The Korean Communications
in Statistics Vol. 5, No. 1, 1998
pp. 25-33

Variable Selection Based on Direction Vectors[1)]

Kyungmee Choi[2)]

Abstract

We review a multivariate version of Kendall's tau based on direction vectors of observations. And with this statistic we propose an analog of the forward variable selection method which selects a set of independent variables for further studies to build the eventual predicting model. This method does not assume the distributions of observations and the linear model and it is strong to the outliers with high asymptotic efficiencies relative to the parametric Pearson's correlation coefficient.

1. Introduction

Let the paired data be $(X_1, Y_1), \cdots, (X_n, Y_n)$, where $X_i \in R^p$ and $Y_i \in R^q$ for $p \geq 1$ and $q \geq 1$, and $i = 1, \cdots, n$ and let X be an $n \times p$ matrix with X_i's as its rows and Y be an $n \times q$ matrix with Y_i's as its columns. Often we want to know if the X and Y are correlated and also we try to find a set of independent variables which best explains the variation among the response variables. Then we build a predicting model for further studies. In the classical parametric Regression Analysis, we assume that the underlying distribution of observations follows the normal. And assuming the linear model, we have used the sum of squares due to regression, say R^2, and the partial F-statistic to test the significance of X on Y.

Without much distributional assumptions Kendall's τ is one of the popular statistics used to measure the correlation of two variables when both X and Y are univariate. Kendall's τ is given by

$$(1) \quad \tau = \frac{2}{\binom{n}{2}} \sum\sum_{i<j} SGN(X_i - X_j) \cdot SGN(Y_i - Y_j) - 1,$$

where SGN(z)=-1 (0) (1) if $z < 0$ ($z=0$) ($z>0$). If all pairs are concordant, Kendall's τ equals 1.0. If all pairs are discordant, the value is -1.0. If it is nearly zero, then like a Pearson's correlation coefficient, it is interpreted to be no monotonic relationship between two variables. If the given observations are independent and identically distributed bivariate random

1) This paper was supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1996.
2) Assistant Professor, Department of Mathematics, Hongik University, Chochiwon, ChungNam, 339-800, Korea.

variables, its asymptotic relative efficiency is $(3/\pi)^2 = .912$, relative to the parametric test that uses Pearson's r as a test statistic (Conover, 1980). Also its robustness is better than the parametric one (Choi, 1995).

In section 2 we are going to review a multivariate version of Kendall's tau introduced by Choi and Marden (1997), which extends the univariate one to a multivariate situation where $X_i \in R^p$ and $Y_i \in R^q$, employing a multivariate ranks defined by Small (1990) and Chaudhury (1996). Then with this statistic, in section 3, we will propose a variable selection method as an analog of the forward variable selection method in the parametric multiple regression. By simulations and examples in section 4 and 5, we can see that this method does not require either distributional or modeling assumptions, but has good robustness and performances.

2. Test Statistics

Let us first define a direction vector $d(z)$ of z by $d(z) = z / \| z \|$. Substituting $d(X_i - X_j)$ and $d(Y_i - Y_j)$ for all the SGN's in Kendall's τ we have a $p \times q$ matrix $\widehat{\tau_M}$ given by

(2) $$\widehat{\tau_M} = \frac{1}{\binom{n}{2}} \sum\sum_{i<j} d(X_i - X_j) d(Y_i - Y_j)^T .$$

Let U be a $p \times \binom{n}{2}$ matrix and V be a $\binom{n}{2} \times q$ matrix which are defined by

(3) $U = d(X_i - X_j)_{1 \le i < j \le n}$ and $V^T = d(Y_i - Y_j)_{1 \le i < j \le n}$.

The indices for both columns of U and rows of V are pairs (ij), and are ordered as follows:

(4) $(12), (13), \cdots, (1n), (23), (24), \cdots, (2n), \cdots, (n-1\ n)$.

