Multivariate data form and descriptive measures

Why multivariate

- As the name implies, multivariate statistical analysis is the study of several random variables simultaneously.
- There is dependence structure in these random variables. Incorporating the inter-dependence structure into the analysis improves the one-at-a-time univariate analysis.
- For large, high dimensional data with many variables, it is often necessary to reduce the data to lower dimensional space that still retains most desirable information in the data.

In fact, statistics has been about reducing observations to a few statistics holding essential information.

Common objectives in multivariate statistical analysis

- Understand and summarize in a simpler manner on data structure, which is often obscured by noise, random perturbations, and measurements errors.
- Understand and summarize in a simpler manner the relationship of one part of data to another.
- Inference from data to a much large population: parameter estimates, hypothesis tests, decisions.
- Univariate and multivariate statistical inference care about similar issues in data analysis and statistical model, such as central locations and variations, difference in treatment effects, outlier detection, and check violations of distribution assumptions, thus the validity of the analysis.

  Characterizing inter-dependence structure and dimension reduction are two key aspects of multivariate analysis.

1 Multivariate data form

Multivariate data analysis deals with measurements of random outcomes that consist of a set of variables for each observation.

Notations

Multivariate observations are often denoted as

$x_{jk} =$ Measurement ($a.k.a.$ observation, outcome, response) of the kth variable on the jth item (or subject)

Commonly, multivariate data are displayed as an array, rows label individual observations, columns index variables.

		variable 1	variable 2	$\cdots$	variable k	$\cdots$	variable p
(observation 1)	item 1	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
(observation 2)	item 2	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(observation j)	item j	x_{j1}	x_{j2}	$\cdots$	x_{jk}	$\cdots$	x_{jp}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(observation n)	item n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

(1)

The data array can be expressed in vector-matrix form.

$$\boldsymbol{X} = [x_{jk}]_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_j' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} \begin{array}{l} \longleftarrow \text{1st (multivariate) observation} \\ \longleftarrow \text{2nd observation} \\ \quad\vdots \\ \longleftarrow j\text{th observation} \\ \quad\vdots \\ \longleftarrow n\text{th observation} \end{array}$$

The convention is to express a vector $\boldsymbol{x}$ as a column. For a vector, the operation "transpose" means to transform a column vector to a row (or to transform a row to a column vector). The notation $\boldsymbol{x}'$ (or $\boldsymbol{x}^T$) denotes the transpose of a vector $\boldsymbol{x}$. Hence the jth row in the data matrix

$$\boldsymbol{x}_j' = [x_{j1} \ \cdots \ x_{jp}]$$

is expressed as the transpose of a vector, consisting of the p components of the jth observation. The column vector corresponding to the jth observation is

$$\boldsymbol{x}_j = \begin{bmatrix} x_{j1} \\ \vdots \\ x_{jp} \end{bmatrix} \in \mathbb{R}^p$$

which is a p-vector in the p-dimensional space of real numbers $\mathbb{R}^p$, where $\in$ means "belong to". Naturally,

$$\boldsymbol{x}_j = [x_{j1} \ \cdots \ x_{jp}]' = (\boldsymbol{x}_j')' = (\boldsymbol{x}_j^T)^T$$

Treating a p-vector as a point in $\mathbb{R}^p$,

$$\boldsymbol{x}_j \text{ (as well as } \boldsymbol{x}_j') \qquad \overset{\text{corresponding to}}{\text{“}=\text{”}} \qquad (x_{j1}, \cdots, x_{jp}) \in \mathbb{R}^p$$

The last expression in the above is in terms of the endpoint coordinates of the vector placed at origin. The coordinate notation is often used interchangeably with the corresponding vector or its transpose when the context is clear.

Observed data and random variables

In terms of notation, lowercase x often indicates an observed value or a realization of a random variable X, which is commonly denoted by a capital letter.

For example, x_{jk} often represents the jth observation of a random variable X_k.

However, lower case is often used as its corresponding random counterpart as well. Therefore whether x_{jk} is a scalar or a random variable depending heavily on the context.

