An extended study to measure dependence with grouped-ordinal variables generated by unobserved non-Normal variables

Emanuela Raffinetti
Department of Economics, Management and Quantitative Methods
University of Milan (Italy)
E-mail: *emanuela.raffinetti@unimi.it*

Abstract

In the last decades, the availability of attitudinal surveys generating data of ordinal (discrete) nature has increasingly risen. Such kind of data may be also associated with responses expressed through grouped-continuous scales. This paper proposes the use of a recent new dependence measure, called MDC_{go}, suitable to all the scenarios where the independent variable is ordinal and the dependent variable is "grouped" into classes. The promising results of the MDC_{go} coefficient behavior in the case of normally and t-Student distributed variables lead us to extend the investigation to the non-normally distributed variables. A Monte Carlo simulation study is built with the aim of assessing the performance of the MDC_{go} coefficient in comparison with the most common dependence coefficients. Additional evidence on the effectiveness of the MDC_{go} coefficient arises from a real application to data on hearth diseases.

Keywords: dependency, ordinal variable, "grouped" variables, unobserved non-normally distributed variables.

1 Introduction

In applied contexts, the assessment of the dependency relationships between variables is typically led by the use of the traditional correlation coefficients, such as the the Pearson's r (see e.g., Pearson 1907), Spearman's r_S (see e.g., Spearman 1904) and Kendall's τ_b (see e.g., Kendall 1938) coefficients. Specifically, the Pearson's correlation coefficient measures the strength of the linear dependence relationships between the variables and, therefore, it is more appropriate if the variable joint distribution is Normal. In the case of variables expressed according to ordered categories, the most suitable measures to evaluate the dependence relationships between the variables are the Spearman's r_S and Kendall's τ_b correlation coefficients.

Currently, the attitudinal surveys appear as the most remarkable ordinal data sources. In many frameworks, the survey scale may be also characterised by variables which belong to certain

classes on a continuous scale. The crucial concern of grouped data relates to the measurement process, since the point value of the variable is not observed. This concern was overcome by the proposal of Raffinetti and Aimar (2019), who developed a new measure, called MDC_{go} (Monotonic Dependence Coefficient for grouped-ordinal data), with attractive features which allow to deal with both ordinal and grouped variables. The analysis of ordinal and grouped variables is connected to the nature of the underlying distribution of the unobserved continuous variables having generated them. Raffinetti and Aimar (2019) resorted to an intensive Monte Carlo simulation study to assess the behavior of the MDC_{go} coefficient in the case of grouped-ordinal variables generated by normally and t-Student unobserved distributed variables. The obtained results highlighted the superiority of the MDC_{go} coefficient in preserving the original dependency information when one variable is discretized and the other variable is grouped. The MDC_{go} coefficient was compared with the Spearman's and Kendall's correlation coefficients, sharing the property of measuring any monotonic dependence relationship. In addition, the comparison was extended by including also the Somers' coefficient (see e.g., Somers, 1962) which, beyond being a monotonic dependence measure for ordinal variables, it appears as an asymmetric index in the sense that it requires, as well as the MDC_{go} coefficient, that the role of the two variables is specified in independent and dependent variables. As discussed by Raffinetti and Aimar (2019), the proposed MDC_{go} coefficient presents as a reformulation of a recent dependence index, called MDC (Monotonic Dependence Coefficient), introduced by Ferrari and Raffinetti (2015) as a dependence measure for continuous and ordinal/tied data.

In this paper we extend the study on the monotonic dependence relationship measurement process by resorting to a Monte Carlo simulation based on samples from bivariate non-Normal distributions for which specific values for the set of parameters (pairwise correlation coefficient, kurtosis and skewness) are fixed. The findings on simulated data, in terms of estimates for the MDC_{go}, Spearman's, Kendall's and Somers' coefficients, are validated through the stochastic dominance Dunn's test and result also coherent with those obtained on actual data.

The present contribution further boosts the implications of the MDC_{go} coefficient from both the theoretical and applied viewpoint. In the former case, the MDC_{go} coefficient appears as an attractive tool for the assessment of the bivariate monotonic dependence relationships when analyzing variables which are not directly measurable, since expressed through subjective scales or, even

if not necessarily unmeasurable, restrict information into classes. Likewise, the MDC_{go} coefficient has an added value in applied research, since it provides a concrete response to a frequent concern in social, psychometric and attitude measurement sciences which typically deal with abstract concepts or non-observable constructs.

The paper is organized as follows. In Section 2, both an overview of the original MDC_{go} coefficient and its re-formalization when dealing with grouped-ordinal data organized in two-way contingency tables are provided. In Section 3, the Monte Carlo simulation results are presented and discussed. In Section 4, an application to heart disease data is illustrated. Section 5 ends the paper with final summarizing comments.

2 Background and proposal

The purpose of the paper is to propose the employment of the MDC_{go} coefficient to assess the strength of the existing monotonic dependence relationships[1] between independent grouped and dependent ordinal variables generated by unobserved underlying non-Normal distributions. As in Raffinetti and Aimar (2019), the MDC_{go} behavior is evaluated in comparison with that associated with the Spearman's r_S, Kendall's τ_b and Somers' Δ coefficients. In the following subsections, a brief overview of these coefficients is provided together with an alternative approach of the MDC_{go} coefficient formalization when dealing with data organized in two-way contingency tables.

2.1 The MDC_{go} coefficient competitors

Spearman's r_S coefficient

The Spearman's r_S correlation coefficient is the most commonly used dependence measure in presence of variables with ordinal nature. As ordinal variables are typically expressed by re-sorting to arbitrary labels based on Likert-type scales (see, Likert 1932 and Norman 2010), the subjectivity issue involved into the assignment of the value to the categories is overcome by the employment of the rank tools. The Spearman's correlation coefficient is computed as the usual

[1] Given two variables Y and X, from the notion of monotonic dependence relationship, it derives that Y and X are linked according to a monotonic functional relationship up to a random noise, i.e. $Y = f(X) + \varepsilon$, with $\varepsilon \sim N(0, \sigma^2)$. In the case of a perfect monotonically dependence relationship between Y and X, $\sigma^2 = 0$.

Pearson's correlation coefficient, by converting the ordered categories into ranks, i.e.

$$r_S = \frac{\sum_{i=1}^{n}(r(x_i)-\bar{r}(x))(r(y_i)-\bar{r}(y))}{\sqrt{\sum_{i=1}^{n}(r(x_i)-\bar{r}(x))^2}\sqrt{\sum_{i=1}^{n}(r(y_i)-\bar{r}(y))^2}}, \quad \text{for } i = 1,\ldots,n, \tag{1}$$

where $\bar{r}(x)$ and $\bar{r}(y)$ are the average ranks of X and Y. Tied categories are treated by taking the average of the positions that they would have otherwise occupied. The Spearman's correlation coefficient is a measure of monotonic dependence bounded in the close range $[-1,+1]$ and crossing value equal to zero in case of independence between the variables.

Kendall's τ_b coefficient

The Kendall's τ_b correlation coefficient represents an alternative to the Spearman's correlation coefficient. Contrary to the Spearman's correlation coefficient, the Kendall's correlation coefficient is based on the notions of concordance and discordance between pairs of observations, i.e.

$$\tau_b = \frac{2(C-D)}{\sqrt{n^2-n-\sum_{i=1}^{n} s_i(s_i-1)}\sqrt{n^2-n-\sum_{i=1}^{n} t_i(t_i-1)}}, \qquad \text{for } i = 1,\ldots,n, \tag{2}$$

where s_i and t_i are the number of tied x_i and y_i values in the i-th tied group, respectively. As well as Spearman's r_S, Kendalls τ_b takes values in a close range $[-1,+1]$.

Both the Spearman's r_S and Kendall's τ_b coefficients are symmetric measures, in the sense that they do not require to specify the role of the two involved variables, that is if the variables are the dependent or independent ones. Statistical literature provides also asymmetrical measures, for which the role of the dependent and independent variables has to be clearly identified. An example is the Somers' Δ coefficient.

Somers' Δ coefficient

The Somers' Δ coefficient is expressed as

$$\Delta = \frac{n_C - n_D}{n_C + n_D + n_{tiedX}}, \tag{3}$$

where n_{tiedX} is the number of tied pairs on X, n_C is the total number of concordant pairs and

n_D is the total number of discordant pairs. Somers' Δ corresponds to the Goodman's γ coefficient penalized for pairs tied only on the independent variable X (see, e.g. Goodman and Kruskal 1954).