Thus

(5) $$\widehat{\tau_M} = \frac{1}{\binom{n}{2}} UV .$$

To study its properties, it is easier to deal with a vector instead of the matrix. So now we define $\hat{\tau}$ as a $pq \times 1$ vector formed by stacking columns of $\widehat{\tau_M}$ as follows:

(6) $$\hat{\tau} = \frac{1}{\binom{n}{2}} \begin{bmatrix} UV^1 \\ UV^2 \\ \cdots \\ UV^q \end{bmatrix},$$

where $V^1, \cdots, V^q$ are columns of the matrix V. Note that all the components in the statistic

$\hat{\tau}$ are bounded since they are averages of vectors of unit length.

To find the covariance matrix of $\hat{\tau}$, let us define $A\otimes B$, the Kronecker product of $r\times s$ matrix A and $u\times v$ matrix B, given by

$$(7)\quad A\otimes B\equiv\begin{bmatrix} b_{11}A & b_{12}A & \cdots & b_{1v}A \\ & \cdots & & \\ b_{u1}A & b_{u2}A & \cdots & b_{uv}A \end{bmatrix}.$$

Then from the usual U-statistic theory, we can get the covariance matrix of $\hat{\tau}$ such as

$$(8)\quad n\Sigma_{\hat{\tau}}=4\Sigma_X\otimes\Sigma_Y$$

for i.i.d. Z_1, Z_2, and Z_3, and Σ_Z is defined by $\Sigma_Z\equiv E\left[d(Z_1-Z_2)d(Z_1-Z_3)^T\right]$. So for any consistent and scale-equivariant estimator of the covariance matrix $\widehat{\Sigma}_{\hat{\tau}}$, its asymptotic distribution under the null hypothesis is

$$(9)\quad W_n\equiv\hat{\tau}^T\widehat{\Sigma}_{\hat{\tau}}^{-1}\hat{\tau}\rightarrow\chi^2_{pq}.$$

We here suggest a useful estimator of the covariance matrix as a block matrix given by

$$(10)\quad \widehat{\Sigma}_{\hat{\tau}}=(\widehat{\Sigma}^{rs}_{\hat{\tau}})_{r,s=1,\cdots,q}.$$

Here for $d(z)_k$ as the k^{th} element of the direction vector $d(z)$, the $p\times p$ matrix $\widehat{\Sigma}^{rs}_{\hat{\tau}}$, located at the r^{th} block in the row and the s^{th} block in the column, is

$$\begin{aligned}\binom{n}{2}^2\widehat{\Sigma}^{rs}_{\hat{\tau}}=&\sum_{i=1}^{n-1}\sum_{j\neq i=i+1}^{n}d(X_i-X_j)d(X_i-X_l)^Td(Y_i-Y_j)_r\,d(Y_i-Y_l)_s\\&+\sum_{j=2}^{n}\sum_{i\neq k=1}^{j-1}d(X_i-X_j)d(X_k-X_j)^Td(Y_i-Y_j)_r\,d(Y_k-Y_j)_s\\&+\sum_{j=2}^{n-1}\sum_{i=1}^{j-1}\sum_{i\neq k=j+1}^{n}d(X_i-X_j)d(X_k-X_j)^Td(Y_i-Y_j)_r\,d(Y_k-Y_j)_s\\&+\sum_{j=2}^{n}\sum_{j=i+1}^{n}\sum_{j\neq k=1}^{i-1}d(X_i-X_j)d(X_i-X_k)^Td(Y_i-Y_j)_r\,d(Y_i-Y_k)_s\end{aligned}\quad(11)$$

For this statistic $\hat{\tau}$, we can consider two different models : fixed X or random X. But both models give the same asymptotic efficiencies relative to the parametric tests in the end (Choi and Marden, 1997). That is, its asymptotic efficiencies relative to the parametric multiple and multivariate test statistics are bigger than 1 when both distributions of X and Y are heavy-tailed. When both distributions of X and Y are light-tailed or normal, they are less than 1, but they are still quite competitive. This W_n is orthogonal transformation invariant and is pretty robust to outliers. Also it can test the significance of the relationship between X and Y without assuming normality or linearity.