Remarks on multivariate data

- In the above n-by-p data array, each observation is a row $\boldsymbol{x}_j'$, the transpose of the observed p-variate vector.

  An alternative layout of multivariate data exchanges the rows and columns: every column is an observation of p-components, every row consists of n observed values of one component variable of the p-variate vector. The advantage of this alternative layout is that the observed p-multivariate vector remains to be a column vector.
- Multivariate data analysis are useful when the component variables are correlated.
- Usually the measurements x_{ik}'s are real numbers, unless specified otherwise.

- Classical multivariate application focuses on the case $n > p$, with fixed p and fixed n.

  Classical multivariate asymptotic results considers the case when $n \to \infty$, for fixed p.

- Comparison with univariate data matrix and regression models

Let $z_1, \cdots, z_r$ denote explanatory variables or inputs (rather than x's, to avoid overuse of x notation here). Let Y be a univariate response variable in a linear regression model. The model conditioned on input variables is

$$Y = \beta_0 + \beta_1 z_1 + \cdots + z_r\beta_r + \varepsilon = \boldsymbol{z}'\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2). \tag{2}$$

Coefficients $\beta_1, \cdots, \beta_r$ are model parameters. Assuming the observed independent outcome of Y is y_j when z variables are set at $z_j' = (z_{j1}, \cdots, z_{jr})$, for $j = 1, \cdots, n$. The model in terms of the observed data becomes

$$y_j = \beta_0 + \beta_1 z_{j1} + \cdots + \beta_r z_{jr} + \varepsilon_j, \qquad j = 1, \cdots, n. \tag{3}$$

In element-wise vector-matrix form, the model to be fitted (to estimate the parameters) with the observed data can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_j \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & \cdots & z_{1r} \\ 1 & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_{j1} & \cdots & z_{jr} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_{n1} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_j \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The abbreviated form in matrix notation is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$$

Indexed the matrices by their dimensions,

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{Z}_{n\times(1+r)}\boldsymbol{\beta}_{(1+r)\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \qquad \boldsymbol{\varepsilon}_{n\times 1} \sim N(0, \sigma^2 \boldsymbol{I}_n)$$

In multivariate case, the jth observed outcome is a p-variate vector $(y_{j1}, \cdots, y_{jp})$ instead of a univariate y_j, and each component variable is fitted with a model formatted as (2) and (3). The multivariate linear regression model in terms of observed data has the form

$$\begin{bmatrix} y_{11} & \cdots & y_{1p} \\ y_{21} & \cdots & y_{2p} \\ \vdots & \cdots & \vdots \\ y_{j1} & \cdots & y_{jp} \\ \vdots & \cdots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & \cdots & z_{1r} \\ 1 & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_{j1} & \cdots & z_{jr} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_{n1} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_{01} & \cdots & \beta_{0p} \\ \beta_{11} & \cdots & \beta_{1p} \\ \vdots & \vdots & \vdots \\ \beta_{r1} & \cdots & \beta_{rp} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} & \cdots & \epsilon_{1p} \\ \epsilon_{21} & \cdots & \epsilon_{2p} \\ \vdots & \cdots & \vdots \\ \epsilon_{j1} & \cdots & \epsilon_{jp} \\ \vdots & \cdots & \vdots \\ \epsilon_{n1} & \cdots & \epsilon_{np} \end{bmatrix}$$

The corresponding matrix notation for the multivariate regression model with n observations is

$$\boldsymbol{Y}_{n\times p} = \boldsymbol{Z}_{n\times(1+r)}\boldsymbol{\beta}_{(1+r)\times p} + \boldsymbol{\epsilon}_{n\times p}$$

where for each $k = 1, \cdots, p$,

$$\boldsymbol{\epsilon}_k = [\epsilon_{1k} \ \cdots \ \epsilon_{nk}]' \sim N(0, \sigma_k^2 \boldsymbol{I}_n)$$

Multivariate regression is of interest when the component variables have dependence structures.