2.2 The MDC_{go} coefficient formalization approach for two-way contingency tables

As discussed by Raffinetti and Aimar (2019), some concerns in the dependence relationship measurement process may arise in presence of grouped variables whose actual points are unobserved. In such a case, Raffinetti and Aimar (2019) introduced a new revisited version of the "Monotonic Dependence Coefficient" (MDC), originally proposed by Ferrari and Raffinetti (2015) as a dependence measure for quantitative and ordinal/tied data, they called "Monotonic Dependence Coefficient for grouped-ordinal data" (MDC_{go}). Specifically, the MDC_{go} allows to preserve the original continuous nature of the grouped variable since the terms involved in its computation are represented by the actual mid-points of each grouped variable class. In this way, the grouped variable is not reduced to its ordinal information, as would happen with the Spearman's, Kendall's and Somers' coefficients which only take into account the position of the observations (rank) or the count of observations that have concordant/discordant ranks.

Suppose the available data are organized in a two-way contingency table reporting information on the phenomenon Y under study (dependent variable) and on the predictor X (independent variable) affecting the phenomenon. Let Y be a grouped variable including h classes (groups) with lower and upper bounds indicated with y_{l_j} and y_{u_j}, for $j = 1, \ldots, h$, and X an ordinal variable expressed through k ordered categories (see Table 1).

By denoting with $n_{j\cdot}$ and $n_{\cdot i}$ the row and column marginal frequency distributions, respectively, it results that, for $j = 1, \ldots, h$ and $i = 1, \ldots, k$, $\sum_{j=1}^{h} n_{j\cdot} = \sum_{i=1}^{k} n_{\cdot i} = n$, where n corresponds to the total number of statistical units. The aggregate phenomenon described by the grouped variable can be re-expressed in terms of single statistical units, as follows:

Table 1: Data organized in a two-way contingency table

	X				
Y	$x_1 = 1$	$\dots$	$x_i = i$	$\dots$	$x_k = k$
$[y_{l_1}; y_{u_1})$	n_{11}	$\dots$	n_{1i}	$\dots$	n_{1k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[y_{l_j}; y_{u_j})$	n_{j1}	$\dots$	n_{ji}	$\dots$	n_{jk}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[y_{l_h}; y_{u_h})$	n_{h1}	$\dots$	n_{hi}	$\dots$	n_{hk}

$$Y = \left\{ \underbrace{[y_{l_1}; y_{u_1}), \dots, [y_{l_1}; y_{u_1})}_{n_1}, \dots, \underbrace{[y_{l_j}; y_{u_j}), \dots, [y_{l_j}; y_{u_j})}_{n_j - n_{j-1}}, \dots, \underbrace{[y_{l_h}; y_{u_h}), \dots, [y_{l_h}; y_{u_h})}_{n_h - n_{h-1}} \right\}, \quad (4)$$

where $n_1, \dots, n_j - n_{j-1}, \dots, n_h - n_{h-1}$ are the absolute frequencies associated with the first group, the j-th group until the last group. Moreover, n_1 is the position of the last unit in the first group, n_j is the position of the last unit in the j-th group, n_h is the position of the last unit in the last group. Similarly, the independent variable X can be translated into single units, as

$$X = \left\{ \underbrace{x_1 = 1, \dots, x_1 = 1}_{m_1}, \dots, \underbrace{x_i = i, \dots, x_i = i}_{m_i - m_{i-1}}, \dots, \underbrace{x_k = k, \dots, x_k = k}_{m_k - m_{k-1}} \right\}, \quad (5)$$

where $m_1, \dots, m_i - m_{i-1}, \dots, m_k - m_{k-1}$ represent the absolute frequencies associated with the first ordered, the i-th ordered until the last ordered category, while $m_1, \dots, m_i, \dots, m_k$ correspond to the position of the last unit in the first, i-th and last ordered category, respectively. Note that $n = \sum_{j=1}^{h} n_{j\cdot} = \sum_{j=1}^{h} (n_j - n_{j-1})$, with $n_{j-1} = n_0 = 0$ if $j = 1$, and similarly $n = \sum_{i=1}^{k} n_{\cdot i} = \sum_{i=1}^{k} (m_i - m_{i-1})$, with $m_{i-1} = m_0 = 0$ if $i = 1$. As a consequence, it derives that $n_h = m_k = n$.

Given the grouped nature of variable Y, the related tendency measure is provided by the mid-point of each class, $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_h$, such that $\bar{y}_1 < \bar{y}_2 < \dots < \bar{y}_h$. Consequently, the grouped variable Y in (4) can be re-written as

$$\bar{Y} = \left\{ \underbrace{\bar{y}_1, \ldots, \bar{y}_1}_{n_1}, \underbrace{\bar{y}_2, \ldots, \bar{y}_2}_{n_2 - n_1}, \ldots, \underbrace{\bar{y}_h, \ldots, \bar{y}_h}_{n_h - n_{h-1}} \right\}. \tag{6}$$

The monotonic dependence relationship between the grouped dependent and the ordinal independent variables is assessed by comparing the grouped variable mid-points values, which in one case are arranged in a non-decreasing sense and in the other case are re-ordered according to the the ordered categories of the independent variable. The ordinal nature of the independent variable leads to tied values, that are data charcaterised by a frequency distribution. Since by construction the MDC_{go} coefficient involves an ordering process based on the ordered categories, the suggestion of Raffinetti and Ferrari (2015) is considered, meaning that the mid-points of each class corresponding to the same independent variable ordered category are substituted by their mean value. By denoting with $\bar{Y}^*$ the $\bar{Y}$ values resulting from the previous manipulation and re-arranged according to the k ordered categories taken by the X variable, it results that

$$\bar{Y}^* = \left\{ \underbrace{\bar{y}^*_1, \ldots, \bar{y}^*_1}_{m_1}, \ldots, \underbrace{\bar{y}^*_i, \ldots, \bar{y}^*_i}_{m_i - m_{i-1}}, \ldots, \underbrace{\bar{y}^*_k, \ldots, \bar{y}^*_k}_{m_k - m_{k-1}} \right\}. \tag{7}$$

As in equation (5), $m_1, \ldots, m_i, \ldots, m_k$ represent the position of the last unit in the class related to the $\bar{Y}^*$ values associated with the first, i-th and last ordered category. Analogously, $m_1, \ldots, m_i - m_{i-1}, \ldots, m_k - m_{k-1}$ are the absolute frequencies associated with the $\bar{Y}^*$ values corresponding to the first, the i-th until the last ordered category. Also in this case equality $m_k = n$ holds.

The MDC_{go} coefficient is formalized as

$$MDC_{go} = \frac{2\sum_{z=1}^{k} \sum_{i=m_{z-1}+1}^{m_z} i\bar{y}^*_z - n(n+1)\bar{\mu}}{2\sum_{j=1}^{h} \sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j - n(n+1)\bar{\mu}}, \tag{8}$$

where: $\bar{\mu} = (1/n)\sum_{i=1}^{n} \bar{y}_i$, $n_{j-1} = n_0 = 0$ for $j = 1$; $m_{z-1} = m_0 = 0$ for $z = 1$; $\bar{y}_j$ is the mid-point of the j-th class, i.e. $\bar{y}_j = \frac{y_{l_j} + y_{u_j}}{2}$, with y_{l_j} and y_{u_j} corresponding to the lower and upper bounds of the class, and such that $\bar{y}_j < \bar{y}_{j+1}\ \forall j = 1, \ldots, h$; $\bar{y}^*_z$ is the mean of the $\bar{Y}$ values belonging to the same ordered category z, with $z = 1, \ldots, k$, and such that $\bar{y}^*_z = \frac{\sum_{r=1}^{m_z - m_{z-1}} \bar{y}_r}{m_z - m_{z-1}}$ where $m_z - m_{z-1}$ is the absolute frequency associated with the z-th ordered category.

Remark 1. *The MDC_{go} coefficient is always well-defined if the condition $h > 1$ is fulfilled, mean-*

ing that data are organized in at least two groups. In the case of only one group ($h = 1$), the denominator in equation (8) becomes equal to zero and the MDC_{go} coefficient is undetermined (for more details, see Raffinetti and Aimar, 2019).

