

Studies in Computational Intelligence 753

Vladik Kreinovich
Songsak Sriboonchitta
Nopasit Chakpitak *Editors*

Predictive Econometrics and Big Data

 Springer



Studies in Computational Intelligence

Volume 753


Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at <http://www.springer.com/series/7092>



Vladik Kreinovich · Songsak Sriboonchitta
Nopasit Chakpitak
Editors

Predictive Econometrics and Big Data

 Springer

Editors

Vladik Kreinovich
Computer Science Department
University of Texas at El Paso
El Paso, TX
USA

Nopasit Chakpitak
International College
Chiang Mai University
Chiang Mai
Thailand

Songsak Sriboonchitta
International College
Chiang Mai University
Chiang Mai
Thailand

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-319-70941-3 ISBN 978-3-319-70942-0 (eBook)
<https://doi.org/10.1007/978-3-319-70942-0>

Library of Congress Control Number: 2017959632

© Springer International Publishing AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland



Preface

Econometrics is a branch of economics that uses mathematical (especially statistical) methods to analyze economic systems, to forecast economic and financial dynamics, and to develop strategies for achieving desirable economic performance.

Traditional econometric techniques have been focused on the quantitative description of economic phenomena. However, the ultimate goal of econometrics—as well as the ultimate goal of science in general—is to predict future development of economics and to develop strategies that optimize the future state of economics. It is therefore desirable to develop techniques that are specifically aimed at predicting economic phenomena. Such predictive econometric techniques—and their applications to real-life economic and financial situations—are one of the main foci of this volume.

Another focus of this book is related to the fact that in the modern world, in which computers are ubiquitous, the amount of economic-related data generated and processed by these computers has grown exponentially. The amount of available economic data is so huge that many traditional statistical data processing algorithms are no longer capable of processing all these data in real time. To process this data, we need to utilize “big data” techniques specifically developed for processing such huge amounts of data and we need to develop big data versions of the state-of-the-art econometric techniques and algorithms. This is a new and promising direction in econometrics. Big data is the main subject of this volume’s keynote paper by Dr. Chaitanya Baru from the US National Science Foundation.

In addition to papers on predictive econometric techniques and on big data applications, this book also contains applications of more traditional statistical techniques to econometric problems.

We hope that this volume will help practitioners to learn how to apply new predictive and big data econometric techniques and help researchers to further improve the existing predictive and big data techniques and to come up with new ideas on how econometric techniques can utilize large amounts of data to make more accurate predictions.

We want to thank all the authors for their contributions and all anonymous referees for their thorough analysis and helpful comments.

The publication of this volume is partly supported by the Chiang Mai School of Economics (CMSE), Thailand. Our thanks go to Dean Pirut Kanjanakaron and CMSE for providing crucial support. Our special thanks go to Prof. Hung T. Nguyen for his valuable advice and constant support.

We would also like to thank Prof. Janusz Kacprzyk (Series Editor) and Dr. Thomas Ditzinger (Senior Editor, Engineering/Applied Sciences) for their support and cooperation in this publication.

September 2017

Vladik Kreinovich
Songsak Sriboonchitta
Nopasit Chakpitak

Contents

Keynote Address

Data in the 21st Century	3
Chaitanya Baru	

Fundamental Theory

Model-Assisted Survey Estimation with Imperfectly Matched Auxiliary Data	21
F. Jay Breidt, Jean D. Opsomer, and Chien-Min Huang	
COBra: Copula-Based Portfolio Optimization	36
Marc S. Paoletta and Paweł Polak	
Multiple Testing of One-Sided Hypotheses: Combining Bonferroni and the Bootstrap	78
Joseph P. Romano and Michael Wolf	
Exploring Message Correlation in Crowd-Based Data Using Hyper Coordinates Visualization Technique	95
Tien-Dung Cao, Dinh-Quyen Nguyen, and Hien Duy Tran	
Bayesian Forecasting for Tail Risk	122
Cathy W. S. Chen and Yu-Wen Sun	
Smoothing Spline as a Guide to Elaborate Explanatory Modeling	146
Chon Van Le	
Quantifying Predictive Uncertainty Using Belief Functions: Different Approaches and Practical Construction	157
Thierry Dencoux	
Kuznets Curve: A Simple Dynamical System-Based Explanation	177
Thongchai Dumrongpokaphan and Vladik Kreinovich	

A Calibration-Based Method in Computing Bayesian Posterior Distributions with Applications in Stock Market	182
Dung Tien Nguyen, Son P. Nguyen, Uyen H. Pham, and Thien Dinh Nguyen	
How to Estimate Statistical Characteristics Based on a Sample: Nonparametric Maximum Likelihood Approach Leads to Sample Mean, Sample Variance, etc.	192
Vladik Kreinovich and Thongchai Dumrongpookaphan	
How to Gauge Accuracy of Processing Big Data: Teaching Machine Learning Techniques to Gauge Their Own Accuracy	198
Vladik Kreinovich, Thongchai Dumrongpookaphan, Hung T. Nguyen, and Olga Kosheleva	
How Better Are Predictive Models: Analysis on the Practically Important Example of Robust Interval Uncertainty	205
Vladik Kreinovich, Hung T. Nguyen, Songsak Sriboonchitta, and Olga Kosheleva	
Quantitative Justification for the Gravity Model in Economics	214
Vladik Kreinovich and Songsak Sriboonchitta	
The Decomposition of Quadratic Forms Under Skew Normal Settings	222
Ziwei Ma, Weizhong Tian, Baokun Li, and Tonghui Wang	
Joint Plausibility Regions for Parameters of Skew Normal Family	233
Ziwei Ma, Xiaonan Zhu, Tonghui Wang, and Kittawit Autchariyapanitkul	
On Parameter Change Test for ARMA Models with Martingale Difference Errors	246
Haejune Oh and Sangyeol Lee	
Agent-Based Modeling of Economic Instability	255
Akira Namatame	
A Bad Plan Is Better Than No Plan: A Theoretical Justification of an Empirical Observation	266
Songsak Sriboonchitta and Vladik Kreinovich	
Shape Mixture Models Based on Multivariate Extended Skew Normal Distributions	273
Weizhong Tian, Tonghui Wang, Fengrong Wei, and Fang Dai	