Multivariate analysis focuses on multivariate responses $\boldsymbol{Y}_{n\times p}$, with or without explanatory variables.

More details will be in the section "Multivariate Linear Regression".

- Multivariate method has also been used on the explanatory variables, such as predictor reduction in regression.
- Commonly the notations $\boldsymbol{X}, X$, and x are used for multivariate matrix, random variable or vector, and observed outcome, respectively.

2 Descriptive statistics

For $k = 1, \cdots, p$, the sample mean for the kth variable based on n observations is

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$$

(Question: Is it meaningful to also consider sample mean for each observation across vector components $k = 1, \cdots, p$?)

The overall sample mean vector $\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}$.

For $k = 1, \cdots, p$, the sample variance for the kth variable based on n observations is

$$s_k^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2, \qquad \left(\text{Alternatively} \quad s_k^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2\right)$$

The denominator $n - 1$ instead of n is to ensure unbiasedness of s_k^2 as an estimator of the population variance σ_k^2.

The sample covariance and sample correlation between the ith and kth variables based on n observations are

$$s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k), \qquad r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{s_{ik}}{s_i \times s_k}$$

The overall sample variance-covariance matrix and correlation matrix based on n observations are

$$\boldsymbol{S} = [s_{ik}]_{p\times p} \quad with \quad s_{kk} = s_k^2, \qquad \boldsymbol{R} = [r_{ik}]_{p\times p} \quad with \quad r_{kk} \equiv 1.$$

From matrix algebra point of view, the sample covariance and correlation matrix can be viewed as linear mappings of white noise to the observed data with given dependence structure in terms of variance-covariance.

For anyone yearning for a single number summary of data variation,

$$|\boldsymbol{S}| = det(\boldsymbol{S})$$

is the generalized sample variance, which is the determinant of the sample variance-covariance matrix.

Vector-matrix representation

- The overall sample mean vector can be written as a sum of n vectors:

$$\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} x_{j1}/n \\ \vdots \\ \sum_{j=1}^{n} x_{jp}/n \end{bmatrix} = \frac{1}{n}\sum_{j=1}^{n}\begin{bmatrix} x_{j1} \\ \vdots \\ x_{jp} \end{bmatrix} = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j$$

  where $\boldsymbol{x}_j$ is the jth observation, a p-vector $\in \mathbb{R}^p$.

- The sample covariance matrix $\boldsymbol{S}$ can be expressed as a sum of n matrices.

$$\boldsymbol{S} = [s_{ik}]_{p\times p} = \left[\frac{1}{n-1}\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k)\right]_{p\times p} = \frac{1}{n-1}\sum_{j=1}^{n}\left[(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k)\right]_{p\times p} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{x}_j-\bar{\boldsymbol{x}})(\boldsymbol{x}_j-\bar{\boldsymbol{x}})'$$

  where we use the vector product

$$(\boldsymbol{x}_j-\bar{\boldsymbol{x}})(\boldsymbol{x}_j-\bar{\boldsymbol{x}})' = \begin{bmatrix} x_{j1}-\bar{x}_1 \\ x_{j2}-\bar{x}_2 \\ \vdots \\ x_{jp}-\bar{x}_p \end{bmatrix} [x_{j1}-\bar{x}_1, x_{j2}-\bar{x}_2, \cdots, x_{jp}-\bar{x}_p] = \Big[(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k)\Big]_{i,k=1,\cdots,p}$$

  Note that

$$(\boldsymbol{x}_j-\bar{\boldsymbol{x}})(\boldsymbol{x}_j-\bar{\boldsymbol{x}})'$$

  is a $p\times p$ matrix for each fixed j, $\boldsymbol{x}_j$ is the jth observation vector, $\bar{\boldsymbol{x}}$ is the sample mean vector.

Vector-matrix expressions can be useful. For example, they are used in the proof of positive-definiteness of $\boldsymbol{S}$ below, in the derivation of properties of sample covariance and correlation matrix.