Remark 2. *Note that the term $\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^*$ at the numerator in (8) can be expressed as follows*

$$\begin{aligned}\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* &= \sum_{i=1}^{m_1} i\bar{y}_1^* + \sum_{i=m_1+1}^{m_2} i\bar{y}_2^* + \ldots + \sum_{i=m_{k-1}+1}^{m_k} i\bar{y}_k^* \\ &= \underbrace{1\cdot\bar{y}_1^* + \ldots + m_1\bar{y}_1^*}_{m_1} + \underbrace{(m_1+1)\bar{y}_2^* + \ldots + m_2\bar{y}_2^*}_{m_2-m_1} + \ldots + \underbrace{(m_{k-1}+1)\bar{y}_k^* + \ldots + m_k\bar{y}_k^*}_{m_k-m_{k-1}} \\ &= \sum_{i=1}^{n} i\bar{y}_i^*,\end{aligned}$$

where $\bar{y}_i^* = \left\{\underbrace{\bar{y}_1^*,\ldots,\bar{y}_1^*}_{m_1},\underbrace{\bar{y}_2^*,\ldots,\bar{y}_2^*}_{m_2-m_1},\ldots,\underbrace{\bar{y}_k^*,\ldots,\bar{y}_k^*}_{m_k-m_{k-1}}\right\}$ *and* $n = \sum_{z=1}^{k}(m_z - m_{z-1})$, *with* $m_{z-1} = m_0 = 0$ *if* $z = 1$. *Similarly, the term* $\sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j$ *at the denominator in (8) can be re-expressed as* $\sum_{i=1}^{n} i\bar{y}_i$, *with* $\bar{y}_i = \left\{\underbrace{\bar{y}_1,\ldots,\bar{y}_1}_{n_1},\underbrace{\bar{y}_2,\ldots,\bar{y}_2}_{n_2-n_1},\ldots,\underbrace{\bar{y}_h,\ldots,\bar{y}_h}_{n_h-n_{h-1}}\right\}$ *and* $n = \sum_{j=1}^{h}(n_j - n_{j-1})$, *with* $n_{j-1} = n_0 = 0$ *if* $j = 1$.

To better illustrate the MDC_{go} coefficient, we now introduce a tutorial example to clarify its computation in the case of data organized in two-way contingency tables.

Tutorial example. Suppose to consider $n = 20$ statistical units on which information are collected in terms of a dependent variable Y, split into $h = 3$ equal width classes, and an independent variable X, expressed according to $k = 2$ ordered categories. Specifically, data are reported in Table 2.

Follow the steps below:

1. translate the aggregate data in Table 2 into single statistical units, as displayed in the first and second columns of Table 3;

2. determine the mid-points $\bar{Y}$ of each class, as reported in the fourth column of Table 3;

Table 2: A tutorial example

	X	
Y	$x_1 = 1$	$x_2 = 2$
$[0;100)$	2	2
$[100;200)$	4	6
$[200;300)$	1	5

Table 3: Data expressed in terms of single statistical units

Y	X	I	$\bar{Y}$	$\bar{Y}^*$
$[0;100)$	1	1	50	135.71
$[0;100)$	1	2	50	135.71
$[0;100)$	2	3	50	135.71
$[0;100)$	2	4	50	135.71
$[100;200)$	1	5	150	135.71
$[100;200)$	1	6	150	135.71
$[100;200)$	1	7	150	135.71
$[100;200)$	1	8	150	173.08
$[100;200)$	2	9	150	173.08
$[100;200)$	2	10	150	173.08
$[100;200)$	2	11	150	173.08
$[100;200)$	2	12	150	173.08
$[100;200)$	2	13	150	173.08
$[100;200)$	2	14	150	173.08
$[200;300)$	1	15	250	173.08
$[200;300)$	2	16	250	173.08
$[200;300)$	2	17	250	173.08
$[200;300)$	2	18	250	173.08
$[200;300)$	2	19	250	173.08
$[200;300)$	2	20	250	173.08

3. as X is expressed according to two pairs of equal categories, replace the $\bar{Y}$ values corresponding to the same ordered category by their mean values, i.e. $135.71 \approx \frac{50 \cdot 2 + 150 \cdot 4 + 250 \cdot 1}{7}$, for $X = 1$, and $173.08 \approx \frac{50 \cdot 2 + 150 \cdot 6 + 250 \cdot 5}{13}$, for $X = 2$;

4. compute the MDC_{go} numerator, by arranging the $\bar{Y}$ values updated through the procedure reported at step 3 according to the corresponding X ordered categories (last column of Table 3). Finally, compute the MDC_{go} denominator by considering the original $\bar{Y}$ values (fourth column of Table 3).

Property 1. *The MDC_{go} coefficient takes values in the close range* $[-1,+1]$.

Proof. Property 1 can be proved by resorting to inequalities proposed by Marshall et al. (2011), i.e.

$$\sum_{s=1}^{i} \bar{y}_s^* \geq \sum_{s=1}^{i} \bar{y}_s \tag{9}$$

and

$$\sum_{s=1}^{i} \bar{y}_s^* \leq \sum_{s=1}^{i} \bar{y}_{n+1-s}. \tag{10}$$

To verify that $MDC_{go} \leq +1$ and $MDC_{go} \geq -1$, we focus on the following inequalities

$$\underbrace{\sum_{z=1}^{k} \sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^*}_{\sum_{i=1}^{n} i\bar{y}_i^*} \leq \underbrace{\sum_{j=1}^{h} \sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j}_{\sum_{i=1}^{n} i\bar{y}_i} \tag{11}$$

and

$$\underbrace{\sum_{z=1}^{k} \sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^*}_{\sum_{i=1}^{n} i\bar{y}_i^*} \geq -\underbrace{\sum_{j=1}^{h} \sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j}_{\sum_{i=1}^{n} i\bar{y}_i} + n(n+1)\bar{\mu}. \tag{12}$$

Inequality in (11) can be proved by looking at inequality in (9). Since inequality in (9) is intuitively true $\forall i = 1,\ldots,n$, then also

$$\sum_{i=1}^{n} \sum_{s=1}^{i} \bar{y}_s^* \geq \sum_{i=1}^{n} \sum_{s=1}^{i} \bar{y}_s \tag{13}$$

holds. As $\sum_{i=1}^{n} \sum_{s=1}^{i} \bar{y}_s^* = n(n+1)\bar{\mu} - \sum_{i=1}^{n} i\bar{y}_i^*$ and $\sum_{i=1}^{n} \sum_{s=1}^{i} \bar{y}_s = n(n+1)\bar{\mu} - \sum_{i=1}^{n} i\bar{y}_i$ are verified, relation in (13) becomes

$$n(n+1)\bar{\mu} - \sum_{i=1}^{n} \bar{y}_i^* \geq n(n+1)\bar{\mu} - \sum_{i=1}^{n} \bar{y}_i$$

which gives $\sum_{i=1}^{n} i\bar{y}_i^* \leq \sum_{i=1}^{n} i\bar{y}_i$.

Inequality in (12) can be proved in a similar way, by considering inequality in (10). As it results that $\sum_{s=1}^{i} \bar{y}_s^* \leq \sum_{s=1}^{i} \bar{y}_{n+1-i}$, $\forall i = 1,\ldots,n$, then it follows that

$$\sum_{i=1}^{n}\sum_{s=1}^{i}\bar{y}_s^* \leq \sum_{i=1}^{n}\sum_{s=1}^{i}\bar{y}_{n+1-s}. \tag{14}$$

The term on the right side of inequality (14) can be rewritten as

$$\sum_{i=1}^{n}\sum_{s=1}^{i}\bar{y}_{n+1-s} = n(n+1)\bar{\mu} - \sum_{i=1}^{n} i\bar{y}_{n+1-i}, \tag{15}$$

where $\sum_{i=1}^{n} i\bar{y}_{n+1-i} = n(n+1)\bar{\mu} - \sum_{i=1}^{n} i\bar{y}_i$. Thus, from (14), we obtain that $n(n+1)\bar{\mu} - \sum_{i=1}^{n} i\bar{y}_i^* \leq \sum_{i=1}^{n} i\bar{y}_i \Rightarrow \sum_{i=1}^{n} i\bar{y}_i^* \geq -\sum_{i=1}^{n} i\bar{y}_i + n(n+1)\bar{\mu}$ ■

Remark 3. *If the MDC_{go} coefficient takes values greater than zero, an increasing monotonic dependence relationship arises. On the contrary, if the MDC_{go} coefficient takes values smaller than zero, a decreasing monotonic dependence relationship occurs (see Raffinetti and Aimar, 2019).*

Property 2. *$MDC_{go} = +1$ (perfect monotonic increasing dependence relationship) when*

$$\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* = \sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j. \tag{16}$$

From Remark 2., it results that equation (16) is equivalent to

$$\sum_{i=1}^{n} i\bar{y}_i^* = \sum_{i=1}^{n} i\bar{y}_i. \tag{17}$$

Equality in (16) is fulfilled if $k = h$ and $n_j = m_z$, $\forall j = 1,\ldots,h$ and $z = 1,\ldots,k$, while equality in (17) is fulfilled if $\bar{y}_i^* = \bar{y}_i$, $\forall i = 1,\ldots,n$.

The perfect monotonic increasing dependence relationship scenario is achieved if the number h of groups equals the number k of the ordered categories and, in addition, the units belonging to each j-th group are the same units belonging to the corresponding z-th ordered category.