3. A Variable Selection Procedure

In this section, we propose to use W_n as the criterion to select a set of variables which

explains the response variable Y best when the distributions of observations are unknown, skewed, or heavy-tailed. Also when there are a huge set of explaratory variables, W_n can be a good statistic to be used to select predictors for further studies to build a final model because of its computational simplicity. The suggested process of variable selection is an analog of the well-known parametric forward variable selection method. Before going any further let us define some matrices which we are going to use. They are:

(12) $X^{P(\{k,\cdots,K\})}$: a partitioned matrix formed by k, ···, K th columns (or variables) of X,

(13) $X_i^{P(\{k,\cdots,K\})}$: the ith row of the previous patitioned matrix,

(14) $\widehat{\tau}_M(X^{P(\{k,\cdots,K\})}, Y) = \frac{1}{\binom{n}{2}} \sum\sum_{i<j} d(X_i^{P(\{k,\cdots,K\})} - X_j^{P(\{k,\cdots,K\})}) d(Y_i - Y_j)^T$:

a matrix of Kendall's τ using the previous partioned matrix, and

(15) $W_n(X^{P(\{k,\cdots,K\})}, Y)$ is the test statistic based on the previous partitioned matrix.

Often we know the response variables of interest. So the problem is to find a set of independent variables which are most highly correlated with the pre-determined response variables. Therefore, we fix the dimension of Y at q and try to reduce that of X. However it could be extended to a method of variable selection which chooses a set of X and a set of Y simultaneously without fixing any dimensions beforehand. Let us see how we can develop an analog of the forward variable selection method with a fixed q as follows:

STEP 1. Select a column of X which shows the highest correlation with Y. That is, for k=1, ···,p, select the k^{th} column which has the biggest $W_n(X^{P(\{k\})}, Y)$. Let this k be r. Then test if $W_n(X^{P(\{r\})}, Y)$ is significant with a certain significance level α. If so, let $L_1 = \{r\}$ and go to STEP 2, where L_1 is the set of indices of significant independent variables. If not, we stop this process and conclude that there is no significant explaratory variables.

STEP 2. When $i \geq 2$, for every k not in L_{i-1} get $L^* = L_{i-1} \bigcup \{k\}$ and calculate $W_n(X^{P(L^*)}, Y)$'s. Among these we look for k such that $|\Delta W_n| = |W_n(X^{P(L^*)}, Y) - W_n(X^{P(L_{i-1})}, Y)|$ is the biggest. Then let k be s and test if $W_n(X^{P(L_{i-1} \bigcup \{s\})}, Y)$ is significant. If so, let $L_i = L_{i-1} \bigcup \{s\}$, L_i be L_{i-1}, and repeat STEP 2. Otherwise stop this process with L_{i-1} as an index set of selected independent variables. We repeat this step until there is no more significant variables to be included.

From (eq. 6) we know that $W_n(X^{P(L)}, Y)$ in STEP 1 is asymptotically distributed as $\chi^2_{\gamma q}$ if the partitioned matrix has γ columns in it. One thing to note in STEP 2 is that we use $|\Delta W_n|$ instead of ΔW_n because ΔW_n does not follow any particular distribution, but $|\Delta W_n|$ does so. We are going to discuss it in the next section by some simulation studies.

4. Simulations

In this section we are going to look for the asymptotics of ΔW_n and $|\Delta W_n|$ by some simulations. For computational simplicity, we work with one response variable and four independent variables.

We simulate 100 sets of 35 independent observations of (Y, X_1, X_2, X_3, X_4), where each of five variables are independent and identically distributed as (i) a standard normal and (ii) a Cauchy. Then we calculated 100 W_n's and got 100 values for each difference:

(15) $\Delta W_n^{12/1} = W_n(Y, X_1, X_2) - W_n(Y, X_1)$,

(16) $\Delta W_n^{123/12} = W_n(Y, X_1, X_2, X_3) - W_n(Y, X_1, X_2)$,

(17) $\Delta W_n^{1234/123} = W_n(Y, X_1, X_2, X_3, X_4) - W_n(Y, X_1, X_2, X_3)$,

(18) $\Delta W_n^{123/1} = W_n(Y, X_1, X_2, X_3) - W_n(Y, X_1)$,

(19) $\Delta W_n^{1234/12} = W_n(Y, X_1, X_2, X_3, X_4) - W_n(Y, X_1, X_2)$,

(20) $\Delta W_n^{1234/1} = W_n(Y, X_1, X_2, X_3, X_4) - W_n(Y, X_1)$.