Properties of sample correlation of two component variables

- r_{jk} is the sample correlation of component variables j and k, which is not affected by whether n or $n-1$ is used in the calculation of component variable sample variance s_k^2.
- $r_{ik} \in [-1, 1]$.
- r_{ik} is a scale-invariant measure.
- r_{ik} is the Pearson correlation coefficient, which measures linear and only linear correlation.
- There exist other more general measures of dependence. For example, Kendall's τ and Spearman's ρ, as described in the section below.

Properties of sample covariance and correlation matrix

- Sample covariance matrix $\boldsymbol{S}$ and sample correlation matrix $\boldsymbol{R}$ are symmetric ($\boldsymbol{S}' = \boldsymbol{S}, \boldsymbol{R}' = \boldsymbol{R}$).

- $\boldsymbol{S}$ and $\boldsymbol{R}$ and positive semi-definite, which means $\boldsymbol{v}'\mathbf{S}\boldsymbol{v} \geq 0$ and $\boldsymbol{v}'\mathbf{R}\boldsymbol{v} \geq 0$, for any p-vector $\boldsymbol{v}$.

  Proof. Use the useful expression of $\boldsymbol{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{x}_j-\bar{\boldsymbol{x}})(\boldsymbol{x}_j-\bar{\boldsymbol{x}})'$,

$$\boldsymbol{v}'\mathbf{S}\boldsymbol{v} = \frac{1}{n-1}\sum_{j=1}^{n}\boldsymbol{v}'(\boldsymbol{x}_j-\bar{\boldsymbol{x}})(\boldsymbol{x}_j-\bar{\boldsymbol{x}})'\boldsymbol{v} = \frac{1}{n-1}\sum_{j=1}^{n}\left|\boldsymbol{v}'(\boldsymbol{x}_j-\bar{\boldsymbol{x}})\right|^2 \geq 0$$

  for any p-vector $\boldsymbol{v}$. Therefore covariance matrix $\boldsymbol{S}$ is positive semi-definite.
  To show that correlation matrix $\boldsymbol{R}$ is positive semi-definite, replace each x_{jk} by $x_{jk}^* = x_{jk}/s_k$ for $k = 1, \cdots, p$, and $j = 1, \cdots, n$. Then the sample covariance matrix of the n p-variate observations of $\boldsymbol{x}_j^*$ is $\boldsymbol{S}^* = \boldsymbol{R}$, the sample correlation matrix of the original $\boldsymbol{x}_j$'s. Thus

$$\boldsymbol{v}'\mathbf{R}\boldsymbol{v} = \boldsymbol{v}'\mathbf{S}^*\boldsymbol{v} \geq 0$$

  by the positive semi-definite property of covariance matrix. Therefore correlation matrix $\boldsymbol{R}$ is also positive semi-definite.

  □

- An alternative proof of the positive semi-definite property

  Notice that $y_j = \boldsymbol{v}'\boldsymbol{x}_j = \boldsymbol{x}_j'\boldsymbol{v}$ can be viewed as the jth observation of a univariate variable Y for $j = 1, \cdots, n$. Then for any p-vector $\boldsymbol{v}$,

$$\begin{aligned}\boldsymbol{v}'\mathbf{S}\boldsymbol{v} &= \frac{1}{n-1}\sum_{j=1}^{n}\boldsymbol{v}'(\boldsymbol{x}_j-\bar{\boldsymbol{x}})(\boldsymbol{x}_j-\bar{\boldsymbol{x}})'\boldsymbol{v} \\ &= \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{v}'\boldsymbol{x}_j-\boldsymbol{v}\bar{\boldsymbol{x}})(\boldsymbol{x}_j{}'\boldsymbol{v}-\bar{\boldsymbol{x}}'\boldsymbol{v}) \\ &= \frac{1}{n-1}\sum_{j=1}^{n}(y_j-\bar{y})(y_j-\bar{y}) = \frac{1}{n-1}\sum_{j=1}^{n}(y_j-\bar{y})^2\end{aligned}$$

  This is the sample covariance of the univariate variable y, which is always non-negative.