Property 3. *$MDC_{go} = -1$ (perfect monotonic decreasing dependence relationship), when*

$$\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* = \sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} (n+1-i)\bar{y}_j. \tag{18}$$

Equation in (18) can be rewritten as

$$\sum_{i=1}^{n} i\bar{y}_i^* = \sum_{i=1}^{n} i\bar{y}_{(n+1-i)} \tag{19}$$

Equality in (18) is fulfilled if $k = h$ *and* $n_j = n_{k+1-j}$, $\forall j = 1,\ldots,h$ *and* $z = 1,\ldots,k$, *while equality in (19) is fulfilled if* $\bar{y}_i^* = \bar{y}_{(n+1-i)}$, $\forall i = 1,\ldots,n$.

The perfect monotonic decreasing dependence relationship scenario is reached if the number h of groups equals the number k of the ordered categories and, in addition, the units belonging to each j-th group are the same units belonging to the corresponding $(k+1-j)$*-th ordered category.*

Property 4. $MDC_{go} = 0$ *(independence), when*

$$2\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* = n(n+1)\bar{\mu}, \tag{20}$$

that is, if $k = 1$.

Due that all the statistical units belong to the same ordered category, $2\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* = 2\sum_{z=1}^{1}\sum_{i=1}^{m_1} i\bar{y}_1^* = 2\sum_{i=1}^{n} i\bar{y}_i^* = 2\sum_{i=1}^{n} i\bar{\mu}$. *Since* $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, *it derives that* $2\sum_{i=1}^{n} i\bar{\mu} = n(n+1)\bar{\mu}$ *and the result in (20) follows.*

As an example of the independence scenario, consider data related to $n = 10$ *statistical units distributed as in Table 4.*

Table 4: Example of independence

	X	
Y	$x_1 = 1$	$x_2 = 2$
$[0;100)$	3	0
$[100;200)$	2	0
$[200;300)$	5	0

From data provided in Table 4, it results that $\bar{\mu} = 170$ *and* $\bar{y}_i^* = 170$, $\forall i = 1,\ldots,10$. *Thus, MDC*$_{go}$ *numerator in (8) numerically becomes* $(1\cdot 170+\ldots+10\cdot 170)-10(10+1)170 = 18,700-18,700 = 0$.

Remark 4. *The more the MDCgo coefficient moves away from value zero, the more the monotonic dependence relationship between the variables arises.*

3 Assessment on simulated data

The purpose of assessing the most performing dependence measure in the case of grouped-ordinal variables with unobserved non-Normal distributions is achieved following the framework presented in Raffinetti and Aimar (2019). In Raffinetti and Aimar (2019), the authors provided an extensive Monte Carlo simulation study by sampling from both bivariate Normal distributions and t-Student distributions, in this last case to take into account the effect associated with a leptokurtic distribution characterized by a higher kurtosis than the Normal distribution.
This section is structured in two subsections. Subsection 3.1 introduces the procedure to generate data from non-Normal distributions and specifies the considered Monte Carlo simulation conditions; Subsection 3.2 illustrates the simulation findings and validate them through the Dunn's test.

3.1 Simulation conditions

We consider a family of bivariate non-Normal distributions according to the method proposed by Vale and Maurelli (1983) and translated into an `R` code by Zopluoglu (2011). The first contribution in generating non-Normal variables is due to Fleishman (1978), who defines in the univariate case, a non-Normal variable Y as a linear combination of the first three powers of a standard Normal variable X. Specifically,

$$Y = a + bX + cX^2 + dX^3, \tag{21}$$

where a, b, c and d are real-valued polynomial coefficients that provide the specified non-Normal distributional form for Y. The polynomial coefficients are obtained through the following system of non-linear equations:

$$
\begin{aligned}
&a+c=0\\
&b^2+6bd+2c^2+15d^2-1=0\\
&2c(b^2+24bd+105d^2+2)-\gamma_1=0\\
&24[bd+c^2(1+b^2+28bd)+d^2(12+48bd+141c^2+225d^2)]-\gamma_2=0,
\end{aligned}
$$

where γ_1 is the population skewness and γ_2 is the population excess kurtosis set by the user. The system of non-linear equations can be solved by resorting to the R software. More details can be found in Astivia (2019), who provided an example of R code to derive the polynomial coefficients corresponding to the values chosen for the skewness and kurtosis.

Vale and Maurelli (1983) developed the Fleishman's method for generating multivariate non-Normal distributions with specified inter-correlations and marginal means, variances, skewness, and kurtosis. The general procedure to simulate multivariate non-Normal variables is to start by generating Normal variables with a specific intermediate correlation matrix. Let $\mathbf{X}$ be an n-dimensional random vector defined as

$$
\mathbf{X}=\mathbf{AZ}+\mathbf{b}, \tag{22}
$$

where $\mathbf{Z}$ is the n-dimensional random vector including random variables distributed according to a standard Normal distribution, $\mathbf{A}$ is an $n\times n$ matrix and $\mathbf{b}$ is the mean vector of $\mathbf{X}$. The matrix $\mathbf{A}$ has to be determined in order that the vector $\mathbf{X}$ transformed from $\mathbf{Z}$ is characterised by a specific intermediate correlation matrix. To do this, the first step is to express the covariance matrix of $\mathbf{X}$, pointed out with $\mathbf{C}$, in terms of matrix $\mathbf{A}$, i.e.

$$
\mathbf{C}=\mathrm{Cov}(\mathbf{X})=\mathrm{Cov}(\mathbf{AZ}+\mathbf{b})=\mathbf{A}\mathrm{Cov}(\mathbf{Z})\mathbf{A}^{'}=\mathbf{AIA}^{'}=\mathbf{AA}^{'}, \tag{23}
$$

where $\mathbf{I}$ is the identity matrix.

Matrix $\mathbf{A}$ can be then determined by resorting to the spectral decomposition of the covariance matrix $\mathbf{C}$, according to which the matrix $\mathbf{C}$ is decomposed into the product of three other matrices as follows

$$\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}' = (\mathbf{U}\mathbf{D}^{1/2})(\mathbf{U}\mathbf{D}^{1/2})', \tag{24}$$

where $\mathbf{U}$ is a real or complex unitary matrix including the eigenvectors of $\mathbf{C}$ and $\mathbf{D}$ is the diagonal matrix including the eigenvalues of $\mathbf{C}$. By combining equation (23) with equation (24), it derives that $\mathbf{A} = \mathbf{U}\mathbf{D}^{1/2}$. In this way, $\mathbf{X}$ is transformed from $\mathbf{Z}$ by using the transformation matrix $\mathbf{A}$, and is characterised by the desired covariance matrix.
Based on the previous premises the Vale and Maurelli's algorithm, referred to the bivariate case, can be summarized through six main steps (see, e.g. Wicklin, 2013):

- ***Step 1:*** set the skewness and kurtosis values for the marginal distributions and compute the polynomial Fleishman coefficients for each;

- ***Step 2:*** specify the desired correlation $r_{Y_1Y_2}$ between the two non-Normal variables Y_1 and Y_2 and compute the intermediate correlation $\rho_{X_1X_2}$ for the Normal variables X_1 and X_2 by solving for the roots of a third order polynomial equation involving the Fleishman's coefficients, i.e. by solving $r_{Y_1Y_2} = \rho_{X_1X_2}(b_1b_2 + 3b_1d_2 + 3d_1b_2 + 9d_1d_2) + \rho^2_{X_1X_2}(2c_1c_2) + \rho^3_{X_1X_2}(6d_1d_2)$ with respect to $\rho_{X_1X_2}$;

- ***Step 3:*** resort to the spectral decomposition $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}'$ to compute $\mathbf{A} = \mathbf{U}\mathbf{D}^{1/2}$;

- ***Step 4:*** simulate uncorrelated standard Normal variables $\mathbf{Z} = (Z_1, Z_2)$;

- ***Step 5:*** use $\mathbf{A}$ and $\mathbf{Z}$ to form correlated variables $\mathbf{X} = (X_1, X_2)$;

- ***Step 6:*** apply the Fleishman cubic transformation to obtain the final data $\mathbf{Y} = (Y_1, Y_2)$ from a distribution with the given correlations and where the marginal distributions have the specified skewness and kurtosis.

According to the Vale and Maurelli's method, the two parameters that affect the normality condition are skewness and kurtosis, where the kurtosis is measured by the parameter excess of kurtosis. All the more the excess of kurtosis moves from value 0, all the more the normality condition becomes weaker. This happens also for the skewness parameter: indeed, a value greater than zero of skewness translates into an asymmetrical distribution.
This family of distributions is particularly interesting since it allows us to vary the skewness

and kurtosis parameters in order to assess their impact on each index dynamic when the normality condition is violated. Based on these considerations, we first consider samples from bivariate non-Normal distributions, by fixing the values taken by the pairwise correlation coefficient ρ, and then discretize one of the two variables and group the other variable. Analogously to Raffinetti and Aimar (2019), we choose three different values of ρ: $\rho = \{0.2, 0.5, 0.8\}$. Four combinations of the skewness (sk) and kurtosis (ku) parameters are set: $\{ku = (2,2), sk = (1.5, 1.5)\}$, $\{ku = (2,2), sk = (2,2)\}$, $\{ku = (3,3), sk = (1.5, 1.5)\}$ and $\{ku = (3,3), sk = (2,2)\}$. In addition, for each of the three ρ values, the X variable is discretized according to the following discretization procedures:

(i) discretization with equal-width intervals (EW);

(ii) discretization with uniform probability (U);

(iii) discretization with asymmetrical probability (A).

