Journal of Statistical Computation and Simulation, 2016
Vol. 86, No. 9, 1643–1659, http://dx.doi.org/10.1080/00949655.2015.1077386

Gaussian copula distributions for mixed data, with application in discrimination

F. Jiryaie[a], N. Withanage[b], B. Wu[c] and A.R. de Leon[d]*

[a]*Department of Statistics, Shahid Beheshti University, Tehran, Iran;* [b]*Department of Economics & Statistics, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka;* [c]*PAREXEL International, Billerica, MA, USA;* [d]*Department of Mathematics & Statistics, University of Calgary, Calgary, Canada*

(*Received 28 August 2014; accepted 25 July 2015*)

The construction of a joint model for mixed discrete and continuous random variables that accounts for their associations is an important statistical problem in many practical applications. In this paper, we use copulas to construct a class of joint distributions of mixed discrete and continuous random variables. In particular, we employ the Gaussian copula to generate joint distributions for mixed variables. Examples include the robit-normal and probit-normal-exponential distributions, the first for modelling the distribution of mixed binary-continuous data and the second for a mixture of continuous, binary and trichotomous variables. The new class of joint distributions is general enough to include many mixed-data models currently available. We study properties of the distributions and outline likelihood estimation; a small simulation study is used to investigate the finite-sample properties of estimates obtained by full and pairwise likelihood methods. Finally, we present an application to discriminant analysis of multiple correlated binary and continuous data from a study involving advanced breast cancer patients.

Keywords: conditional grouped continuous model; general location model; logit model; normal distribution; probit model; *t*-distribution

1. Introduction

Many statistical applications involve the collection and analysis of multivariate data comprising a mixture of discrete and continuous variables. Examples can be found in medicine (where continuous laboratory measurements may be included with such variables as the presence or absence of a certain symptom for each patient), in health studies (where data may involve a patient's choice of healthcare unit, his state of health, his global quality of life, along with a number of quantitative health-related variables), and in many other fields. Multivariate modelling of such data often leads to complications in practice due to a relative lack of standard models.

Factorization models directly specify the joint distribution as the product of a conditional distribution of one set of variables and a marginal distribution of the other. General location models (GLOMs) [1] are based on conditional Gaussian distributions and have received much attention in the literature. They assume conditional normality of continuous components and an arbitrary distribution for discrete components. A reverse factorization entails specifying a latent continuous multivariate distribution from which discrete variables are

*Corresponding author. Email: adeleon@ucalgary.ca