  Consequently the above equations provide another proof that $\boldsymbol{S}$ is positive semi-definite.

- If $\boldsymbol{v}'\mathbf{S}\boldsymbol{v} > 0$, $\boldsymbol{v}'\mathbf{R}\boldsymbol{v} > 0$, for any p-vector $\boldsymbol{v} \neq 0$, then $\boldsymbol{S}$ and $\boldsymbol{R}$ and (strictly) positive definite.

- When $\boldsymbol{S}$ is only positive semi-definite but not positive definite, then there is $\boldsymbol{v} \neq 0, Var(\boldsymbol{X}\boldsymbol{v}) = \boldsymbol{v}'\mathbf{S}\boldsymbol{v} = 0$. Thus $\boldsymbol{X}\boldsymbol{v}$ has component $\equiv c$, a constant (with probability 1). In other words, the p columns of $\boldsymbol{X}$ are linearly dependent, $\boldsymbol{X}$ lies on a lower dimension hyperplane, perpendicular to a non-zero vector in $\mathbb{R}^p$.

- Let $diag(\boldsymbol{S}) = diag(s_1^2, \cdots, s_p^2)$ be the diagonal matrix with $s_1^2, \cdots, s_p^2$ as the diagonal elements and zero elsewhere. Define $D = (diag(\mathbf{S}))^{1/2} = diag(s_1, \cdots, s_p)$. Then (exercise)

$$\boldsymbol{S} = D\boldsymbol{R}D, \qquad \boldsymbol{R} = D^{-1}\boldsymbol{S}D^{-1}$$

Remarks

- If s_{ik} uses denominator n instead of $n-1$, the notation for the corresponding covariance matrix $[s_{ik}]_{p\times p}$ is $\boldsymbol{S}_n$ instead of $\boldsymbol{S}$.

- Sometimes we may consider using centered data $\boldsymbol{X}_c = [x_{jk} - \bar{x}_j]_{n\times p}$. then the covariance matrix has a simplified and useful expression

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}_c'\boldsymbol{X}_c$$

- A geometric interpretation of the generalized variance is $|\boldsymbol{S}| = V_p^2/(n-1)^p$, where V_p is the volume generated by the p deviation (or centered) vectors.

3 Alternative measures of correlatedness and dependence

The sample correlation coefficient formula is the Pearson correlation coefficient which measures linear correlation. There are other metrics aim at measuring more general dependence structures among variables.

3.1 Kendall's rank correlation coefficient τ

Kendall's rank correlation coefficient ($a.k.a.$ Kendall's τ) measures similarity of two variables by comparing the relative ordering of the two sets of ranks.

Let $(x_i, y_i), i = 1, \cdots, n$ be observations of a bivariate random vector (X, Y). Consider all $\frac{1}{2}n(n-1)$ pairs $\{(x_i, y_i), (x_j, y_j)\}$ with $i < j$. The pair is called concordant if $(x_j - x_i)(y_j - y_i) > 0$ and discordant if $(x_j - x_i)(y_j - y_i) < 0$. Kendall's tau for the data (sample) is defined as

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

where n_c is the number of concordant pairs and n_d is the number of discordant pairs.

Remarks on Kendall's τ

- $\tau \in [-1, 1]$. $\tau = 1$ is the case of perfect agreement, $\tau = -1$ is the case of complete ranking reversal.
- There are various modifications of τ in the case of ties.
- $n_c - n_d$ can be written as $\sum_{i<j} sign(x_j - x_i) sign(y_j - y_i)$, where the $sign$ function is defined as

$$sign(x) = \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0. \end{cases}$$

  Concordant pairs have $sign(x_j - x_i) = sign(y_j - y_i)$, discordant pairs $sign(x_j - x_i) = -sign(y_j - y_i)$. Thus

$$\tau = \frac{\sum_{i<j} sign(x_j - x_i) \cdot sign(y_j - y_i)}{\frac{1}{2}n(n-1)}$$