The number of the k categories is let vary to each discretization scenario, by setting $k = \{3, 5, 7\}$. We use the same notation proposed by Raffinetti and Aimar (2019) to denote the variable X discretized into three, five and seven ordered categories. Specifically, we point out with: EW3, EW5 and EW7, the cases of three, five and seven equal-width categories; U3, U5 and U7, the cases of three, five and seven uniform categories; A3, A5 and A7, the cases of three, five and seven asymmetrical categories, respectively. The variable Y is transformed into a grouped variable characterized by h equal width classes. The number h of groups is set by fixing $h = \{3, 5, 7\}$. The sample size is $n = 500$ and the number of iterations is equal to 10,000. Variable X is discretized into asymmetrical categories as suggested by Raffinetti and Aimar (2019) and displayed in Table 5, where the number k of categories is provided together with the related probabilities p_{asym}.

Table 5: Number k of categories for discretization with asymmetrical probability p_{asym}

k	p_{asym}						
3	0.10	0.30	0.60				
5	0.05	0.10	0.15	0.20	0.50		
7	0.05	0.05	0.10	0.10	0.20	0.25	0.25

3.2 Simulation findings

Before discussing the obtained simulation results, a remark is required. When dealing with non-normally distributed data, the pairwise correlation coefficient ρ does not represent the benchmark value as in the case of normally distributed data. If no discretization and grouping data transformation is introduced, we expect that the Monotonic Dependence Coefficient appears as the most performing coefficient taking values higher than the selected pairwise correlation coefficient values (see, e.g. Ferrari and Raffinetti, 2015). The discretization and grouping process, instead, provides a shrinkage in the dependence relationship measurement. Therefore, our aim is to detect which dependence coefficient reaches the highest values in all the considered scenarios.
The simulation findings are displayed through the boxplots representing the distributions of the four indices referred to the scenarios considered for ρ, ku, sk, h and k. We remark that, for the sake of brevity and without loss of generality, we report only the boxplots associated with the kurtosis and skewness combination $\{ku = (3,3), sk = (2,2)\}$, due to the similarity with the findings related to the other combinations.
Boxplots in Figure 1, Figure 2 and Figure 3 describe the distribution of the four coefficients for the pairwise coefficients $\rho = 0.2$, $\rho = 0.5$ and $\rho = 0.8$, respectively, which are graphically denoted with a horizontal green dashed line.

Figure 1 about here

Figure 2 about here

Figure 3 about here

We evaluate the performance of the MDC_{go}, τ_b, r_S and Δ coefficients by looking at their median values arising from the boxplots. Specifically, the higher is the median value of a coefficient, the better is its capability in catching the existing dependence relationship. When considering the discretization process applied to the X variable (equal width, uniform and asymmetrical categories) and the grouping process of the Y variable (equal width intervals), in most cases the MDC_{go} index performs better than the Spearman's, Kendall's and Somers' coefficients, highlighting its attitude in catching the dependence relationships in presence of grouped-ordinal data. The only scenarios in which the MDC_{go} performs worse than its competitors arise when

$\rho = \{0.2, 0.5.0.8\}$, $Y = \{5, 7\}$ and $X =$EW3. In these cases, the Somers' Δ coefficient reaches the highest values. The MDC_{go} seems to perform as well as its considered competitors when $\rho = 0.2$, $Y = 3$ and $X =$EW3. In general, the MDC_{go} coefficient is characterized by a higher variability and, as the value of ρ increases, it results as the most performing monotonic dependence measure.

The considerations reported above need to be confirmed into an inferential perspective by checking if the mean (median) ranks of the four indices are significantly different from each other. In order to determine which coefficients are significantly different, we compare multiple pairs of the coefficient mean (median) ranks. The Dunn's Multiple Comparison Test is then applied (see, e.g. Dunn [1961], 1964). The Dunn's test is typically known as a "post hoc" non parametric test, since implemented after an ANOVA. Given v groups, the number of possible comparisons to take into account equals to $c = v(v-1)/2$. In our context, $v = 4$ since the analysis is focused on four monotonic dependence coefficients. Thus, the test is extended to $c = 6$ comparisons (more in detail, MDC_{go} vs r_S, MDC_{go} vs τ_b, MDC_{go} vs Δ, r_S vs τ_b, r_S vs Δ and τ_b vs Δ). The null hypothesis for the test is that there is no difference between the mean (median) ranks of the coefficients while the alternative hypothesis for the test is that there is a difference between the mean (median) ranks of the coefficients. The presence of multiple comparisons changes the meaning of the significance level α, which is adjusted by resorting to the Bonferroni's correction procedure according to which the adjusted significance level α is obtained as $1/c$. In our case, suppose to consider a significance level $\alpha = 0.05$, so that the Bonferroni adjusted significance level becomes $\alpha = 0.05/6 = 0.00833 \approx 0.01$.

By rejecting the Dunn's test null hypothesis, the findings provided by the boxplots displaying the coefficient ordering and consequently the associated capability in catching the monotonic dependence relationship in presence of grouped-ordinal data, are validated. The null hypothesis of equality between the mean (median) ranks of the coefficients is always rejected except for the compared pairs reported in Figures 4, 5, 6, 7, 8, 9, and 10 where the dot charts of the adjusted p-values of the Dunn's test are represented.

Figures 4, 5, 6, 7, 8, 9, and 10 about here

In the dot charts, the red-dashed line indicates the 0.05 threshold (the original significance level α) and the coefficients compared on a pairwise basis are indicated on the left-hand side. In

order to reject the null hypothesis of equality between the mean (median) ranks of the coefficients, the p-values have to be smaller than 0.01. From Figure 4, it results that r_S and τ_b, Δ and τ_b and r_S and Δ are not differently performing while, contrary to what previously stated, the MDC_{go} coefficient seems to have a different performance with respect to its competitors. In the cases reported in Figures 6, 7 and 8 it derives that both Δ and τ_b are similar in their performance. Finally, from Figures 5, 9 and 10 it arises that there is no significant difference between r_S and τ_b.

According to the findings based on the Dunn's test, the ordering of the coefficients provided by the boxplots is confirmed highlighting how in most scenarios the MDC_{go} is the most appropriate index to measure the monotonic dependence relationship with grouped-ordinal data generated by unobservable phenomena with underlying non-Normal distributions.

4 A real data example

We introduce an illustrative application to provide a further example of how our proposal works when dealing with real data. The considered dataset represents a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa, and is available at the link `https://www.kaggle.com/emilianito/saheart`. The whole version of these data is described in a contribution by Rousseauw et al. (1983).

The data frame includes 462 observations on the following 10 variables: sbp (systolic blood pressure); tobacco (cumulative tobacco - kg); ldl (low density lipoprotein cholesterol); adiposity (a measure of fat); famhist (family history of heart disease - Present, Absent); typea (type-A behavior - a score); obesity (a mesaure of obesity); alcohol (current alcohol consumption); age (age at onset); chd (coronary heart disease). In the original work by Rousseauw et al. (1983), the response variable is the coronary heart disease which takes 0/1 values. The purpose was to detect the main factors which could lead to a coronary hearth disease. In our perspective, the response variable has to be a continuous grouped variable and, consequently, we select as response variable the variable describing the systolic blood pressure (sbp). Specifically, we aim at assessing if the systolic blood pressure may be affected by the patient medical parameters and demographic features, such as the low density lipoprotein cholesterol (ldl), adiposity, obesity and age. The joint distributions of sbp and each single explanatory variable are displayed in

Figure 11, from which it arises that data are non-normally distributed.

Figure 11 about here

We apply the MDC_{go}, Somers' Δ, Kendall's τ_b and Spearman's r_S coefficients to evaluate if to an increase in the medical parameters and age of the patients corresponds an increase of the systolic blood pressure. Suppose that the observed response and explanatory variables, even if generated by unobserved non-normally distributed variables, are not available at the individual level but only at the group level (response variable) and through a Likert-scale (explanatory variables). To do this, we split the sbp variable into three equal width classes denoting the low-normal, normal-high and high systolic blood pressure. Specifically, we consider the following classes: $[101,140)$, $[140,179)$ and $[179,218)$. Analogously, the predictors ldl, adiposity, obesity and age are discretizing into five ordered categories whose distribution is reported in Table 6.

