The Korean Communications
in Statistics Vol. 5, No. 2, 1998
pp. 411-416

A Simple Nonparametric Test of Complete Independence[1)]

Cheolyong Park[2)]

Abstract

A simple nonparametric test of complete or total independence is suggested for continuous multivariate distributions. This procedure first discretizes the original variables based on their order statistics, and then tests the hypothesis of complete independence for the resulting contingency table. Under the hypothesis of independence, the chi-squared test statistic has an asymptotic chi-squared distribution.

We present a simulation study to illustrate the accuracy in finite samples of the limiting distribution of the test statistic. We compare our method to another nonparametric test of complete independence via a simulation study. Finally, we apply our method to the residuals from a real data set.

1. Introduction

Let $\boldsymbol{Y}_i=(Y_{i1}, Y_{i2}, \ldots Y_{ip})$, $i=1,2,\ldots,n$, be a random sample from a continuous distribution function $F(\boldsymbol{y})$, $\boldsymbol{y}\in R^p$. We want to test the null hypothesis of complete or total independence

$$H_0: F(y_1, y_2, \ldots, y_p)=\prod_{j=1}^{p} F_j(y_j), \forall\ \boldsymbol{y}\in R^P, \tag{1}$$

where F_j is the j-th marginal distribution function of F. Our primary interest is on the multivariate case; i.e. $p\geq 3$. Several nonparametric tests of multivariate independence have been suggested. Some tests are based on linear rank statistics (e.g., Puri, Sen and Gokhale(1970) and Sinha and Wieand(1977) among others), on nonlinear rank statistics (e.g., Sirahata and Wakimoto(1984) among others), and on the empirical distribution or characteristic function (e.g., Blum, Kieffer, and Rosenblatt(1961) and Csörgö (1985) among others). Except for the tests based on linear rank statistics, most other tests are not so easy to compute or their limiting distributions are not easily computable. The test statistics by Puri, Sen and Gokhale(1970) essentially test pairwise independence. The test statistics by Sinha and

1) This research was in part supported by the Bisa Research Grant of Keimyung University in 1997.
2) Assistant Professor, Department of Statistics, Keimyung University, Taegu 704-701, Korea

Wieand(1977) test multivariate independence, are easy to compute for some cases, and are asymptotically normal.

The objective of this study is to suggest a nonparametric test of complete independence, which is easy to compute, is able to detect multivariate dependence (not just pairwise dependence), and has a well-known limiting distribution. This procedure first discretizes the original variables based on their order statistics and then tests the hypothesis of complete independence for the resulting contingency table. It is shown by Park(1998) that, under the null hypothesis, the chi-squared test statistic has an asymptotic chi-squared distribution. In Section 2, we will present our method and its results in detail. In Section 3, we will present a simulation study to show the accuracy in finite samples of the limiting distribution of the test statistic and compare our method to another easy-to-compute test statistic of Sinha and Wieand(1977) via a simulation study. Also we will apply our method to examine the residuals from fitting a time series model to geyser data.

2. The Method and Its Results

We let $Y=(Y_{ij})$ denote an $n\times p$ raw data matrix. The n rows of Y are a random sample from a continuous multivariate distribution. We will refer to the columns of Y as the 'coordinates' of Y.

We discretize the matrix Y to obtain an $n\times p$ matrix T whose entries T_{ij} are all integers in $\{1,2,\ldots,d\}$. The discretization is applied separately to each of the coordinates of Y. We utilize the sample quantiles of each coordinate to divide the values in each coordinate into say, d categories with approximately equal size. In other words, the discretized matrix T is given by

$$T_{ij}=k, \quad \text{if } (k-1)n/d < R_{ij} \le kn/d, \tag{2}$$

where R_{ij} is the rank of Y_{ij} among $Y_{1j}, Y_{2j}, \ldots, Y_{nj}$.

We can form a contingency table from the n rows of the discretized matrix T. This contingency table contains d^p cells corresponding to the possible p-tuples of integers in $\{1,2,\ldots,d\}$. We have n observations distributed among these d^p cells. Under the null hypothesis, the expected number of observations in any given cell is approximately equal to n/d^p (if d divides n exactly, the expectation is exactly equal to n/d^p). We use $\pi=(\pi_1,\pi_2,\ldots,\pi_p)$, with $1\le\pi_i\le d$ for all i, to denote a particular cell in our table. For each cell π, the cell count U_π is defined by

$$U_\pi=\sum_{i=1}^{n} I(T_{i\cdot}=\pi), \tag{3}$$

where $T_{i\cdot}$ is the i-th row of T. This chi-squared test statistic is defined to be

$$X^2 = \sum_{\pi} \frac{(U_{\pi} - n/d^p)^2}{n/d^p} . \tag{4}$$

The test statistic tests the hypothesis of complete independence for the contingency table of the cell counts. We note that all marginal sums of the table are fixed by the discretizing scheme and that, under the hypothesis of independence, the distribution of the cell counts is equal to the conditional distribution of a multinomial given the marginals sums equal to the fixed margins. Park(1998) has shown that, under the hypothesis, the limiting distribution of X^2 is the chi-squared distribution with $d^p - 1 - p(d-1)$ degrees of freedom.