Plausibility Regions on Parameters of the Skew Normal Distribution Based on Inferential Models	287
Xiaonan Zhu, Baokun Li, Mixia Wu, and Tonghui Wang	
Measures of Mutually Complete Dependence for Discrete Random Vectors	303
Xiaonan Zhu, Tonghui Wang, S. T. Boris Choy, and Kittawit Autchariyapanitkul	
Applications	
To Compare the Key Successful Factors When Choosing a Medical Institutions Among Taiwan, China, and Thailand	321
Tzong-Ru (Jiun-Shen) Lee, Yu-Ting Huang, Man-Yu Huang, and Huan-Yu Chen	
Forecasting Thailand's Exports to ASEAN with Non-linear Models	339
Petchaluck Boonyakunakorn, Pathairat Pastpipatkul, and Songsak Sriboonchitta	
Thailand in the Era of Digital Economy: How Does Digital Technology Promote Economic Growth?	350
Noppasit Chakpitak, Paravee Maneejuk, Somsak Chanaim, and Songsak Sriboonchitta	
Comparing Linear and Nonlinear Models in Forecasting Telephone Subscriptions Using Likelihood Based Belief Functions	363
Noppasit Chakpitak, Woraphon Yamaka, and Songsak Sriboonchitta	
Volatility in Thailand Stock Market Using High-Frequency Data	375
Saowaluk Duangin, Jirakom Sirisrisakulchai, and Songsak Sriboonchitta	
Technology Perception, Personality Traits and Online Purchase Intention of Taiwanese Consumers	392
Massoud Moslehpour, Ha Le Thi Thanh, and Pham Van Kien	
Analysis of Thailand's Foreign Direct Investment in CLMV Countries Using SUR Model with Missing Data	408
Chalerm Jaitang, Paravee Maneejuk, Aree Wiboonpongse, and Songsak Sriboonchitta	
The Role of Oil Price in the Forecasts of Agricultural Commodity Prices	422
Rossarin Osathanunkul, Chatchai Khiengamdee, Woraphon Yamaka, and Songsak Sriboonchitta	

Does Forecasting Benefit from Mixed-Frequency Data Sampling Model: The Evidence from Forecasting GDP Growth Using Financial Factor in Thailand	430
Natthaphat Kingnetr, Tanaporn Tungtrakul, and Songsak Sriboonchitta	
A Portfolio Optimization Between US Dollar Index and Some Asian Currencies with a Copula-EGARCH Approach	443
Ji Ma, Jianxu Liu, and Songsak Sriboonchitta	
Technical Efficiency Analysis of China's Agricultural Industry: A Stochastic Frontier Model with Panel Data	454
Ji Ma, Jianxu Liu, and Songsak Sriboonchitta	
Empirical Models of Herding Behaviour for Asian Countries with Confucian Culture	464
Munkh-Ulzii, Massoud Moslehpour, and Pham Van Kien	
Forecasting the Growth of Total Debt Service Ratio with ARIMA and State Space Model	492
Kobpongkit Navapan, Petchaluck Boonyakunakorn, and Songsak Sriboonchitta	
Effect of Macroeconomic Factors on Capital Structure of the Firms in Vietnam: Panel Vector Auto-regression Approach (PVAR)	502
Nguyen Ngoc Thach and Tran Thi Kim Oanh	
Emissions, Trade Openness, Urbanisation, and Income in Thailand: An Empirical Analysis	517
Rossarin Osathanunkul, Natthaphat Kingnetr, and Songsak Sriboonchitta	
Analysis of Risk, Rate of Return and Dependency of REITs in ASIA with Capital Asset Pricing Model	536
Rungrapee Phadkantha, Woraphon Yamaka, and Roengchai Tansuchat	
Risk Valuation of Precious Metal Returns by Histogram Valued Time Series	549
Pichayakone Rakpho, Woraphon Yamaka, and Roengchai Tansuchat	
Factors Affecting Consumer Visiting Spa Shop: A Case in Taiwan	563
Meng-Chun Susan Shen, I-Tien Chu, and Wan-Tran Huang	
The Understanding of Dependent Structure and Co-movement of World Stock Exchanges Under the Economic Cycle	573
Songsak Sriboonchitta, Chukiat Chaiboonsri, and Jittima Singvejsakul	
The Impacts of Macroeconomic Variables on Financials Sector and Property and Construction Sector Index Returns in Stock Exchange of Thailand Under Interdependence Scheme	590
Wilawan Srichaikul, Woraphon Yamaka, and Roengchai Tansuchat	

Measures of Mutually Complete Dependence for Discrete Random Vectors

Xiaonan Zhu¹, Tonghui Wang^{1(✉)}, S. T. Boris Choy²,
and Kittawit Autchariyapanitkul³

¹ Department of Mathematical Sciences, New Mexico State University,
Las Cruces, USA

{xzhu, twang}@nmsu.edu

² Discipline of Business Analytics, The University of Sydney, Sydney, Australia
boris.choy@sydney.edu.au

³ Faculty of Economics, Maejo University, Chiang Mai, Thailand
Kittawit_a@mju.ac.th

Abstract. In this paper, a marginal-free measure of mutually complete dependence for discrete random vectors through subcopulas is defined, which generalizes the corresponding results for discrete random variables. Properties of the measure are studied and an estimator of the measure is introduced. Several examples are given for illustration of our results.

Keywords: Discrete random vector · Mutually complete dependence
Dependence measure · Subcopula

1 Introduction

Complete dependence (or functional dependence) is an important concept in many aspects of our life, such as econometrics, insurance, finance, etc. Recently, measures of (mutually) complete dependence have been defined and studied by many authors, e.g. Rényi [8], Schweizer and Wolff [9], Lancaster [5], Siburg and Stoimenov [12], Trutschnig [16], Dette et al. [3], Tasena and Dhompongsa [14], Shan et al. [11], Tasena and Dhompongsa [15] and Boonmee and Tasena [2]. However, measures in above papers have some drawbacks. Some measures only work for continuous random variables or vectors and some measures rely on marginal distributions (See Sect. 2 for a summary of several important measures). To the best of our knowledge, none of previously proposed measures are marginal-free and can describe (mutually) complete dependence for discrete random vectors. To overcome this issue, in this paper, we define a marginal-free measure of (mutually) complete dependence for discrete random vectors by using subcopulas, which extends the corresponding results of discrete random variables given in [11] to multivariate cases.

This paper is organized as follows. Some necessary concepts and definitions, and several measures of (mutually) complete dependence are reviewed briefly in Sect. 2. A marginal-free measure of (mutually) complete dependence for discrete random vectors is defined and properties of this measure are studied in Sect. 3. An estimator of the measure is introduced in Sect. 4.

2 Preliminaries

Let (Ω, \mathcal{A}, P) be a probability space, where Ω is a sample space, \mathcal{A} is a σ -algebra of Ω and P is a probability measure on \mathcal{A} . A *random variable* is a measurable function from Ω to the real line \mathbb{R} , and for any integer $n \geq 2$, an *n-dimensional random vector* is a measurable function from Ω to \mathbb{R}^n . For any $a = (a_1, \dots, a_n)$ and $b = (b_1, \dots, b_n) \in \mathbb{R}^n$, we say $a \leq b$ if and only if $a_i \leq b_i$ for all $i = 1, \dots, n$. Let X and Y be random vectors defined on the same probability space. X and Y are said to be *independent* if and only if $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ for all x and y . Y is *completely dependent* (CD) on X if Y is a measurable function of X almost surely, i.e., there is a measurable function ϕ such that $P(Y = \phi(X)) = 1$. X and Y are said to be *mutually completely dependent* (MCD) if X and Y are completely dependent on each other.