Table 6: Distribution of the discretized explanatory variables

Categories	1	2	3	4	5
ldl	0.38	0.47	0.12	0.03	0
adiposity	0.10	0.19	0.32	0.29	0.10
obesity	0.1	0.58	0.28	0.03	0.01
age	0.15	0.16	0.18	0.23	0.28

Results based on the MDC_{go} and the remaining coefficients are included in Table 7. Findings in Table 7 show the attitude of the MDC_{go} coefficient in measuring the monotonic dependence relationship between blood systolic pressure and cholesterol level, adiposity, obesity and age. Specifically, If on the one hand, the most impacting factors on the blood systolic pressure are adiposity (0.351) and age (0.388), on the other hand obesity seems not to greatly affect the blood systolic pressure. This consideration holds also if referring to the Spearman's, Kendall's and Somers' coefficients. Moreover, it arises that typically the following relation $MDC_{go} > r_S > \tau_b > \Delta$ is fulfilled for all the predictors. The MDC_{go} coefficient provides an improvement in the monotonic dependence relationship measurement equal to: the 2%, 9% and 20% with respect to r_S, τ_b and Δ for the ldl variable; the 13.1%, 22.6% and 38.5% with respect to r_S, τ_b and Δ for the adiposity variable; the 5.2%, 11.4% and 19.1% with respect to r_S, τ_b and Δ for the obesity variable; the 15.3%, 25% and 41.5% with respect to r_S, τ_b and Δ for the age variable.

Table 7: Results

Explanatory variable	MDC_{go}	r_s	τ_b	Δ
ldl	0.214	0.209	0.195	0.171
adiposity	0.351	0.305	0.272	0.216
obesity	0.220	0.209	0.195	0.178
age	0.388	0.329	0.291	0.227

5 Concluding remarks

In this paper an extended study for assessing the monotonic dependence relationships when dealing with grouped-ordinal variables generated by unobserved non-Normal variables and organized in two-way contingency tables is presented. The focus is on the use of the MDC_{go} coefficient, recently developed for measuring dependence with grouped-ordinal data with unobserved underlying Normal and t-Student distributions, in comparison with the most commonly standard dependence measures (Spearman's, Kendall's and Somers' coefficients).

In order to assess the MDC_{go} behavior with respect to the considered competitors (Spearman's, Kendall's and Somers' coefficients), a Monte Carlo simulation study based on several scenarios is led, by first sampling from bivariate non-Normal distributions characterized by specific pairwise correlation coefficient ($\rho = \{0.2, 0.5, 0.8\}$) kurtosis ($ku = (2,2), ku = (3,3)$) and skewness ($sk = (1.5, 1.5), sk = (2,2)$) parameters, and then by discretizing the variable selected as the independent one and grouping the variable considered as the dependent one. On average, the MDC_{go} coefficient reaches the highest values showing its capability in catching the actual monotonic dependence relationship with grouped-ordinal data generated by non-Normal distributions.

Finally, an application on hearth disease data provides evidence on the coherence between the MDC_{go} simulation results and those on real data.

We believe that the proposed measure provides a reliable tool to be used by researchers in the fields of psychology, sociology and other social sciences to assess the actual monotonic dependence relationships between variables built on latent constructs and then involving the loss of basic information.

Acknowledgments

Acknowledgments go to the two anonymous reviewers for their helpful comments and suggestions that allowed to improve the paper.

References

Astivia, O.L.O. 2019. “Issues, problems and potential solutions when simulating continuous, non-normal data in the social sciences”. *PsyArXiv*: 1-33. doi:10.31234/osf.io/frmnx

Dunn, O. J. 1961. “Multiple comparisons among means”. *Journal of the American Statistical Association* 56 (293): 52-64. doi:10.2307/2282330.

Dunn, O. J. 1964. “Multiple comparisons using rank sums”. *Technometricsn* 6 (3): 241-252. doi:10.2307/1266041.

Ferrari, P. A., and E. Raffinetti. 2015. “A different approach to Dependence Analysis”, *Multivariate Behavioral Research* 50 (2): 248-264. doi:10.1080/00273171.2014.973099.

Fleishman, A. I. 1978. “A method for simulating non-normal distributions”. *Psychometrika* 43: 521-532. doi:10.1007/BF02293811.

Goodman L. A., and W. H. Kruskal. 1954. “Measures of association for cross classifications”. *Journal of American Statistical Association* 49: 732-764. doi:10.1080/01621459.1954.10501231.

Kendall, K. 1938. “A New Measure of Rank Correlation”. *Biometrika* 30 (12): 81-89. doi:10.2307/2332226.

Likert, R. 1932. “A Technique for the measurement of attitudes”, *Archives of Psychology* 22 (140): 1-55.

Marshall, A. W., I. Olkin and C. A. Arnold. 2011. *Inequalities: Theory of Majorization and Its Applications*. Springer.

Norman, G. 2010. “Likert scales, levels of measurement and the law of statistics”. *Advances in Health Science Education* 15: 625-632. doi:10.1007/s10459-010-9222-y.

Pearson, K. 1907. *Mathematical Contributions to the Theory of Evolution, XVI. On Further Methods of Determining Correlation: Draper’s Research Memoirs, Biometric Series, IV*. Cambridge: Cambridge University Press.

Raffinetti, E., and F. Aimar. 2019. “MDC_{go} takes up the association/correlation challenge for grouped ordinal data”. *AStA Advances in Statistical Analysis* 103 (4): 527-561. doi:10.1007/s10182-018-00341-1.

Rousseauw, J., J. du Plessis, A. Benade, P. Jordaan, J. Kotze, and J. Ferreira. 1983. “Coronary risk factor screening in three rural communities”. *South African Medical Journal* 64: 430-436.

Somers, R.H. 1962. “A new asymmetric measure of association for ordinal variables”. *American Sociological Review* 27 (6): 799-811. doi:10.2307/2090408. (1962)

Spearman, C. 1904. “The proof and measurement of correlation between two things”, *American Journal of Psychology* 15: 72-101. doi:10.2307/1422689.

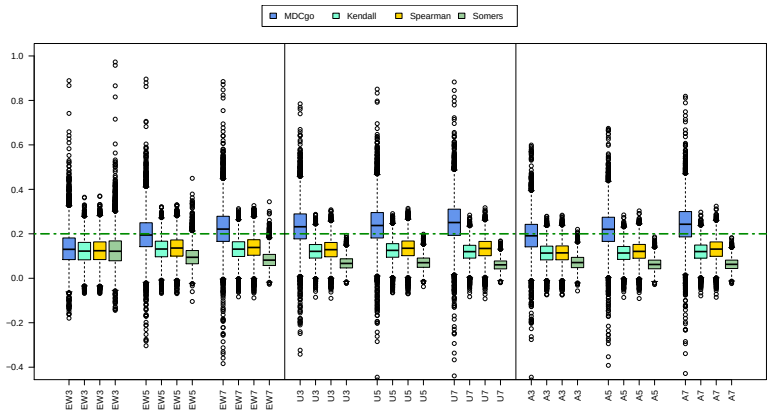
Vale, C. D., and V. A. Maurelli. 1983. “Simulating multivariate nonnormal distributions”. *Psychometrikas* 48 (3): 465-471. doi:10.1007/BF02293687.

Wicklin, R. 2013 “Simulating Data with SAS”. SAS Institute Inc., Cary, North Carolina.

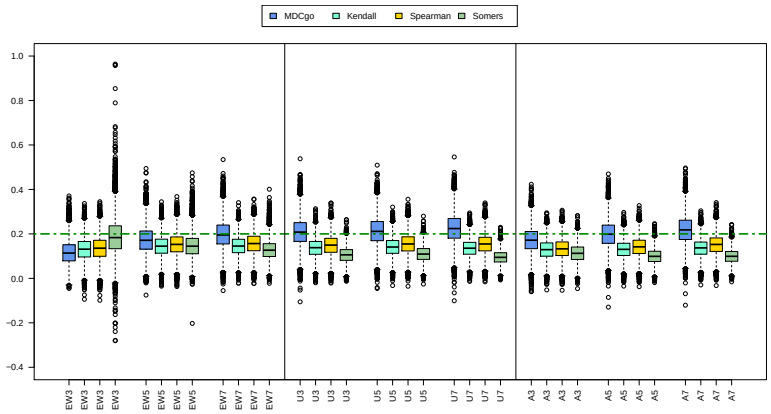
Zopluoglu, C. 2011. “Application in R: Generating Multivariate Non-normal Variables”. `https://www.dropbox.com/s/ldcu4f3mnf89fby/gennonnormal.r`.

