Journal of Multivariate Analysis 128 (2014) 210–225

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Kendall's tau for hierarchical data

H. Romdhani, L. Lakhal-Chaieb, L.-P. Rivest *

Département de Mathématiques et de Statistique, Université Laval, Pavillon Alexandre-Vachon 1045, av. de la Médecine,
Québec G1V 0A6, Canada

ARTICLE INFO

Article history:
Received 19 February 2013
Available online 4 April 2014

AMS subject classifications:
primary 62H20
secondary 62H15

Keywords:
Exchangeable copula
Hierarchical Archimedean copula
Hierarchical meta-elliptical copula
Intra class independence test
Kendall's tau
Pitman efficiency

ABSTRACT

This paper is concerned with hierarchical data having three levels. The level 1 units are nested in the level 2 units or subclusters which are themselves nested in the level 3 clusters. The model for this data is assumed to fulfill some symmetry assumptions. The level 1 units within each subcluster are exchangeable and a permutation of the subclusters belonging to the same cluster leaves the model unchanged. We are interested in measuring the dependence associated to clusters and subclusters respectively. Two exchangeable Kendall's tau are proposed as non parametric measures of these two associations and estimators for these measures are proposed. Their asymptotic properties are then investigated under the proposed hierarchical model for the data. These statistics are then used to estimate the intra-class correlation coefficients for data drawn from elliptical hierarchical distributions. Hypothesis tests for the cluster and subcluster effects based on the proposed estimators are developed and their performances are assessed using Pitman efficiencies and a Monte Carlo study.

1. Introduction

Hierarchical data structures are commonly found in many application areas of statistics especially in social sciences but also in other fields such as economics, finance and risk management. The hierarchy arises naturally from the organization of the data: the variable of interest is observed on units that are grouped into subclusters that are themselves grouped into clusters. Education provides a well known example where scores are observed on students which are clustered in schools and the schools themselves are grouped in geographical regions. In this paper, students, regions and schools are called units, clusters and subclusters, respectively.

Statistical models for hierarchical data characterize the association within and between subclusters. Multilevel models, also known as hierarchical or nested linear models, investigate the variation at different levels of the hierarchy. The standard reference for these models is [4]; most of the models presented there are based on the normality assumption. In survey sampling, a multistage sample design is considered in the presence of a hierarchical data structure [13, chapter 4]. In this context, the dependence levels associated to the subclusters and to the clusters respectively, are typically of prime interest as the precision of the statistical methods applied to these data depend on the strength of this dependence.

In the bivariate case, Kendall's tau is a measure of association defined as the probability of concordance minus the probability of discordance. Joe [6] defined an ordering of multivariate concordance and constructed a multivariate Kendall's tau; see also [2]. For clustered data, Romdhani, Rivest and Lakhal-Chaieb [12] introduced an exchangeable version of Kendall's tau as a measure of intra cluster association. The empirical counterpart of the latter association measure is

* Corresponding author.
E-mail addresses: hela.romdhani.1@ulaval.ca (H. Romdhani), Lajmi.Lakhal@mat.ulaval.ca (L. Lakhal-Chaieb), Louis-Paul.Rivest@mat.ulaval.ca (L.-P. Rivest).

http://dx.doi.org/10.1016/j.jmva.2014.03.016
0047-259X/© 2014 Published by Elsevier Inc.

computed by considering all possible pairs of bivariate observations such that the two elements of a pair come from different clusters. The properties of the resulting estimator were investigated under an exchangeable multivariate distribution model.

This paper considers three level hierarchical data. It proposes two association measures based on the exchangeable Kendall's tau. The first one, associated to the subclusters, measures the association between two units from the same subcluster and is defined by considering pairs of bivariate vectors drawn from two subclusters coming from different clusters. The second one is related to the clusters and measures the association between two units from the same cluster but different subclusters. It is defined by considering pairs of bivariate vectors coming from two different clusters such that each observation of a bivariate observation is drawn from two different subclusters. As will be seen in Sections 4 and 5, these statistics allow one to perform valid inference for hierarchical data when the normality assumption is questionable.

The sampling distributions of the two association measures based on the exchangeable Kendall's tau are investigated under a general model for hierarchical data. These models are presented in Section 2 along with the different association measures. In Section 3, we give estimators for the two exchangeable Kendall's tau and investigate their asymptotic distributions. Section 4 presents two estimators for the intra class correlation coefficients for elliptical distributions; one is the standard moment estimator and the other is based on Kendall's tau. Section 5 investigates tests for the cluster and subcluster effects based on Kendall's tau estimators proposed in Section 3. The cluster effect test's performance is assessed using Pitman efficiencies and a simulation study is conducted for the two tests. Section 6 provides a numerical example. Proofs and technical details can be found in the appendices.

