolmogorov-Smirnov test (KS) and proposed test (prop).

3. $X \sim \text{Ber}(p)$, $Y \sim \begin{cases} N(0,1), & \text{if } X = 0, \\ N(5,1.5), & \text{if } X = 1. \end{cases}$

We set $p = 0.3, 0.4, 0.5$ and sample size $n = 50, 100, 200, 500$. We present the simulation result in Table 3. The power improves as the correlation coefficients increase. In most cases, true positive rates of the proposed test are greater than those of Kolmogorov-Smirnov test. In particular, when $p = 0.5$, $n = 200$ in the first simulation and when $p = 0.5$, $n = 100$ in the second simulation, the power of the proposed test is better than the Kolmogorov-Smirnov test. The only results with the bold face represent the case that the true positive rates of the proposed test are less than the of Kolmogorov-Smirnov test. It is difficult to apply Kernel density estimation when we have very very few observations contained in a subset conditional on a specific $X = x$. In the first simulation in Table 3, when $p = 0.3$ and $n = 50$, around 15 observations have $X = 1$. It is hard to expect that the probability density is fitted well using only 15 observations. In that case, our method does not show better results than the Kolmogorov-Smirnov test. However, the proposed method works better if we have sufficient number of observations to apply kernel density estimation.

5. Data analysis

The structure learning procedure of Bayesian networks investigates the independence of variables and then identifies conditional independence between two variables. In this section, we identify independence of variables in Pima Indian diabetes data using the proposed nonparametric method because we cannot assume a conditional normal distribution for conditional continuous random variables. The Pima Indian diabetes data set was donated in the UCI machine learning repository. In the data set, there are no missing values of 768 observations, but some values of glucose, pressure, triceps, insulin, blood pressure and body mass index

Table 4: Variables in Pima Indian diabetes data

Variables	Distribution assumption	Estimated distribution	Description
glucose	normal	$N(122.63, 30.82^2)$	
pressure	normal	$N(70.66, 12.48^2)$	
triceps	normal	$N(29.15, 10.50^2)$	
mass	normal	$N(33.09, 7.02^2)$	
log(insulin)	normal	$N(4.81, 0.70^2)$	'insulin' is highly right skewed. A log transformation is taken.
log(pedigree)	normal	$N(-0.84, 0.63^2)$	'pedigree' is highly right skewed. A log transformation is taken.
age	negative binomial	$20 + nb(\hat{r} = 1.38, \hat{p} = 0.11)$	'age' is greater than 20 and an integer. A location transformation is taken.
pregnant	discrete	No specific distribution	Bootstrapping is used
diabetes	Bernoulli	Ber(0.33)	

are zero. In R package 'mlbench' (Leisch *et al.*, 2009), the obscure values are all substituted as missing values in the data set 'PimaIndiansDiabetes2'. Excluding those missing values, we have the number of observations $n = 392$. We describe the variables in Table 4. This data set includes both discrete and continuous variables. After log transforming highly skewed continuous variable 'insulin' and 'pedgree', we assume that continuous variables distributed as normal. The variable 'age' has an integer value and greater than 20; therefore we consider it as a location transformed negative binomial distribution. For the variable 'pregnant', we could not assume a specific distribution, so we use bootstrap when we generate random numbers for independence test. We assume a Bernoulli distribution because 'diabetes' has only two outcomes $1, 0$. According to the distribution assumption, we estimated the parameters which are given in Table 4.

We test the following three types of paired random variables.

1. A normal random variable and a normal random variable

   We carry out the Fisher's Z test for Pearson's correlation ρ among normal random variables. In R package 'pcalg' (Kalisch *et al.*, 2019), we can use the function 'condIndFisherZ' and set that no given variable is given. In the following, independence test of i^{th} and j^{th} of data set 'pima' is carried out when $\alpha = 0.01$ and $n = 392$.