When the distributions are normal or cauchy, QQ-plots and histograms of these statistics do not give any implication that these differences would follow a certain distribution. We also see that the ratio of any two of these statistics divided by their corresponding degrees of freedoms do not follow any particular distributions. However QQ-plots of $|\Delta W_n|$ show that these absolute differences follow some chi-square distributions with proper degrees of freedom. That is, $|\Delta W_n^{12/1}|$ is χ^2_1, $|\Delta W_n^{123/12}|$ is χ^2_1, $|\Delta W_n^{1234/123}|$ is χ^2_1, $|\Delta W_n^{123/1}|$ is χ^2_2, $|\Delta W_n^{1234/12}|$ is χ^2_2, and $|\Delta W_n^{1234/1}|$ is χ^2_3. Thus we are going to use absolute differences $|\Delta W_n|$ in the place of the partial F-statistic of the aprametric regression analysis to test wheather we will include the selected variable in the model or not and conduct the variable selection process in the next section.

5. Examples

In this section we are going to use two sets of data. One is simulated and the other is Moore's data (Moore, 1975 ; Weisberg, 1985). Both examples show that the proposed statistics based on direction vectors lead us to select independent variables just like the parametric one. Note here that the application of W_n to observations does not require either distributional or modeling assumptions.

Let us begin with some definitions we are going to use. Let *C(a:b)* be *(b-a)* integers from *a* to *b* and *rnorm(n)* be n random values from the standard normal distribution. And the significance level for the inclusion of a variable in the model is taken to be 0.25 throughout this section.

Example 1. Let us simulate 30 observations of (Y, X_1, X_2, X_3, X_4) in the following way. X_1 is from the standard normal distribution, X_2 is C(-15:14) + 0.01* rnorm(30), X_3 is cos(C(1:30)) + 0.01* rnorm(30), X_4 is 0.2 * C(1:30)^1.2 + 0.01 * rnorm(30), and Y is calculated by $X_1 + X_2 + X_3^2 + \exp(X_4)$. Then we obtain both W_n's and R^2 as in the following Table 1.

The third column contains test statistics related to a multivariate version of Kendall's tau. We first select X_2 or X_4 since they are significant and the same. Whichever we choose as the first variable to be included, the next variable for the inclusion is going to be the other one since $W_n(X_2, X_4)$ is the biggest among models with two variables. The variable selection procedure with the tau will stop at this step since $|\Delta W_n| = |W(X_1, X_2, X_4) - W(X_2, X_4)|$ is not big enough to include X_1.

From the fourth column of the Table 1 we select X_4 as the first variable to be included. Then X_2 is selected which shows the highest R^2 among models with X_2 and the second candidate. Also the partial F-statistic of the least square fit shows that X_2 is significant. The variable selection procedure stops here again because $R^2(X_1, X_2, X_4)$ is the biggest among other models of three variables, two of them being X_2 and X_4, but X_1 is not significant anymore.

Therefore in Example 1 we can conclude that both methods provide the same set of predicting variables. One interesting thing to note here is that $W_n(X_2, X_4)$ is not necessarily greater than either $W_n(X_2)$ or $W_n(X_4)$, which means that W_n is not monotonic in terms of number of variables.

The next example is about the real data analyzed by Moore (1975) and Weisberg (1985). This example shows that the monotonic translation of Y is not necessary when we use the multivariate Kendall's tau because the monotonic transformation does not change the signs of

$(Y_i - Y_j)$'s.

<Table 1> W_n and R^2 with (Y, X_1, X_2, X_3, X_4)

STEP	MODEL	W_n	R^2
1	1	0.00286472	0.0001
	2	59.6782 *	0.3737
	3	0.000318302	0.0112
	4	59.6782 *	0.4102
2	12	57.6834	0.3740
	13	0.0471743	0.0112
	14	53.5555	0.4106
	23	57.3149	0.3812
	24	59.6641 *	0.6331 *
	34	52.3736	0.4168
3	123	55.8287	0.3821
	124	57.8938	0.6365
	134	49.3569	0.4179
	234	57.5747	0.6340
4	1234	56.2104	0.6379

Note : The model with * is the best among others in each step and significant at the given significance level.