3. Simulation Studies and an Example

In this section, we present a simulation study to illustrate the behavior in finite samples of the limiting distribution of X^2. We then compare our method to a test statistic of Sinha and Wieand(1977) via a simulation study with pairwise independence only. Finally, we apply our method to the residuals from geyser data.

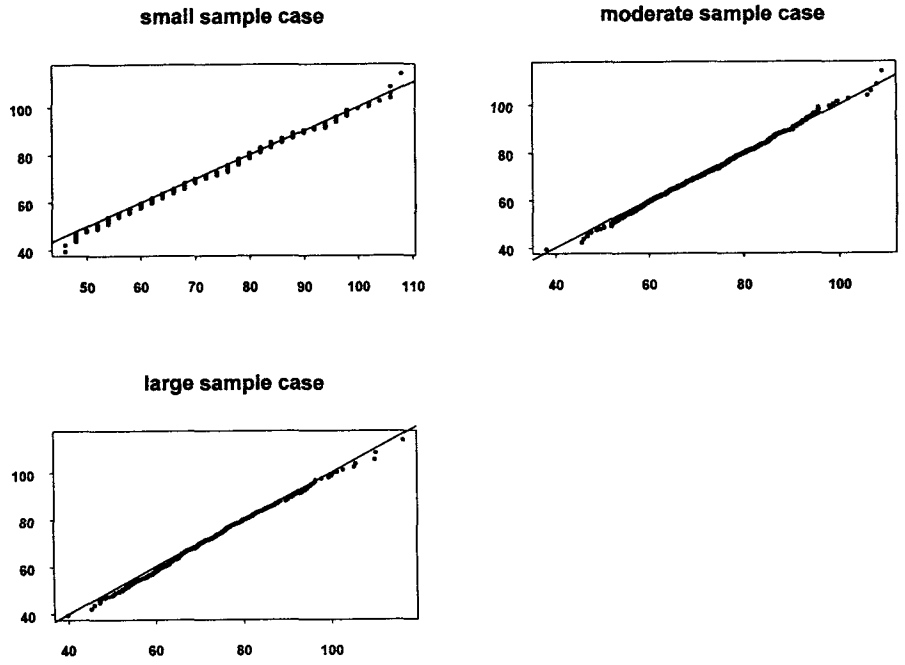
3.1 A Simulation Study on Limiting Distribution

It is well known that, under the null hypothesis, the distribution of the cell counts with all margins fixed is equal to the conditional distribution of a multinomial given that the marginal sums are equal to the fixed margins. Thus our method is nonparametric in the sense that it does not depend on the continuous parent distribution. Thus we have chosen the multivariate normal distribution to generate data sets.

To illustrate the accuracy of the limiting distribution of X^2, we take the number of variables p to be four, and the number of categories d to be three. X^2 values are computed from the contingency tables of the cell counts with $d^p = 81$ cells. The limiting distribution of X^2 is the chi-squared distribution with 72 degrees of freedom. We consider three sample sizes, $n = 81$, 405, and 810, which are denoted as small, moderate, and large samples, respectively. We note that average numbers of observations per cell are 1, 5, and 10 for small, moderate, and large samples, respectively.

For each sample size n, we generate 500 matrices Y whose entries are i.i.d. standard normal variates and calculate X^2 for each of them. The chi-squared probability plots for the three sample sizes are given in Figure 1.

Figure 1. Quantile-Quantile Plots for Small, Moderate, and Large Sample Cases



In each plot, the order statistics (horizontal axis) are plotted against the quantiles of the $\chi^2(72)$ distribution (vertical axis) with a reference line having slope 1 and intercept 0. Examining the plots, we find that all points fall quite near the reference line and so the limiting distribution is a very good approximation even for the small sample case.

3.2 A Simulation Study With Pairwise Independence Only

Since the test statistics by Sinha and Wieand(1977) are for testing multivariate independence and are easy to compute for some cases, we will compare our method to one of their test statistics when observations are pairwise independent but are not completely independent. The test statistic being compared to our method is the Spearman's version

$$S_n^{(p)} = \frac{1}{n} \sum_{j=1}^{n} \prod_{i=1}^{p} (R_{ij}/n),$$

where R_{ij} is defined in (2). This statistic is quite easy to compute and has an asymptotic normal distribution with mean 2^{-p} and variance $(3^{-p} - 4^{-p}(p+3)/3)/n$.