Let E_1, \dots, E_n be nonempty subsets of \mathbb{R} and Q a real-valued function with the domain $\text{Dom}(Q) = E_1 \times \dots \times E_n$. Let $[a, b] = [a_1, b_1] \times \dots \times [a_n, b_n]$ such that all vertices of $[a, b]$ belong to $\text{Dom}(Q)$. The Q -volume of $[a, b]$ is defined by

$$\mathcal{V}_Q([a, b]) = \sum \text{sgn}(c)Q(c),$$

where the sum is taken over all vertices $c = (c_1, \dots, c_n)$ of $[a, b]$, and

$$\text{sgn}(c) = \begin{cases} 1, & \text{if } c_i = a_i \text{ for an even number of } i\text{'s,} \\ -1, & \text{if } c_i = a_i \text{ for an odd number of } i\text{'s.} \end{cases}$$

An *n-dimensional subcopula* (or *n-subcopula* for short) is a function C with the following properties [7].

- (i) The domain of C is $\text{Dom}(C) = D_1 \times \dots \times D_n$, where D_1, \dots, D_n are nonempty subsets of the unit interval $I = [0, 1]$ containing 0 and 1;
- (ii) C is *grounded*, i.e., for any $u = (u_1, \dots, u_n) \in \text{Dom}(C)$, $C(u) = 0$ if at least one $u_i = 0$;
- (iii) For any $u_i \in D_i$, $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$, $i = 1, \dots, n$;
- (iv) C is *n-increasing*, i.e., for any $u, v \in \text{Dom}(C)$ such that $u \leq v$, $\mathcal{V}_C(\{u, v\}) \geq 0$.

For any n random variables X_1, \dots, X_n , by Sklar's Theorem [13], there is a unique n -subcopula such that

$$H(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)), \quad \text{for all } (x_1, \dots, x_n) \in \overline{\mathbb{R}}^n,$$

where $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, H is the joint cumulative distribution function (c.d.f.) of X_1, \dots, X_n , and F_i is the marginal c.d.f. of X_i , $i = 1, \dots, n$. In addition, if X_1, \dots, X_n are continuous, then $\text{Dom}(C) = I^n$ and the unique C is called the *n-copula* of X_1, \dots, X_n . For more details about the copula theory, see [7].

Next, we are going to recall some measures of MCD and CD, which are equal to 0 if and only if two random variables (or vectors) are independent, and equal to 1 if

and only if they are MCD or CD. In 2010, Siburg and Stoimenov [12] defined an MCD measure for continuous random variables as

$$\omega(X, Y) = (3\|C\|^2 - 2)^{\frac{1}{2}}, \quad (1)$$

where X and Y are continuous random variables with the copula C and $\|\cdot\|$ is the Sobolev norm of bivariate copulas given by

$$\|C\| = \left(\int \int |\nabla C(u, v)|^2 du dv \right)^{\frac{1}{2}},$$

where $\nabla C(u, v)$ is the gradient of $C(u, v)$.

In 2013, Tasena and Dhompangsa [14] generalized Siburg and Stoimenov's measure to multivariate cases as follows. Let X_1, \dots, X_n be continuous variables with the n -copula C . Define

$$\delta_i(X_1, \dots, X_n) = \delta_i(C) = \frac{\int \dots \int [\partial_i C(u_1, \dots, u_n) - \pi_i C(u_1, \dots, u_n)]^2 du_1 \dots du_n}{\int \dots \int \pi_i C(u_1, \dots, u_n) (1 - \pi_i C(u_1, \dots, u_n)) du_1 \dots du_n},$$

where $\partial_i C$ is the partial derivative on the i th coordinate of C and $\pi_i C : I^{n-1} \rightarrow I$ is defined by $\pi_i C(u_1, \dots, u_{n-1}) = C(u_1, \dots, u_{i-1}, 1, u_i, \dots, u_{n-1})$, $i = 1, 2, \dots, n$. Let

$$\delta(X_1, \dots, X_n) = \delta(C) = \frac{1}{n} \sum_{i=1}^n \delta_i(C). \quad (2)$$

Then δ is an MCD measure of X_1, \dots, X_n .

In 2015, Shan et al. [11] considered discrete random variables. Let X and Y be two discrete random variables with the subcopula C . An MCD measure of X and Y is given by

$$\mu_t(X, Y) = \left(\frac{\|C\|_t^2 - L_t}{U_t - L_t} \right)^{\frac{1}{2}}, \quad (3)$$

where $t \in [0, 1]$ and $\|C\|_t^2$ is the discrete norm of C defined by

$$\|C\|_t^2 = \sum_i \sum_j \left\{ \left(tC_{\Delta i, j}^2 + (1-t)C_{\Delta i, j+1}^2 \right) \frac{\Delta v_j}{\Delta u_i} + \left(tC_{i, \Delta j}^2 + (1-t)C_{i+1, \Delta j}^2 \right) \frac{\Delta u_i}{\Delta v_j} \right\},$$

$$C_{\Delta i, j} = C(u_{i+1}, v_j) - C(u_i, v_j), \quad C_{i, \Delta j} = C(u_i, v_{j+1}) - C(u_i, v_j),$$

$$\Delta u_i = u_{i+1} - u_i, \quad \Delta v_j = v_{j+1} - v_j,$$

$$L_t = \sum_i (tu_i^2 + (1-t)u_{i+1}^2) \Delta u_i + \sum_j (tv_j^2 + (1-t)v_{j+1}^2) \Delta v_j,$$

and

$$U_t = \sum_i (tu_i + (1-t)u_{i+1}) \Delta u_i + \sum_j (tv_j + (1-t)v_{j+1}) \Delta v_j.$$

In 2016 Tasena and Dhompongsa [15] defined a measure of CD for random vectors. Let X and Y be two random vectors. Define

$$\omega_k(Y|X) = \left[\int \int \left| F_{Y|X}(y|x) - \frac{1}{2} \right|^k dF_X(x) dF_Y(y) \right]^{\frac{1}{k}},$$

where $k \geq 1$. The measure of Y CD on X is given by

$$\bar{\omega}_k(Y|X) = \left[\frac{\omega_k^k(Y|X) - \omega_k^k(Y^\perp|X^\perp)}{\omega_k^k(Y|Y) - \omega_k^k(Y^\perp|X^\perp)} \right]^{\frac{1}{k}}, \quad (4)$$

where X^\perp and Y^\perp are independent random vectors with the same distributions as X and Y , respectively.