   ```
   library(pcalg)
   corMatrix <- cor(pima[,1:6]);
   condIndFisherZ(i,j,NULL,corMatrix,n,0.01)
   ```

   We conclude that all paired normal random variables are not independent. The results are given in the left top part of the Table 5.

2. a discrete random variable and a discrete random variable

   For paired discrete random variables, we apply Pearson chi square test in the following way.

   ```
   chisq.test(table(age,pregnant))
   chisq.test(table(age,diabetes))
   ```

Table 5: Result of independence test

		Continuous variables						Discrete variables		
		glucose	pressure	triceps	mass	log (insulin)	log (pedigree)	age	pregnant	diabetes
Continuous variables	glucose	0	1	1	1	1	1	0	1	1
	pressure	1	0	1	1	1	1	0	1	1
	triceps	1	1	0	1	1	1	0	0	1
	mass	1	1	1	0	1	1	1	1	1
	log(insulin)	1	1	1	1	0	1	0	1	1
	log(pedigree)	1	1	1	1	1	0	1	1	1
Discrete variables	age	0	0	0	1	0	1	0	1	1
	pregnant	1	1	0	1	1	1	1	0	1
	diabetes	1	1	1	1	1	1	1	1	0

1 indicates that paired variables are not independent, while 0 indicates that paired variables are independent.

```
chisq.test(table(pregnant,diabetes))
```

We obtain that all paired discrete random variables are not independent based on $\alpha = 0.01$. Those results are given in the right bottom part of the Table 5.

3. A normal random variable and a discrete random variable (mixed type)

We have three discrete random variables, which are 'age', 'pregnant', and 'diabetes'. For diabetes = 1, there are 262 observations and for diabetes = 0, there are 130 observations. It is easy to do kernel density estimation for conditional continuous random variables for all categories of diabetes because the number of observations in each case is sufficient. However, some values of 'age' and 'pregnant' have too small number of observations to conduct kernel density estimation. It is impossible to conduct kernel density estimation when the number of observation is less than three. In practice, we perform the proposed independence test for the case that the number of observations is greater than four, when given a specific value of age or a specific value of pregnant. In Figure 3, we can see that twenty four out of a total forty two kinds for 'age' has more than five observations. There are 352 observations, which correspond to 89.80% of the total observations of 392. Similarly, we exclude observations which has the cell count of given pregnant value less than five, those are pregnant ≥ 13. Then we use 386 observations to perform independence tests, which is 98.47% of total observations. The results of independence test are given in Table 5. For example, let x and y be pregnant and triceps, respectively. We calculated the overlapped area $\widehat{f(y|x)}$ and $\widehat{f(y)}$ for used values $x = 1, 2, \ldots, 12$ in the test. For all $x = 1, 2, \ldots, 12$, all overlapped areas are greater than 1^{st} percentile based on the independently generated y and x. Thus, we concluded that the variable 'triceps' and 'pregnant' are independent.

In Appendix, we provide the R code to do independence test. First, we use 'den.adj' function to estimate density $f(y|x)$ and $f(y)$ using y's. It often happens that numerical integration of estimated density is slightly different from one. To adjust it, we calculate the numerical sum 'C' and divide the density by 'C'. The function 'alpha_percentile' finds the threshold corresponding to lower α percentile of overlapped area of $\widehat{f(y|x)}$ and $\widehat{f(y)}$ obtained from independently generated random numbers through 2,000 simulations. The function 'area_overlap_mar_cond' calculates overlapped area of estimated marginal density