Example 2. An experiment was conducted to model oxygen uptake (O2UP) in milligrams of oxygen per minute. There are five chemical measurements to be used as five independent variables : biological oxygen demand (BOD), total Kjeldahl nitrogen (TKN), total solids (TS), total volatile solids (TVS), and chemical oxygen demand (COD), each measured in milligrams per liter (Moore, 1975). 20 values were observed. With Log(O2UP) as the response variable, Weisberg (1985) calculated several statistics including R^2 to find a proper set of variables that should be further studied with the eventual goal of building a prediction model. We are going to calcualte the multivariate Kenall's tau with the original O2UP as the response variable and compare two results. As mentioned before in Example 1, values of Kendall's tau does not change by the Log transformation. The following Table 2 contains W_n and R^2.

From the Table 2, TS is the first variable selected since $W_n(TS) = 16.9779$ which is greater than $\chi^2_{1.0.25}$. With TS in the model, COD is the second variable of interest because

$W_n(TS, COD)$ is the biggest among statistics of two-variable models, one of them being TS, and $|W_n(TS, COD) - W_n(TS)| > \chi^2_{1,0.25}$. Then our selection procedure stops here since $|W(TS, TVS, COD) - W(TS, COD)|$ is not significant enough to include BOD as the third predictor. So we stop with TS and COD as variables for the further study. Using R^2, we conduct the usual variable selection process and stop also with two variables TS and COD. Both methods select the same predictors. An additional merit of using Kendall's tau is that we do not need to transform the response variable.

<Table 2> W_n and R^2

STEP	MODEL	W_n	R^2
1	BOD	14.6878	.598
	TKN	0.464211	.008
	TS	16.9779 *	.697
	TVS	16.7292	.505
	COD	14.4095	.693
2	BOD, TS	17.3013	.716
	TKN, TS	17.0708	.710
	TS, TVS	18.3978	.712
	TS, COD	18.6159 *	.786 *
3	BOD, TS, COD	18.7208	.790
	TKN, TS, COD	18.6987	.805
	TS, TVS, COD	18.9302	.790

Note: The model with * is the best among others in each step and significant at the given significance level.

6. Conclusion.

The conventional variable selection method uses the sum of squares due to regression to select the best candidate for inclusion in the model and partial F-statistic to test its significance. But to use these we have to assume the normality of observations and the linear relationship between independent variables and response variables. Also they are well known to be sensative to the outliers. So we suggested to use a multivariate version of Kendall's tau which uses only the directions of vectors. Then the variable selection process with it can be used when distributions of observations are unknown, skewed, or heavy-tailed without any modeling assumptions. Also because of its calculational simplicity and speed, we can conduct the variable selection process in the super large data set. From the examples we have studied so far, we also see that this statistic stays the same after the monotonic trasformations of X or Y.

However the asymptotics of differences of W_n's are not simple to figure out in theory. So at this point we greatly rely on the simulation studies. Also a method to select both independent and response variables simultaneously is left for the future study.

References

[1] Chaudhury, P. (1996) On a geometric notion of quanltiles for multivariate data, *Journal of the American Statistical Association*, 91, 862-872.

[2] Choi, K. (1995). *Nonparametric Multivariate Multisample Tests for the Location Problem and Multivariate Regression Based on Directions of Data*, Unpublished Ph.D Dissertation, University of Illinois at Urbana - Champaign.

[3] Choi, K. and Marden, J. (1997). A multivariate version of Kendall's tau, a manuscript in revision.

[4] Conover W. J. (1980). *Practical Nonparametric Statistics*, 2nd ed., John Wiley and Sons, Inc.

[5] Moore, J. A. (1975). *Total Biochemical Oxygen Demand of Animal Manures*, Unpublished dissertation, University of Minnesota.

[6] Small, C. G. (1990). A survey of Multidimensional medians, *International Statistical Review*, 58,263-277.

[7] Weisberg, S. (1985). *Applied Linear Regression*, John Wiley and Sons, Inc.