In the same period, Boonmee and Tasena [2] defined a measure of CD for continuous random vectors by using *linkages* which were introduced by Li et al. [6]. Let X and Y be two continuous random vectors with the linkage C . The measure of Y being completely dependent on X is defined by

$$\zeta_p(Y|X) = \left[\int \int \left| \frac{\partial}{\partial u} C(u, v) - \Pi(v) \right|^p du dv \right]^{\frac{1}{p}}, \quad (5)$$

where $\Pi(v) = \prod_{i=1}^n v_i$ for all $v = (v_1, \dots, v_n) \in \mathbb{R}^n$.

From above summaries we can see that measures given by (1), (2) and (5) only work for continuous random variables or vectors. The measure defined by (3) only works for bivariate discrete random variables. The measure given by (4) relies on marginal distributions of random vectors. Thus it is worth considering marginal-free measures of CD and MCD for discrete random vectors.

3 An MCD Measure for Discrete Random Vectors

In this section, we identify $\mathbb{R}^{n_1+\dots+n_k}$ with $\mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_k}$, where n_1, \dots, n_k are positive integers. So any n -dimensional random vector X can be viewed as a tuple of n random variables, i.e., $X = (X_1, \dots, X_n)$, where X_1, \dots, X_n are random variables. Also, if $Y = (Y_1, \dots, Y_m)$ is an m -dimensional random vector, we use (X, Y) to denote the $(n+m)$ -dimensional random vector $(X_1, \dots, X_n, Y_1, \dots, Y_m)$. Let $\psi = (\psi_1, \dots, \psi_n): \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a function. ψ is said to be strictly increasing if and only if each component $\psi_i: \mathbb{R} \rightarrow \mathbb{R}$ is strictly increasing, i.e., for any a_i and $b_i \in \mathbb{R}$ such that $a_i < b_i$, we have $\psi_i(a_i) < \psi_i(b_i)$, $i = 1, \dots, n$.

We will focus on *discrete random vectors* in this section, i.e., they can take on at most a countable number of possible values. Let $X = (X_1, \dots, X_n) \in \mathcal{L}_1 \subseteq \mathbb{R}^n$ and $Y = (Y_1, \dots, Y_m) \in \mathcal{L}_2 \subseteq \mathbb{R}^m$ be two discrete random vectors defined on the same probability space (Ω, \mathcal{A}, P) . Their *joint* c.d.f. H , *marginal* c.d.f.'s F and G , and *marginal probability mass functions* (p.m.f.) f and g are defined, respectively, as follows.

$$H(x, y) = P(X \leq x, Y \leq y), \quad F(x) = P(X \leq x), \quad G(y) = P(Y \leq y),$$

$$f(x) = P(X = x), \quad \text{and} \quad g(y) = P(Y = y), \quad \text{for all } x \in \mathbb{R}^n \text{ and } y \in \mathbb{R}^m.$$

Also, we use F_i and f_i to denote the marginal c.d.f. and p.m.f. of the i th component X_i of X , $i = 1, \dots, n$, and use G_j and g_j to denote the marginal c.d.f. and p.m.f. of the j th component Y_j of Y , $j = 1, \dots, m$, respectively. For simplicity, we assume that $f(x) \neq 0$ and $g(y) \neq 0$ for all $x \in \mathcal{L}_1$ and $y \in \mathcal{L}_2$. If C is the subcopula of the $(n+m)$ -dimensional random vector (X, Y) , i.e.,

$$H(x, y) = C(u(x), v(y)), \quad \text{for all } x \in \mathbb{R}^n \text{ and } y \in \mathbb{R}^m,$$

where

$$u(x) = (u_1(x_1), \dots, u_n(x_n)) = (F_1(x_1), \dots, F_n(x_n)) \in I^n,$$

and

$$v(y) = (v_1(y_1), \dots, v_m(y_m)) = (G_1(y_1), \dots, G_m(y_m)) \in I^m,$$

for all $x \in \mathcal{L}_1$ and $y \in \mathcal{L}_2$, then C is said to be the *subcopula of X and Y* . In addition, for each vector $e \in E_1 \times \dots \times E_n$, where E_i is a countable subset of \mathbb{R} , let e_L be the *greatest lower bound* of e with respect to the coordinate-wise order, i.e., $e_L = (e'_1, \dots, e'_n)$ such that if there exists some element in E_i that is less than e_i , then e'_i is the greatest element in E_i so that $e'_i < e_i$, otherwise $e'_i = e_i$, $i = 1, \dots, n$. We use 1_n and ∞_n to denote the n -dimensional constant vector $(1, \dots, 1)$ and $(\infty, \dots, \infty) \in \mathbb{R}^n$.

To construct desired measures, a distance between two discrete random vectors is defined as follows.

Definition 1. Let X and Y be discrete random vectors. The *distance between the conditional distribution of Y given X and marginal distribution of Y* is defined by

$$\omega^2(Y|X) = \sum_{y \in \mathcal{L}_2} \sum_{x \in \mathcal{L}_1} [P(Y \leq y|X = x) - G(y)]^2 f(x)g(y). \quad (6)$$

From the above definition, we can obtain the following two results.

Lemma 1. For any discrete random vectors X and Y , we have $\omega^2(Y|X) \leq \omega_{\max}^2(Y|X)$, where

$$\omega_{\max}^2(Y|X) = \sum_{y \in \mathcal{L}_2} [G(y) - (G(y))^2] g(y).$$

Proof. By the definition,

$$\begin{aligned} \omega^2(Y|X) &= \sum_{y \in \mathcal{L}_2} \sum_{x \in \mathcal{L}_1} [P(Y \leq y|X = x) - G(y)]^2 f(x)g(y) \\ &= \sum_{y \in \mathcal{L}_2} \sum_{x \in \mathcal{L}_1} [P(Y \leq y|X = x)^2 - 2P(Y \leq y|X = x)G(y) + (G(y))^2] f(x)g(y) \\ &\leq \sum_{y \in \mathcal{L}_2} \sum_{x \in \mathcal{L}_1} [P(Y \leq y|X = x) - 2P(Y \leq y|X = x)G(y) + (G(y))^2] f(x)g(y) \end{aligned}$$

$$\begin{aligned}
&= \sum_{y \in \mathcal{L}_2} \sum_{x \in \mathcal{L}_1} [P(X=x, Y \leq y) - 2P(X=x, Y \leq y)G(y) + (G(y))^2 f(x)] g(y) \\
&= \sum_{y \in \mathcal{L}_2} [G(y) - (G(y))^2] g(y).
\end{aligned}$$

□

Lemma 2. Let X and Y be discrete random vectors. There is a function $\phi : \mathcal{L}_1 \rightarrow \mathcal{L}_2$ such that $\phi(X) = Y$ if and only if $\omega^2(Y|X) = \omega_{\max}^2(Y|X)$, i.e., $P(Y \leq y|X=x) = 0$ or 1 for all $(x, y) \in \mathcal{L}_1 \times \mathcal{L}_2$.