Figures

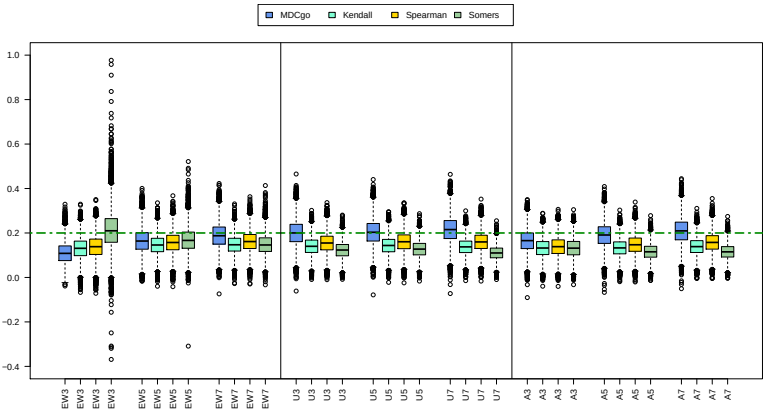
Figure 1: Boxplots of MDC_{go}, Kendall's, Spearman's, and Somers' coefficient distributions in the case of $ku = (3,3)$, $sk = (2,2)$ and $\rho = 0.2$



(a) Boxplots of the distribution of the four coefficients - Y=3 classes

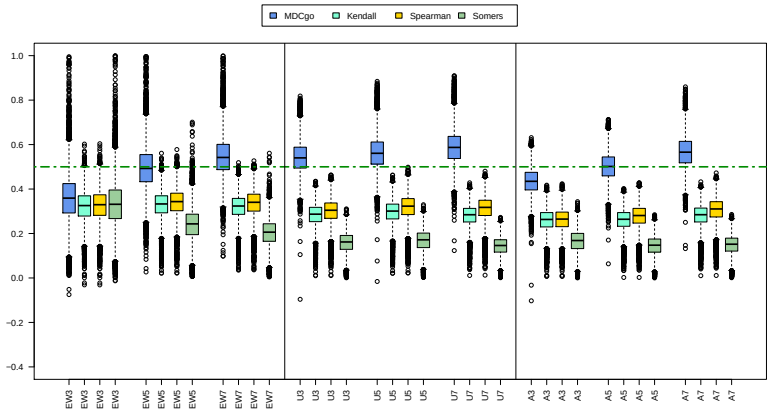


(b) Boxplots of the distribution of the four coefficients - Y=5 classes

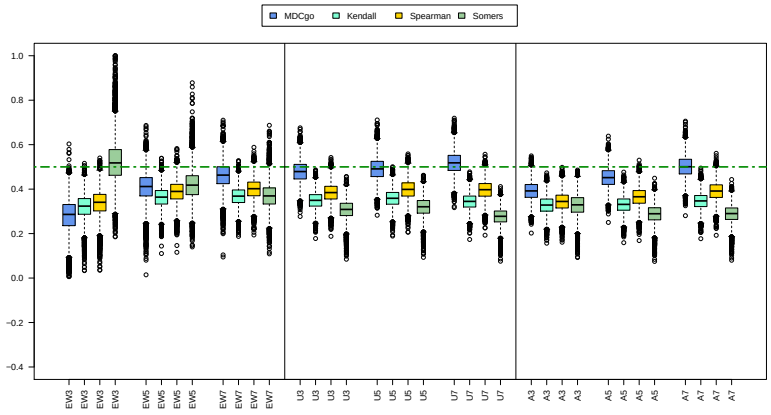


(c) Boxplots of the distribution of the four coefficients - Y=7 classes

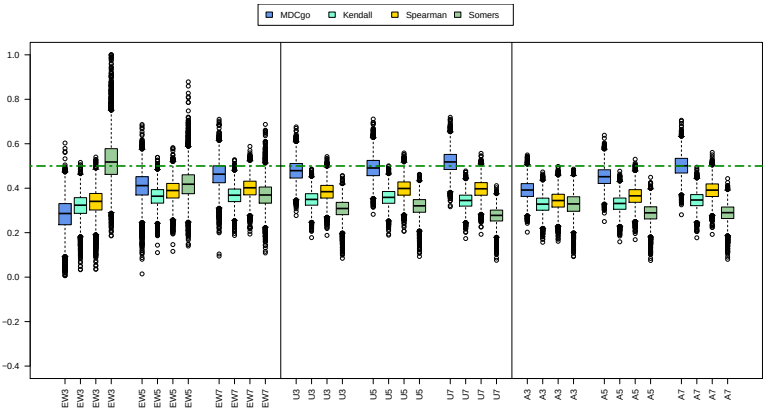
Figure 2: Boxplots of MDC_{go}, Kendall's, Spearman's, and Somers' coefficient distributions in the case of $ku = (3,3)$, $sk = (2,2)$ and $\rho = 0.5$



(a) Boxplots of the distribution of the four coefficients - Y=3 classes

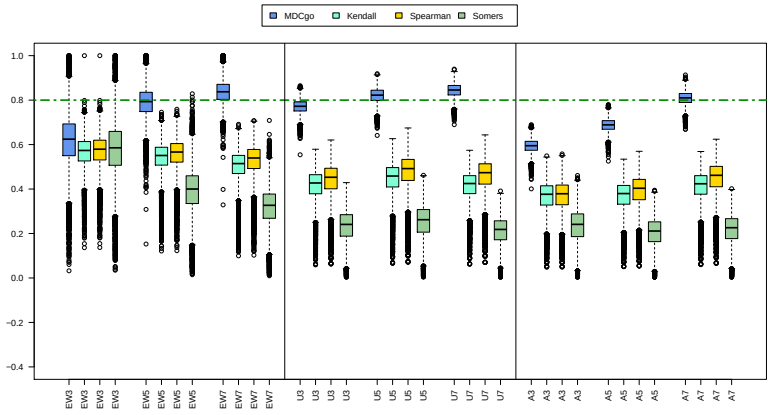


(b) Boxplots of the distribution of the four coefficients - Y=5 classes

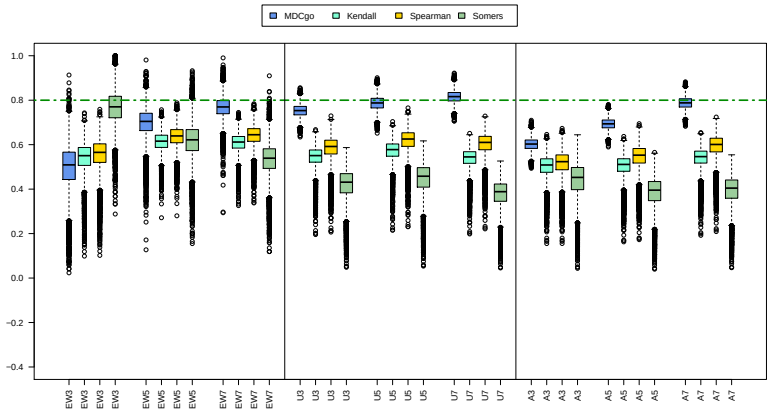


(c) Boxplots of the distribution of the four coefficients - Y=7 classes

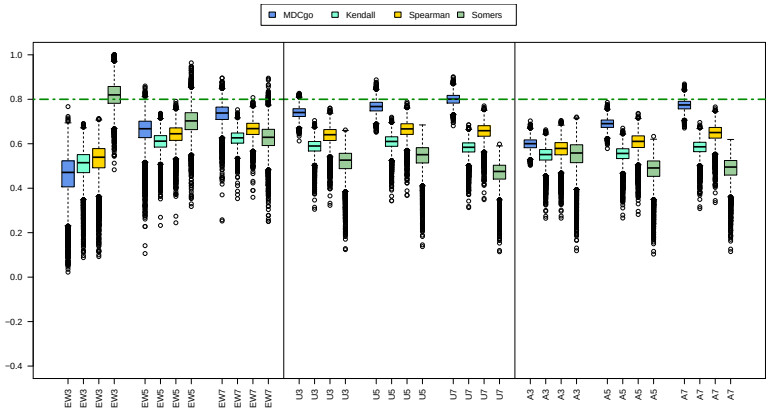
Figure 3: Boxplots of MDC_{go}, Kendall's, Spearman's, and Somers' coefficient distributions in the case of $ku = (3,3)$, $sk = (2,2)$ and $\rho = 0.8$