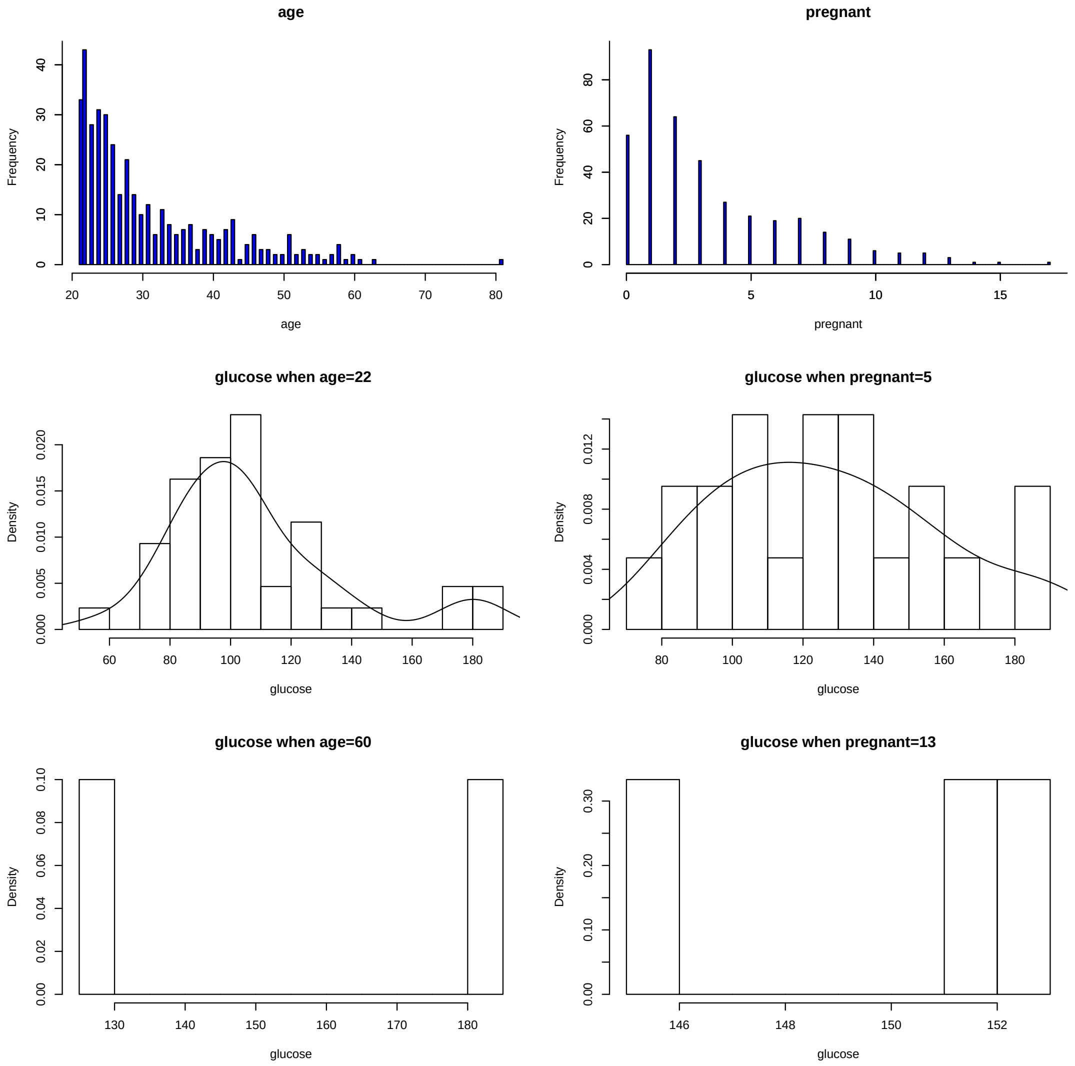

Figure 3: *Histogram of discrete variable, age and pregnant. Examples of some conditional distributions of a given age or a given pregnant value.*

and estimated conditional density of the data. The function 'indtest.cont_discrete' gives the result of independence test of a random variable and a discrete variable.