Proof. For “if” part, suppose that $P(Y \leq y|X=x) = 0$ or 1 for all $(x, y) \in \mathcal{L}_1 \times \mathcal{L}_2$. Then $\sum_{t \leq y} P(Y=t|X=x) = 0$ or 1. So $\sum_{t \leq y} P(X=x, Y=t) = 0$ or $P(X=x)$. Thus there exists a unique $y(x) \in \mathcal{L}_2$, which depends on x , such that $P(X=x, Y=y(x)) = P(X=x)$, and $P(X=x, Y=y) = 0$ for all $y \in \mathcal{L}_2$ with $y \neq y(x)$. Now if we define $\phi(x) = y(x)$ for all $x \in \mathcal{L}_1$, then $\phi(X) = Y$.

For “only if” part, suppose that $\phi(X) = Y$. Fix $x \in \mathcal{L}_1$. It is sufficient to show that $P(X=x, Y=y) = 0$ for all $y \neq \phi(x)$. Suppose that, on the contrary, there is $y' \in \mathcal{L}_2$ such that $y' \neq \phi(x)$ and $P(X=x, Y=y') \neq 0$, then there exists $\omega \in \Omega$ so that $X(\omega) = x$ and $Y(\omega) = y'$. So we have $\phi(X)(\omega) \neq Y(\omega)$. It's a contradiction. □

Now we can define a measure of CD for two discrete random vectors as follows.

Definition 2. For any discrete random vectors X and Y , the *measure of Y being completely dependent on X* is given by

$$\mu(Y|X) = \left[\frac{\omega^2(Y|X)}{\omega_{\max}^2(Y|X)} \right]^{\frac{1}{2}} = \left[\frac{\sum_{y \in \mathcal{L}_2} \sum_{x \in \mathcal{L}_1} [P(Y \leq y|X=x) - G(y)]^2 f(x) g(y)}{\sum_{y \in \mathcal{L}_2} [G(y) - (G(y))^2] g(y)} \right]^{\frac{1}{2}}. \quad (7)$$

Properties of the measure $\mu(Y|X)$ are given as follows.

Theorem 1. For any discrete random vectors X and Y , the measure $\mu(Y|X)$ has the following properties:

- (i) $0 \leq \mu(Y|X) \leq 1$;
- (ii) $\mu(Y|X) = 0$ if and only if X and Y are independent;
- (iii) $\mu(Y|X) = 1$ if and only if Y is a function of X ;
- (iv) $\mu(Y|X)$ is invariant under strictly increasing transformations of X and Y , i.e., if ψ_1 and ψ_2 are strictly increasing functions defined on \mathcal{L}_1 and \mathcal{L}_2 , respectively, then $\mu(\psi_2(Y)|\psi_1(X)) = \mu(Y|X)$.

Proof. Property (i) is obvious by Lemma 1. For Property (ii), note that $\mu(Y|X) = 0$ if and only if $\omega^2(Y|X) = 0$. It is equivalent to $P(Y \leq y|X=x) = G(y)$ for all $(x, y) \in \mathcal{L}_1 \times \mathcal{L}_2$, i.e., X and Y are independent. Property (iii) follows Lemma 2. Lastly, since ψ_1 and ψ_2 are strictly increasing, we have $P(\psi_1(X) \leq \psi_1(x)) = P(X \leq x)$ and $P(\psi_2(Y) \leq \psi_2(y)) = P(Y \leq y)$. Thus Property (iv) holds. □

Remark 1. (i) It can be shown that the measure given by (7) is a discrete version of the measure $\bar{\omega}_k(Y|X)$ for random vectors given by (4) with $k = 2$. The difference here is that based on (7), we are going to define a marginal-free measure by using subcopulas for discrete random vectors.

(ii) The above measure may be simplified into

$$\mu'(Y|X) = \left[\frac{\sum_{y \in \mathcal{L}_2} \sum_{x \in \mathcal{L}_1} [P(Y \leq y|X = x) - G(y)]^2 f(x)}{\sum_{y \in \mathcal{L}_2} [G(y) - (G(y))^2]} \right]^{\frac{1}{2}}. \quad (8)$$

As indicated by Shan [10], $\mu'(Y|X)$ is well defined only if Y is a finite discrete random vector, i.e., if \mathcal{L}_2 is a finite set. Otherwise, $\sum_{y \in \mathcal{L}_2} [G(y) - (G(y))^2]$ may diverge. However, the measure $\mu(Y|X)$ given by (7) is well defined for all discrete random vectors.

(iii) It is easy to see that $\omega_{\max}^2(Y|X) = 0$ if and only if Y is a constant random vector, i.e., if and only if there is $y \in \mathbb{R}^m$ such that $P(Y = y) = 1$. In this case, Y is clearly a function of X . Thus, without loss of generality, we assume that X and Y are not constant random vectors.

Since most multivariate dependence properties of random variables can be determined by their subcopula C , we are going to redefine the measure $\mu(Y|X)$ by using subcopulas such that $\mu(Y|X)$ is free of marginal distributions of X and Y . First, note that for any $x \in \mathcal{L}_1$ and $y \in \mathcal{L}_2$, we have

$$G(y) = H(\infty_n, y) = C(1_n, v(y)), \quad (9)$$

$$f(x) = \mathcal{V}_H([(x_L, \infty_m), (x, \infty_m)]) = \mathcal{V}_C([(u(x)_L, 1_m), (u(x), 1_m)]), \quad (10)$$

$$g(y) = \mathcal{V}_H([(\infty_n, y_L), (\infty_n, y)]) = \mathcal{V}_C([(1_n, v(y)_L), (1_n, v(y))]), \quad (11)$$

and

$$\begin{aligned} P(Y \leq y|X = x) &= \frac{P(X = x, Y \leq y)}{P(X = x)} \\ &= \frac{\mathcal{V}_H([(x_L, y), (x, y)])}{\mathcal{V}_H([(x_L, \infty_m), (x, \infty_m)])} \\ &= \frac{\mathcal{V}_C([(u(x)_L, v(y)), (u(x), v(y))])}{\mathcal{V}_C([(u(x)_L, 1_m), (u(x), 1_m)])} \end{aligned} \quad (12)$$

Thus, from Eqs. (9)–(12), we can redefine $\mu(Y|X)$ as follows.