  If we view (x_i, y_i) as observed values from independent random vectors (X_i, Y_i), then kendal's τ for an i.i.d. (independently identically distributed) random sample can be written as

$$\tau = \frac{\sum_{i<j} sign(X_j - X_i) \cdot sign(Y_j - Y_i)}{\frac{1}{2}n(n-1)}$$

- For independent bivariate random vectors (X_1, Y_1) and (X_2, Y_2) with the same continuous cumulative joint distribution function, Kendall's τ can be defined by the joint probability function,

$$\tau = \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] \; - \; \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

  - for two (continuous) independent bivariate random variables (X_1, Y_1) and (X_2, Y_2) of the same distribution, an alternative expression is $\tau = 2\mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1$ (Exercise).
  - when (X_1, Y_1) and (X_2, Y_2) and independent and have the same continuous cumulative distribution function (CDF) $F(x, y) = \mathbb{P}(X_i < x, Y_i < y)$, then Kendall's τ can be defined by the joint CDF as $\tau = 4 \iint_{\mathbb{R}^2} F(x, y) dF(x, y) - 1$. (Exercise)

- Relation of τ and ρ for normal sample

  If (X, Y) are of bivariate normal distribution with correlation coefficient r, then Kendall's τ gives slightly weaker correlation than r other than the extreme of 0 and ±1, with the relation

$$\tau = \frac{2}{\pi} \arcsin r$$

  (The proof involves a few steps of variable transformations and inverse trigononometric function relations and is omitted here.)

  The formula holds for more general distributions such as elliptical distributions (e.g., see Lindskog et al. 2001, Kendall's tau for Elliptical distributions), and provides a useful alternative for the estimation of r.

- Invariance property

  By definition, Kendall's τ depends on relative orders of data within each variable only, thus it is invariant under strictly monotone transformation of the variables. This property is useful in the analysis of large data, such as in estimation of covariance and precision matrices for samples beyond Gaussian distributions.

3.2 Spearman's rank correlation coefficient ρ

Let $(x_i, y_i), i = 1, \cdots, n$ be the component-wise ranks of n observations $(X_i, Y_i), i = 1, \cdots, n$ of a bivariate random vector (X, Y), each individual variable is ordered and indexed by i, so that x_i is the ith largest among all x's, and y_i is the ith largest among all y's, (or vise versa, meaning letting x_i be the ith smallest among all x's, and y_i be the ith smallest among all y's).

The Spearman's rank correlation coefficient is the Pearson correlation coefficient on the ranks.

$$\rho_s = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{j=1}^n (y_j - \bar{y})^2}}$$

In the absence of ties, Spearman's ρ can be expressed as (exercise)

$$\rho_s = 1 - \frac{6 \sum_{i=1}^n (x_i - y_i)^2}{n(n^2 - 1)}$$

For bivariate random vector (X_1, X_2) of continuous cumulative distribution function $F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$ and of marginal cumulative distribution functions $F_1(x_1) = P(X_1 \leq x_1), F_2(x_2) = P(X_2 \leq x_2)$, Spearman's ρ is the correlation of $F_1(X_1)$ and $F_2(X_2)$,

$$\rho_s = Corr\Big(F_1(X_1), F_2(X_2)\Big)$$

Then

$$\rho_s = 12 \iint F_1(x_1) F_2(x_2) dF(x_1, x_2) - 3$$

Proof. For $i = 1, 2$, for any $u \in [0, 1]$, since F_i is necessarily continuous,

$$P(F_i(X_i) \le u) = P(X_i \le F_i^{-1}(u)) = F_i(F_i^{-1}(u)) = u, \qquad where \quad F_i^{-1}(u) = \inf\{t, F_i(t) \ge u\}.$$