We generate three random variables with pairwise independence but without complete independence in the following way; generate i.i.d. standard normal random variables

X_1, X_2, X_3 and then put $Y_1 = X_1$, $Y_2 = X_2$, and $Y_3 = \text{sign}(X_1 X_2) + X_3$. We put $d = 3$ for our method and consider five sample sizes, $n = 27$, 54, 81, 108, 135. Our method has an average of 1, 2, 3, 4, and 5 observations per cell, respectively. For each n, we generate 200 matrices Y using the above scheme and compute the (asymptotic) p-values of X^2 and $S_n^{(p)}$ for each of them. The number of samples among 200 within some ranges of p-value are summarized in Table 1.

Table 1. Results of a Simulation Study with Pairwise Independence Only

range of p-value		< .05	< .01	< .005	< .001	< .0001	< .00001
$n = 27$	X^2	42	13	8	5	0	0
	$S_n^{(p)}$	93	46	30	14	5	3
$n = 54$	X^2	104	36	26	10	1	0
	$S_n^{(p)}$	104	52	40	19	7	3
$n = 81$	X^2	144	96	80	40	11	3
	$S_n^{(p)}$	100	55	42	24	8	1
$n = 108$	X^2	190	142	125	69	37	10
	$S_n^{(p)}$	125	72	57	28	10	3
$n = 135$	X^2	199	178	166	131	73	34
	$S_n^{(p)}$	138	96	74	41	12	4

Examining the table, we find that $S_n^{(p)}$ is better for small samples like $n = 24$, 54 and our method is better for large samples like $n = 81$, 108, 135. Also we find that, as the sample size becomes larger, the power of our test increases a lot whereas that of $S_n^{(p)}$ does not increase so much.

3.3 An Example Using Geyser Data

In this example, we apply our method to examine the residuals from a fitted model. If a model is correctly chosen, the residuals from the fitted model will be approximately independent. Our method can be used to examine the residuals since it detects the remaining dependence among them. There are two time series available in geyser data and the *waiting time* between eruptions is used for our example (see Azzalini and Bowman(1990) for details).

The time series plot of the waiting time reveals nothing very unusual and we try to fit a time series model to the data. We use an automatic procedure called **AR** in S-Plus to find

one of the 'best' autoregressive models. The procedure used the Akaike information criterion to choose the order of the model. We also use the Yule-Walker equations to estimate the autoregression coefficients. This procedure chooses an AR(2) model. The time series plot for the residuals from AR(2) model does not show any unusual pattern. Both autocorrelation and partial autocorrelation functions up to 25 lags are well inside the error bars.

Now we examine the residuals using our approach. We divide the residuals into subseries of three consecutive residuals and take each subseries as an observation. In this way, we obtain a 99×3 data matrix Y. Our method with $d=3$ leads to $X^2=32.73$ with (asymptotic) p-value 0.036. This p-value gives a warning signal that the residuals might have some dependence. However, the Spearman's version of Sinha and Wieand(1977) gives an (asymptotic) p-value over 0.6 and absolute values of all rank correlations are less than 0.1. Now we further examine the residuals using the lagged plots. The 'lag 1' plot shows a slight tendency that the variance of the residual e_t at time t increases with the value of the residual e_{t-1} at time $t-1$.

References

[1] Azzalini, A., and Bowman, A.W. (1990). A Look at Some data on the Old Faithful Geyser. *Applied Statistics* **39**, 357-65.

[2] Blum, J.R., Kieffer, J., and Rosenblatt, M. (1961). Distribution Free Tests of Independence Based on the Sample Distribution Function. *Annals of Mathematical Statistics* **32**, 485-98.

[3] Csörgö, S. (1985). Testing for Independence by the Empirical Characteristic Function. *Journal of Multivariate Analysis* **16**, 290-99.

[4] Park, C. (1998). The Chi-Squared Test of Independence for a Multi-way Contingency Table with All Margins Fixed. *Journal of the Korean Statistical Society*, **27**, 197-203.

[5] Puri, M.L., Sen, P.K., and Gokhale, D.V. (1970). On a Class of Rank Order Tests for Independence in Multivariate Distributions. *Sankyha Series A.* **32**, 271-98.

[6] Sinha, B.K., and Wieand, H.S. (1977). Multivariate Nonparametric Tests for Independence. *Journal of Multivariate Analysis* **7**, 572-83.

[7] Shirahata, S., and Wakimoto, K. (1984). Asymptotic Normality of a Class of Nonlinear Rank Tests for Independence. *Annals of Statistics* **12**, 1124-29.