6. Discussion

We developed an independence test of a continuous random variable and a discrete random variable. For Bayesian networks, the existing methods are very restricted to analyze mixed type of variables. The R package 'bnlearn' (Scutari *et al.*, 2019) is applicable to only normal variables because that partial correlation can be used as a conditional correlation when all variables are normal (Baba *et al.*, 2014). The R package 'deal' can be used to mixed type of normal variable and discrete variable, but it assumes that conditional continuous distribution is normal. Our proposed method can be applicable to mixed type of variables without assuming any specific distribution for the conditional continuous distribution. The proposed method also has disadvantages. If the number of observations associated with some condition is very small, we cannot apply our method because it is difficult to use kernel density estimation for small number of observations. However, if more than 90% of data are available to perform our proposed method, we can expect our independence test to be practically used. Also, the proposed test is more powerful than Kolmogorov-Smirnov test for sufficient number of observations. Our method can be used for the first step of constraint-based causal structure learning such as PC algorithm (Neapolitan, 2003; Spirtes *et al.*, 2000). In PC algorithm or PC-stable algorithm (Colombo and Maathuis, 2014), first independence test of paired variables is performed. Next, those algorithm find variables are in the conditional independence relationship increasing the number of conditional variables. To find the causal structure, practical conditional independence tests are required. We are planning to examine a nonparametric conditional independence test without assuming a specific distribution for a future work.

Acknowledgements

Mijeong Kim was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2017R1C1B5015186).

Appendix: R function

```
den.adj<-function(x){
  d <- density.default(x, n = 512, cut = 3)
  xx <- d$x  ## 512 evenly spaced points on [min(x) - 3 * d$bw, max(x) + 3
        * d$bw]
  dx <- xx[2L] - xx[1L]  ## spacing / bin size
  yy <- d$y
  f <- approxfun(xx, yy)
  C <- integrate(f, min(xx), max(xx),rel.tol=.Machine$double.eps^.05)$value
  yy1<-yy/C
  return(list(x=na.omit(xx),f,C))
}

###############################################################
alpha_percentile<-function(a,b,alpha){
 ### A: Gaussian (col1-col6, "glucose","pressure","triceps","mass","l_insulin",
        "l_pedigree"),
 ### B: Discrete (col7-col9, "age","pregnant","diabetes")

  A<-dta[,a];   B<-dta[,b]
  B.name<-colnames(dta)[b]
  fit.A<-dta.fitting[[a]]
  t.B<-data.frame(table(B));
  B.unique<-unclass(as.numeric(as.character(t.B[which(t.B$Freq>=5),1])))
  m<-length(B.unique); nsim<-2000;  n<-392; sim1<-matrix(NA,nrow=nsim,ncol=m)

  for(iter in 1:nsim){
    for(j in 1:m){
      current_item<-B.unique[j]
      sample.A<-rnorm(n,mean=fit.A$estimate[1],sd=fit.A$estimate[2])
      if(b==7){
         fit.B<-dta.fitting[[b]]
         sample.B<-rnbinom(n,size=fit.B$estimate[1],mu=fit.B$estimate[2])+20}
      if(b==8){ sample.B<-sample(B,size=n,replace=T)}
      if(b==9){ sample.B<-rbinom(n,size=1,prob=dta.fitting[[9]])}

      dta<-data.frame(sample.A,sample.B)
      c.dta<-dta[sample.B==current_item,1]

      if(length(c.dta)>=3){
       A.list<-den.adj(sample.A)
       A.x<-A.list[[1]];  A.f<-A.list[[2]]; A.list.cond.B<-den.adj(c.dta)
       A.c.x<-A.list.cond.B[[1]]
       A.cond.f<-A.list.cond.B[[2]]

       xnew<-seq(max(c(min(A.x),min(A.c.x))),min(c(max(A.x),max(A.c.x))),
              length.out=512)
       df1<-A.f(xnew)/A.list[[3]]
       df2<- A.cond.f(xnew)/A.list.cond.B[[3]]
       min.df<-na.omit(sapply(1:length(df1), function(t) min(df1[t],df2[t])))
       dx<-xnew[2L]-xnew[1L]
       sim1[iter,j]<-sum(min.df)*dx
     }
   }
 }

  alpha.vector<-rep(NA,m)

  for(j in 1:m){
    temp<-na.omit(sim1[,j]);
    n1<-length(temp)
    alpha.vector[j]<-sort(temp)[round(n1*alpha)]
  }
 return(list(alpha.vector,var_name=B.unique,dta=sim1))
}

###############################################################
area_overlap_mar_cond<-function(a,b){
  A<-dta[,a];   B<-dta[,b]
  t.B<-data.frame(table(B));
  B.unique<-unclass(as.numeric(as.character(t.B[which(t.B$Freq>=5),1])))
  m<-length(B.unique)
  A.list<-den.adj(A);   A.x<-A.list[[1]];   A.f<-A.list[[2]]
  area_overlap<-rep(NA,m)

  for(j in 1:m){
    current_item<-B.unique[j]
    dta<-data.frame(A,B)
    c2.dta<-dta[B==current_item,1]
    A.list.cond.B<-den.adj(c2.dta)
    A.c.x<-A.list.cond.B[[1]]
    A.cond.f<-A.list.cond.B[[2]]
    xnew<-seq(max(c(min(A.x),min(A.c.x))),min(c(max(A.x),max(A.c.x))),length.
          out=512)
    df1<-A.f(xnew)/A.list[[3]]
    df2<- A.cond.f(xnew)/A.list.cond.B[[3]]
    min.df<-na.omit(sapply(1:length(df1), function(t) min(df1[t],df2[t]) ))
    dx<-xnew[2L]-xnew[1L]
    area_overlap[j]<-sum(min.df)*dx
  }
  return(list(area_overlap,var_name=B.unique))
}

###############################################################
indtest.cont_discrete<-function(a,b,alpha){
  threshold<-alpha_percentile(a,b,alpha)
  dta<-threshold[[3]]
  test_value<-area_overlap_mar_cond(a,b)
  diff<-test_value[[1]]-threshold[[1]]
  diff_sign<-ifelse(diff>0,0,1)
  return(list(diff_sign,var_name=test_value[[2]],dta=dta))
}
```