Definition 3. Let X and Y be two discrete random vectors with the subcopula C . Suppose that the domain of C is $\text{Dom}(C) = \mathcal{L}_1' \times \mathcal{L}_2'$, where $\mathcal{L}_1' \subseteq I^n$ and $\mathcal{L}_2' \subseteq I^m$. The measure of Y being completely dependent on X based on C is given by

$$\mu_C(Y|X) = \left[\frac{\omega^2(Y|X)}{\omega_{\max}^2(Y|X)} \right]^{\frac{1}{2}} = \left[\frac{\sum_{v \in \mathcal{L}_2'} \sum_{u \in \mathcal{L}_1'} \left[\frac{\mathcal{V}_C((u_L, v), (u, v))}{\mathcal{V}_C((u_L, 1_m), (u, 1_m))} - C(1_n, v) \right]^2 \mathcal{V}_C((u_L, 1_m), (u, 1_m)) \mathcal{V}_C((1_n, v_L), (1_n, v))}{\sum_{v \in \mathcal{L}_2'} [C(1_n, v) - (C(1_n, v))^2] \mathcal{V}_C((1_n, v), (1_n, v_L))} \right]^{\frac{1}{2}} \quad (13)$$

Remark 2. Based on the same idea, if X and Y are continuous random vectors with the unique copula C , the measure $\mu_C(Y|X)$ given by (13) can be rewritten as

$$\mu_C(Y|X) = \left[\frac{\int \int \left(\frac{\partial C}{\partial C_X} - C_Y \right)^2 \frac{\partial C_X}{\partial u} \frac{\partial C_Y}{\partial v} du dv}{\int C_Y (1 - C_Y) \frac{\partial C_Y}{\partial v} dv} \right]^{\frac{1}{2}},$$

where C_X and C_Y are copulas of X and Y . This is a marginal-free measure of CD for continuous random vectors.

By using $\mu_C(Y|X)$ defined in Definition 3, we can define a marginal-free measure of mutual complete dependence for two discrete random vectors as follows.

Definition 4. For any discrete random vectors X and Y with the subcopula C , the *MCD measure* of X and Y is defined by

$$\mu_C(X, Y) = \left[\frac{\omega^2(Y|X) + \omega^2(X|Y)}{\omega_{\max}^2(Y|X) + \omega_{\max}^2(X|Y)} \right]^{\frac{1}{2}}, \quad (14)$$

where $\omega^2(\cdot|\cdot)$ and $\omega_{\max}^2(\cdot|\cdot)$ are the same as those given in Definition 3.

The properties of the measure $\mu_C(X, Y)$ are given in the following theorem. The proof is straightforward.

Theorem 2. Let X and Y be two discrete random vectors with the subcopula C . The measure $\mu_C(X, Y)$ has following properties,

- (i) $\mu_C(X, Y) = \mu_C(Y, X)$;
- (ii) $0 \leq \mu_C(X, Y) \leq 1$;
- (iii) $\mu_C(X, Y) = 0$ if and only if X and Y are independent;
- (iv) $\mu_C(X, Y) = 1$ if and only if X and Y are MCD;
- (v) $\mu_C(X, Y)$ is invariant under strictly increasing transformations of X and Y .

Remark 3. (i) The insufficiency of 2-copulas to describe joint distributions with given multivariate marginal distributions was discussed by Genest et al. [4]. Let C be a 2-copula. They showed that

$$H(x_1, \dots, x_{n_1}, x_{n_1+1}, \dots, x_{n_1+n_2}) = C(H_1(x_1, \dots, x_{n_1}), H_2(x_{n_1+1}, \dots, x_{n_1+n_2}))$$

defines a $(n_1 + n_2)$ dimensional c.d.f., where $n_1 + n_2 \geq 3$, for all marginal c.d.f.'s H_1 and H_2 with dimensions n_1 and n_2 , respectively, only if C is the bivariate independence copula, i.e.,

$$C(u, v) = uv, \quad \text{for all } (u, v) \in I^2.$$

Thus, in this work, we have to use the $(n + m)$ -subcopula of (X, Y) to construct a marginal-free measure.

- (ii) Both Shan et al. [11] and Tasena and Dhompongsa [15] tried to use copulas to construct measures of functional dependence for discrete random variables or vectors. However, we do not think that copulas should be used to construct measures for discrete random variables or vectors because, for fixed discrete random variables or vectors, the corresponding copulas may not be unique. Thus, as shown in their papers, if we have different copulas for two fixed discrete random variables, copula-based measures may give us different results.
- (iii) Boonmee and Tasena [2] used linkages to construct a marginal-free measure of CD for continuous random vectors, but linkages have some defects. First, linkages are defined for continuous random vectors. Second, to find the linkage of two random vectors, they need to be transformed to uniform random vectors. It is not convenient in applications (See Li et al. [6] for more details of linkages). Thus, in this work, we prefer to use the subcopula of (X, Y) to construct marginal-free measures, since subcopulas are not only good for discrete random vectors but also more popular than linkages.
- (iv) If both X and Y are discrete random variables with the 2-subcopula C , then we have

$$\omega^2(Y|X) = \sum_{v \in \mathcal{L}_2'} \sum_{u \in \mathcal{L}_1'} \left[\frac{C(u, v) - C(u_L, v)^2}{u - u_L} - v \right]^2 (u - u_L)(v - v_L),$$

$$\omega^2(X|Y) = \sum_{u \in \mathcal{L}_1'} \sum_{v \in \mathcal{L}_2'} \left[\frac{C(u, v) - C(u, v_L)^2}{v - v_L} - u \right]^2 (u - u_L)(v - v_L),$$

$$\omega_{\max}^2(Y|X) = \sum_{v \in \mathcal{L}_2'} (v - v^2)(v - v_L) \quad \text{and} \quad \omega_{\max}^2(X|Y) = \sum_{u \in \mathcal{L}_1'} (u - u^2)(u - u_L).$$

In this case, the measure $\mu_C(X, Y) = \left[\frac{\omega^2(Y|X) + \omega^2(X|Y)}{\omega_{\max}^2(Y|X) + \omega_{\max}^2(X|Y)} \right]^{\frac{1}{2}}$ is identical to the measure given by (3) with $t = 0$.

- (v) If both X and Y are continuous random variables, i.e., $\max\{u - u_L, v - v_L\} \rightarrow 0$, then it can be show that

$$\begin{aligned} \mu_C(X, Y) &= \left[\frac{\omega^2(Y|X) + \omega^2(X|Y)}{\omega_{\max}^2(Y|X) + \omega_{\max}^2(X|Y)} \right]^{\frac{1}{2}} \\ &= \left\{ 3 \iint \left[\left(\frac{\partial C}{\partial u} \right)^2 + \left(\frac{\partial C}{\partial v} \right)^2 \right] dudv - 2 \right\}^{\frac{1}{2}}, \end{aligned}$$

which is identical to the measure given by (1).

Next, we use two examples to illustrate our above results.

Example 1. Let $X = (X_1, X_2)$ be a random vector with the distribution given in Table 1. Let $Y = (Y_1, Y_2) = (X_1^2, X_2^2)$. Then the distribution of Y , the joint distribution of X and Y , and the corresponding subcopula are given in Tables 2, 3 and 4, respectively. It is easy to show that $\omega^2(Y|X) = \omega_{\max}^2(Y|X) = 161/1458$, $\omega^2(X|Y) = 2699/38880$, $\omega_{\max}^2(X|Y) = 469/2916$. So $\mu_C(Y|X) = 1$, $\mu_C(X|Y) = 0.6569$ and $\mu_C(X, Y) = 0.8142$.