Therefore $F_i(X_i) \sim U(0, 1)$, the uniform distribution on $(0, 1)$, which is of mean $\frac{1}{2}$ and variance $\frac{1}{12}$.
By the definition of ρ_s,

$$\rho_s = \frac{Cov[F_1(X_1), F_2(X_2)]}{\sqrt{V(F_1(X_1))V(F_2(X_2))}} = 12 \times Cov[F_1(X_1), F_2(X_2)]$$

Using the notation $F(dx) = dF(x) = f(x)dx$ (when corresponding density f exists),

$$\begin{aligned}
Cov[F_1(X_1), F_2(X_2)] &= E\left(F_1(X_1) - \frac{1}{2}\right)\left(F_2(X_2) - \frac{1}{2}\right) \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \left(F_1(x_1) - \frac{1}{2}\right)\left(F_2(x_2) - \frac{1}{2}\right) F(dx_1, dx_2) \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} F_1(x_1)F_2(x_2)F(dx_1, dx_2) - \frac{1}{2}\int_{\mathbb{R}}\int_{\mathbb{R}} F_1(x_1)F(dx_1, dx_2) \\
&\quad - \frac{1}{2}\int_{\mathbb{R}}\int_{\mathbb{R}} F_2(x_2)F(dx_1, dx_2) + \frac{1}{4}\int_{\mathbb{R}}\int_{\mathbb{R}} F(dx_1, dx_2)
\end{aligned}$$

In the last expression, the integral in the second term

$$\int_{\mathbb{R}}\int_{\mathbb{R}} F_1(x_1)F(dx_1, dx_2) = \int_{x_1 \in \mathbb{R}} F_1(x_1) \int_{x_2 \in \mathbb{R}} F(dx_1, dx_2) = \int_{\mathbb{R}} F_1(x_1)F_1(dx_1) = \int_0^1 udu = \frac{1}{2}$$

Similarly, the integral in the third term

$$\int_{\mathbb{R}}\int_{\mathbb{R}} F_2(x_2)F(dx_1, dx_2) = \int_{x_2 \in \mathbb{R}} F_2(x_2) \int_{x_1 \in \mathbb{R}} F(dx_1, dx_2) = \int_0^1 F_2(x_2)F_2(dx_2) = \int_0^1 udu = \frac{1}{2}$$

The last term $= \frac{1}{4}$, and the rest follows. □

An example — Comparisons of correlation measures

Consider the heights (h_i's in cm) and weights (w_i's in kg) of four subjects. $(Height, Weight)$ are dependent bivariate random variables. The observations (measurements) are in the table below, along with the correlations calculated by Pearson, Kendall, and Spearman methods. The values are quite different. Why, exactly?

Subject i	(h_i, w_i)	h_i rank	w_i rank	Concordant pairs	Discordant pairs	Correlation
A	(160, 45)	1	1	AB, AC, AD	BC, BD, CD	Pearson's $r = 0.7$
B	(170, 61)	2	4			Kendall's $\tau = 0.0$
C	(180, 60)	3	3			Spearman's $\rho = 0.2$
D	(190, 59)	4	2			

Discussions

If we change the value 45 in Subject A, in which situations that only one of the three correlations would change?

Example R commands for various correlations:

```
> cor(c(45,61,60,59),c(160,170,180,190),method="pearson")    # default
> cor(c(45,61,60,59),c(160,170,180,190),method="spearman")
> cor(c(45,61,60,59),c(160,170,180,190),method="kendall")
```

Another example

With slight modifications to the example above, the following example produces different signs using different metrics of correlation.

Subject i	(h_i, w_i)	h_i rank	w_i rank	Concordant pairs	Discordant pairs	Correlation
A	(150, 59.1)	1	2	AB, AC	AD, BC, BD, CD	Pearson's $r = 0.005$
B	(170, 61.0)	2	4			Kendall's $\tau = -1/3$
C	(180, 60.0)	3	3			Spearman's $\rho = -0.4$
D	(190, 59.0)	4	1			

Remarks

Two variables can be dependent or “related”' in various manners. It is important to know the measure used quantifying the dependence or correlated-ness.

Note Related contents in the textbook by J&W: chapters 1, 2 and 3.