References

Baba K, Shibata R, and Sibuya M (2004). Partial correlation and conditional correlation as measures of conditional independence, *Australian & New Zealand Journal of Statistics*, **46**, 657–664.

Chakravarti IM, Laha RG, and Roy J (1967). *Handbook of Methods of Applied Statistics* (Vol. I), John Wiley & Sons, New York.

Colombo D and Maathuis MH (2014). Order-independent constraint-based causal structure learning, *The Journal of Machine Learning Research*, **15**, 3741–3782.

Kalisch M, Hauser A, Maechler M, *et al.* (2019). Package 'pcalg'.

Leisch F, Dimitriadou E, Leisch MF, *et al.* (2009). Package 'mlbench'.

Neapolitan RE (2004). *Learning Bayesian Networks*, Pearson Prentice Hall, Upper Saddle

River, NJ.

Neyman J and Pearson ES (1933). On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, **231**, 289–337.

Pearl J, Glymour M, and Jewell NP (2016). *Causal Inference in Statistics: A Primer*, John Wiley & Sons, Chichester.

Pearson K (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **50**, 157–175.

Russell SJ and Norvig P (2003). *Artificial Intelligence: A Modern Approach* (2nd ed), Prentice Hall, Upper Saddle River, N.J., 111–114.

Scutari M, Scutari MM, and MMPC HP (2019). Package ‘bnlearn’.

Scutari M and Denis JB (2014). *Bayesian Networks: with Examples in R*, Chapman and Hall/CRC, Boca Raton.

Silverman BW (1986). *Density Estimation*, Chapman and Hall, London.

Spirtes P, Glymour CN, Scheines R, and Heckerman D (2000). *Causation, Prediction, and Search* (2nd ed), MIT press, Cambridge, Mass.

Received October 26, 2019; Revised February 15, 2020; Accepted February 20, 2020