Table 1. Distribution of X .

X_2	X_1			X_2
	-1	0	1	
-1	1/18	2/18	3/18	6/18
0	1/18	2/18	2/18	5/18
1	1/18	3/18	3/18	7/18
X_1	3/18	7/18	8/18	1

Table 2. Distribution of Y .

Y_2	Y_1		Y_2
	0	1	
0	2/18	3/18	5/18
1	5/18	8/18	13/18
Y_1	7/18	11/18	1

Table 3. Joint distribution of X and Y .

Y	X									Y
	(-1,1)	(-1,0)	(-1,-1)	(0,-1)	(0,0)	(0,1)	(1,-1)	(1,0)	(1,1)	
(0,0)	0	0	0	0	$\frac{2}{18}$	0	0	0	0	$\frac{2}{18}$
(0,1)	0	0	0	$\frac{2}{18}$	0	$\frac{3}{18}$	0	0	0	$\frac{5}{18}$
(1,0)	0	$\frac{1}{18}$	0	0	0	0	0	$\frac{2}{18}$	0	$\frac{3}{18}$
(1,1)	$\frac{1}{18}$	0	$\frac{1}{18}$	0	0	0	$\frac{3}{18}$	0	$\frac{3}{18}$	$\frac{8}{18}$
X	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{2}{18}$	$\frac{2}{18}$	$\frac{3}{18}$	$\frac{3}{18}$	$\frac{2}{18}$	$\frac{3}{18}$	1

Table 4. Subcopula of X and Y .

V	U								$(1,1)$
	$(\frac{3}{18}, \frac{6}{18})$	$(\frac{3}{18}, \frac{11}{18})$	$(\frac{3}{18}, 1)$	$(\frac{10}{18}, \frac{6}{18})$	$(\frac{10}{18}, \frac{11}{18})$	$(\frac{10}{18}, 1)$	$(1, \frac{6}{18})$	$(1, \frac{11}{18})$	
$(\frac{7}{18}, \frac{5}{18})$	0	0	0	0	$\frac{2}{18}$	$\frac{2}{18}$	0	$\frac{2}{18}$	$\frac{2}{18}$
$(\frac{7}{18}, 1)$	0	0	0	$\frac{2}{18}$	$\frac{4}{18}$	$\frac{7}{18}$	$\frac{2}{18}$	$\frac{4}{18}$	$\frac{7}{18}$
$(1, \frac{5}{18})$	0	$\frac{1}{18}$	$\frac{1}{18}$	0	$\frac{3}{18}$	$\frac{3}{18}$	0	$\frac{5}{18}$	$\frac{5}{18}$
$(1, 1)$	$\frac{1}{18}$	$\frac{2}{18}$	$\frac{3}{18}$	$\frac{3}{18}$	$\frac{6}{18}$	$\frac{10}{18}$	$\frac{6}{18}$	$\frac{11}{18}$	1

Example 2. Let $X = (X_1, X_2)$ be a discrete random vector, where X_1 is a geometric random variable with the success rate $p = \frac{1}{2}$, X_2 is a binomial random variable with the number of trials $n = 2$ and the success rate $p = \frac{1}{2}$, and X_1 and X_2 are independent. Let $Y = X_1 - X_2$. Then the joint distribution and subcopula of X and Y are given in Tables 5 and 6. By calculation, $\omega^2(Y|X) = \omega_{\max}^2(Y|X) = 1223/7168$, $\omega^2(X|Y) = 3407543/30965760$ and $\omega_{\max}^2(X|Y) = 1/3$. So $\mu_C(Y|X) = 1$, $\mu_C(X|Y) = 0.3301$ and $\mu_C(X, Y) = 0.5569$.

Table 5. Joint distribution of X and Y .

X	Y								\dots	X
	-1	0	1	2	3	4	5			
(1,0)	0	0	$\frac{1}{2^3}$	0	0	0	0			$\frac{1}{2^3}$
(1,1)	0	$\frac{1}{2^2}$	0	0	0	0	0			$\frac{1}{2^2}$
(1,2)	$\frac{1}{2^3}$	0	0	0	0	0	0			$\frac{1}{2^3}$
(2,0)	0	0	0	$\frac{1}{2^4}$	0	0	0			$\frac{1}{2^4}$
(2,1)	0	0	$\frac{1}{2^3}$	0	0	0	0			$\frac{1}{2^3}$
(2,2)	0	$\frac{1}{2^4}$	0	0	0	0	0			$\frac{1}{2^4}$
(3,0)	0	0	0	0	$\frac{1}{2^5}$	0	0			$\frac{1}{2^5}$
(3,1)	0	0	0	$\frac{1}{2^4}$	0	0	0			$\frac{1}{2^4}$
(3,2)	0	0	$\frac{1}{2^3}$	0	0	0	0			$\frac{1}{2^3}$
(4,0)	0	0	0	0	0	$\frac{1}{2^6}$	0			$\frac{1}{2^6}$
(4,1)	0	0	0	0	$\frac{1}{2^5}$	0	0			$\frac{1}{2^5}$
(4,2)	0	0	0	$\frac{1}{2^6}$	0	0	0			$\frac{1}{2^6}$
(5,0)	0	0	0	0	0	0	$\frac{1}{2^7}$			$\frac{1}{2^7}$
(5,1)	0	0	0	0	0	$\frac{1}{2^6}$	0			$\frac{1}{2^6}$
(5,2)	0	0	0	0	$\frac{1}{2^7}$	0	0			$\frac{1}{2^7}$
\vdots										\vdots
Y	$\frac{1}{2^3}$	$\frac{2^2+1}{2^4}$	$\frac{2^3+1}{2^5}$	$\frac{2^3+1}{2^6}$	$\frac{2^3+1}{2^7}$	$\frac{2^3+1}{2^8}$	$\frac{2^3+1}{2^9}$	\dots		1

Table 6. Subcopula of X and Y .