2. Models for nested data

2.1. A general model for three level data

Let I denote the number of clusters, n_i the number of subclusters in cluster i, $i = 1, \ldots, I$, and m_{ij}, for $j = 1, \ldots, n_i$, the size of the jth subcluster of cluster i. Let $Y_{ij\ell}$, $i = 1, \ldots, I, j = 1, \ldots, n_i$ and $\ell = 1, \ldots, m_{ij}$ denotes the random variable for the ℓth unit of the jth subcluster of cluster i. The random vector representing the data in cluster i is the $N_i \times 1$ vector $\mathbf{Y}_i = (\mathbf{Y}_{i1}^T, \ldots, \mathbf{Y}_{in_i}^T)^T$ where $N_i = \sum_j^{n_i} m_{ij}$ is the total number of observations in the ith cluster and $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijm_{ij}})^T$ denotes the $m_{ij} \times 1$ random vector for subcluster j of cluster i. Measurements from different clusters are assumed to be independent.

The dependence within a cluster is modeled using a family of cumulative distribution functions (cdf) $\{F^{(n)}_{m_1,m_2,\ldots,m_n}(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) : \mathbf{Y}_j \in \Re^{m_j}, \sum_j m_j = N\}$ indexed by $(n, m_1, \ldots, m_n)$. This family is assumed to satisfy the following permutability property: let $\mathbf{y}_j$ represent an $m_j \times 1$ vector for $j = 1, \ldots, n$ and let P_j be a permutation matrix of dimension $m_j \times m_j$ for $j = 1, \ldots, n$. Then

$$F^{(n)}_{m_1,\ldots,m_n}(\mathbf{y}_1, \ldots, \mathbf{y}_n) = F^{(n)}_{m_{\pi(1)},\ldots,m_{\pi(n)}}(P_{\pi(1)}\mathbf{y}_{\pi(1)}, \ldots, P_{\pi(n)}\mathbf{y}_{\pi(n)}),$$

for any permutation $\{\pi(1), \ldots, \pi(n)\}$ of the integers $\{1, \ldots, n\}$. The joint cdf is then invariant to permutations within the subclusters and between the subclusters themselves. We assume in addition that

$$F^{(n)}_{m_1,\ldots,m_n}(\mathbf{y}_1, \ldots, \mathbf{y}_{n-1}, (\infty, \ldots, \infty)^T) = F^{(n-1)}_{m_1,\ldots,m_{n-1}}(\mathbf{y}_1, \ldots, \mathbf{y}_{n-1}).$$

The cdf of the vector $\mathbf{Y}_j$ is given by $F^{(1)}_{m_j}(\mathbf{y}_j)$ and is assumed to be closed under margins that is $F^{(1)}_{m_j}(y_{j1}, \ldots, y_{j,m_j-1}, \infty) = F_{m_j-1}(y_{j1}, \ldots, y_{j,m_j-1})$. In the notation of [8], the cdf $F^{(n)}_{m_1,\ldots,m_n}$ is h-extendible.

These assumptions have several implications. They imply that all the variables $Y_{ij\ell}$ have the same marginal cdf $F(y) = F^{(1)}_1(y)$. The dependence between two units of the same subcluster is characterized by a common bivariate cdf given by $F_s(y_1, y_2) = F^{(1)}_2((y_1, y_2)^T)$ while the common bivariate cdf of two units from two different subclusters of a cluster is $F_c(y_1, y_2) = F^{(2)}_{1,1}(y_1, y_2)$. The indices c and s refer to the clusters and the subclusters levels respectively. Two special cases are of interest. If the N random variables of a cluster are exchangeable, then there are no subcluster effect and $F_c(y_1, y_2) = F_s(y_1, y_2)$. It may also happen that the subclusters within a cluster are independent; this implies $F_c(y_1, y_2) = F(y_1)F(y_2)$. Examples of families of cdf $F^{(n_i)}_{m_{i1},\ldots,m_{in_i}}$ satisfying these conditions are presented in the next two subsections.

2.2. The standard normal and elliptical distributions for hierarchical data

The standard normal random effect model for hierarchical data writes

$$Y_{ij\ell} = \mu + a_i + b_{j(i)} + \epsilon_{ij\ell}, \quad i = 1, \ldots, I, j = 1, \ldots, n_i, \ell = 1, \ldots, m_{ij} \tag{1}$$

where μ is the overall mean, a_i, $b_{j(i)}$ and $\epsilon_{ij\ell}$ are independent normal random variables with respective variances σ_a^2, σ_b^2 and σ^2. A positive value of σ_a^2 induces dependence between subclusters of the same cluster. A positive value of σ_b^2 makes the dependence between units of the same subcluster stronger than the dependence between those coming from two different subclusters of the same cluster. The intra cluster correlations associated to clusters and to subclusters are then respectively