(a) Boxplots of the distribution of the four coefficients - Y=3 classes



(b) Boxplots of the distribution of the four coefficients - Y=5 classes



(c) Boxplots of the distribution of the four coefficients - Y=7 classes

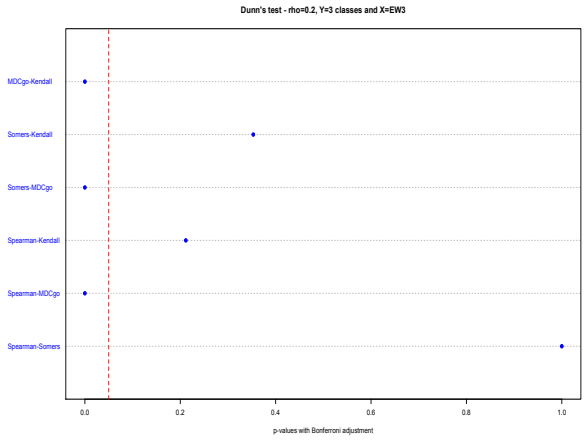


Figure 4: Dunn's test p-values - $\rho = 0.2$, $Y = 3$, $X =$EW3

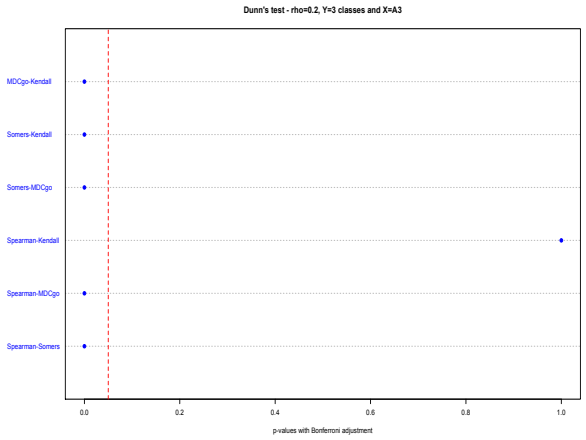


Figure 5: Dunn's test p-values - $\rho = 0.2$, $Y = 3$, $X =$A3

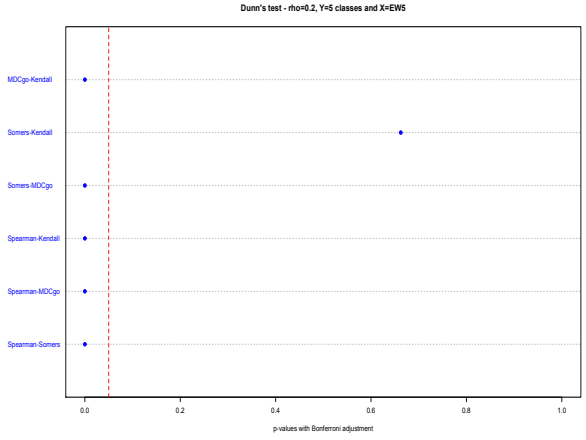


Figure 6: Dunn's test p-values - $\rho = 0.2$, $Y = 5$, $X =$EW5

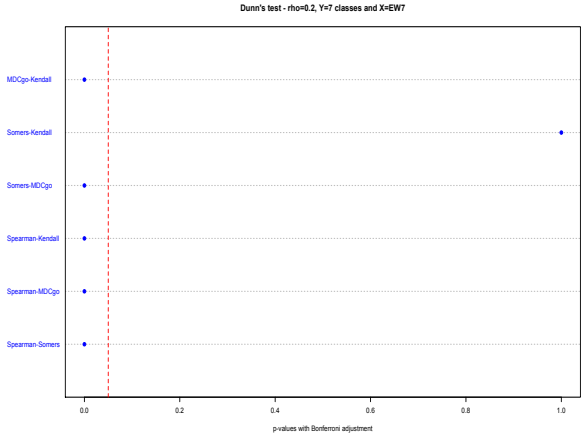


Figure 7: Dunn's test p-values - $\rho = 0.2$, $Y = 7$, $X =$EW7

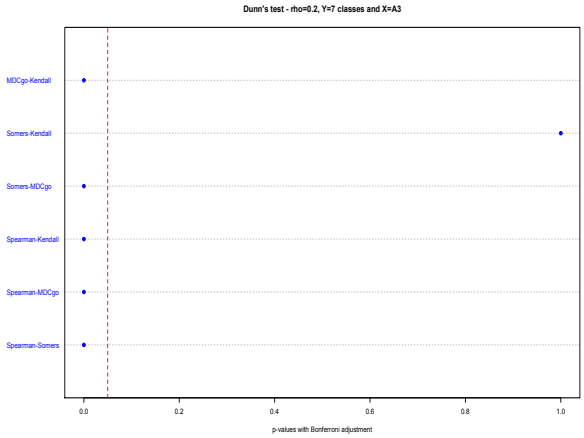


Figure 8: Dunn's test p-values - $\rho = 0.2$, $Y = 7$, $X =$A3

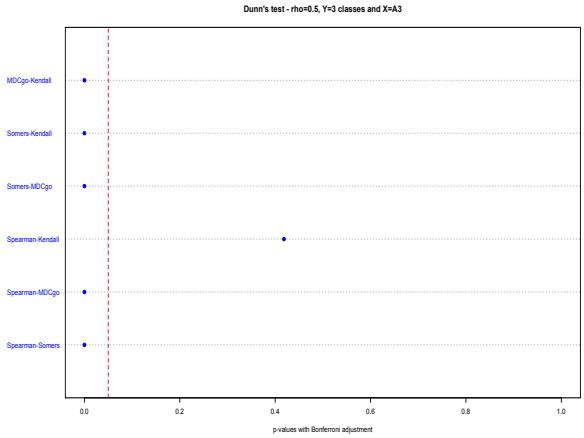


Figure 9: Dunn's test p-values - $\rho = 0.5$, $Y = 3$, $X =$A3

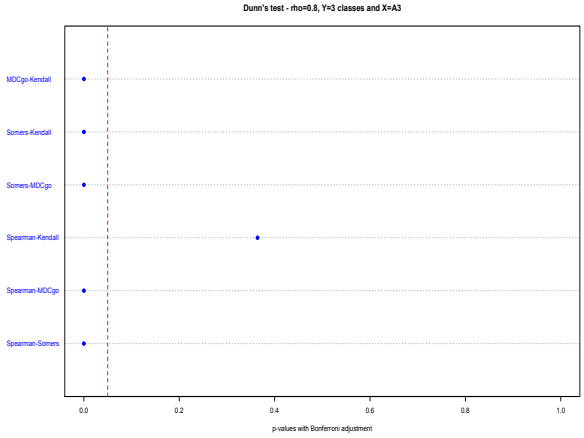


Figure 10: Dunn's test p-values - $\rho = 0.8$, $Y = 3$, $X =$A3

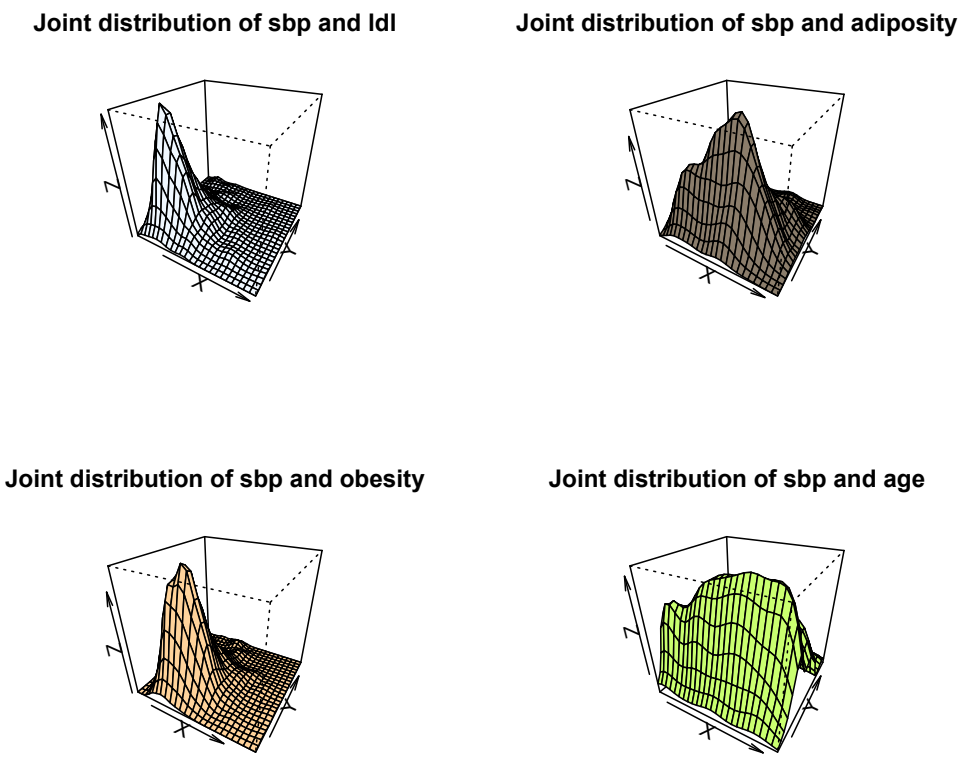


Figure 11: Joint distribution of sbp and each single explanatory variable