U	V					...
	$\frac{1}{2^3}$	$\frac{2^2+2+1}{2^4}$	$\frac{2^4+2^2+2+1}{2^5}$	$\frac{2^5+2^4+2^2+2+1}{2^6}$	$\frac{2^6+2^5+2^4+2^2+2+1}{2^7}$	
$(\frac{1}{2}, \frac{1}{2^2})$	0	0	$\frac{1}{2^3}$	$\frac{1}{2^3}$	$\frac{1}{2^3}$	
$(\frac{1}{2}, \frac{3}{2^2})$	0	$\frac{1}{2^2}$	$\frac{2+1}{2^3}$	$\frac{2+1}{2^3}$	$\frac{2+1}{2^3}$	
$(\frac{1}{2}, 1)$	$\frac{1}{2^3}$	$\frac{2+1}{2^3}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	
$(\frac{2+1}{2^2}, \frac{1}{2^2})$	0	0	$\frac{1}{2^3}$	$\frac{2+1}{2^3}$	$\frac{2+1}{2^3}$	
$(\frac{2+1}{2^2}, \frac{3}{2^2})$	0	$\frac{1}{2^2}$	$\frac{1}{2}$	$\frac{2^2+1}{2^4}$	$\frac{2^2+1}{2^4}$	
$(\frac{2+1}{2^2}, 1)$	$\frac{1}{2^3}$	$\frac{2^2+2+1}{2^4}$	$\frac{6+2^2+1}{2^4}$	$\frac{2+1}{2^2}$	$\frac{2+1}{2^2}$	
$(\frac{2^2+2+1}{2^2}, \frac{1}{2^2})$	0	0	$\frac{1}{2^3}$	$\frac{2+1}{2^4}$	$\frac{2^2+2+1}{2^5}$	
$(\frac{2^2+2+1}{2^2}, \frac{3}{2^2})$	0	$\frac{1}{2^2}$	$\frac{1}{2}$	$\frac{2^2+1}{2^3}$	$\frac{2^4+2^2+1}{2^5}$	
$(\frac{2^2+2+1}{2^2}, 1)$	$\frac{1}{2^3}$	$\frac{2^2+2+1}{2^4}$	$\frac{2^4+2^2+2+1}{2^5}$	$\frac{2^4+2^3+2+1}{2^5}$	$\frac{2^2+2+1}{2^3}$	
...						

4 Estimators of $\mu(Y|X)$ and $\mu(X, Y)$

In this section, we are going to construct estimators of measures $\mu(Y|X)$ and $\mu(X, Y)$. Let $X \in \mathcal{L}_1$ and $Y \in \mathcal{L}_2$ be two discrete random vectors and $[n_{xy}]$ be their observed multi-way contingency table. Suppose that the total number of observation is n . For every $x \in \mathcal{L}_1$ and $y \in \mathcal{L}_2$, let n_{xy} , n_x and n_y be numbers of observations of (x, y) , x and y , respectively, i.e., $n_x = \sum_{y \in \mathcal{L}_2} n_{xy}$ and $n_y = \sum_{x \in \mathcal{L}_1} n_{xy}$. If we define $\hat{p}_{xy} = n_{xy}/n$, $\hat{p}_x = n_x/n$, $\hat{p}_y = n_y/n$, $\hat{p}_{y|x} = \hat{p}_{xy}/\hat{p}_x = n_{xy}/n_x$ and $\hat{p}_{x|y} = \hat{p}_{xy}/\hat{p}_y = n_{xy}/n_y$, then estimators of measures $\mu(Y|X)$ and $\mu(X, Y)$ can be defined as follows.

Definition 5. Let $X \in \mathcal{L}_1$ and $Y \in \mathcal{L}_2$ be two discrete random vectors with a multi-way contingency table $[n_{xy}]$. Estimators of $\mu(Y|X)$ and $\mu(X, Y)$ are given by

$$\hat{\mu}(Y|X) = \left[\frac{\hat{\omega}^2(Y|X)}{\hat{\omega}_{\max}^2(Y|X)} \right]^{\frac{1}{2}},$$

and

$$\hat{\mu}(X, Y) = \left[\frac{\hat{\omega}^2(Y|X) + \hat{\omega}^2(X|Y)}{\hat{\omega}_{\max}^2(Y|X) + \hat{\omega}_{\max}^2(X|Y)} \right]^{\frac{1}{2}},$$

where

$$\begin{aligned} \hat{\omega}^2(Y|X) &= \sum_{y \in \mathcal{L}_2} \sum_{x \in \mathcal{L}_1} \left[\sum_{y' \leq y} (\hat{p}_{y'|x} - \hat{p}_{y'}) \right]^2 \hat{p}_x \hat{p}_y, \\ \hat{\omega}_{\max}^2(Y|X) &= \sum_{y \in \mathcal{L}_2} \left[\sum_{y' \leq y} \hat{p}_{y'} - \left(\sum_{y' \leq y} \hat{p}_{y'} \right)^2 \right] \hat{p}_y, \end{aligned}$$

and $\hat{\omega}^2(X|Y)$ and $\hat{\omega}_{\max}^2(X|Y)$ are similarly defined as $\hat{\omega}^2(Y|X)$ and $\hat{\omega}_{\max}^2(Y|X)$ by interchanging X and Y .

From the above definition, we have the following result. The proof is trivial.

Theorem 3. Let X and Y be discrete random vectors. Estimators $\hat{\mu}(Y|X)$ and $\hat{\mu}(X, Y)$ have following properties,

- (i) $0 \leq \hat{\mu}(Y|X), \hat{\mu}(X, Y) \leq 1$;
- (ii) X and Y are empirically independent, i.e., $\hat{p}_{xy} = \hat{p}_x \cdot \hat{p}_y$ for all $(x, y) \in \mathcal{L}_1 \times \mathcal{L}_2$ if and only if $\hat{\mu}(Y|X) = \hat{\mu}(X, Y) = 0$;
- (iii) $\hat{\mu}(Y|X) = 1$ if and only if Y is a function of X . And $\hat{\mu}(X, Y) = 1$ if and only if X and Y are functions of each other.

Next, we use two example to illustrate our results.

Example 3. Suppose that we have the following multi-way contingency tables. Then from Table 7, we have $\hat{\mu}(Y|X) = 0.0516$, $\hat{\mu}(X|Y) = 0.0762$ and $\hat{\mu}(X, Y) = 0.0642$, so X and Y have very weak functional relations. However, from Table 8, we have $\hat{\mu}(Y|X) = 0.5746$, $\hat{\mu}(X|Y) = 0.0465$ and $\hat{\mu}(X, Y) = 0.3485$, so the functional dependence of Y on X is much stronger than the functional dependence of X on Y .

Example 4. The data given in Table 9 [1] is from a survey conducted by the Wright State University School of Medicine and the United Health Services in Dayton, Ohio. The survey asked students in their final year of a high school near Dayton, Ohio,

Table 7. Contingency table of X and Y .

Y	X				$n_{\cdot y}$
	(1, 1)	(1, 2)	(2, 1)	(2, 2)	
(1, 1)	10	20	5	10	45
(1, 2)	15	25	10	5	55
(2, 1)	5	35	10	5	55
(2, 2)	25	5	10	5	45
$n_{x \cdot}$	55	85	35	25	200

Table 8. Contingency table of X and Y .

Y	X				$n_{\cdot y}$
	(1, 1)	(1, 2)	(2, 1)	(2, 2)	
(1, 1)	43	2	3	40	88
(1, 2)	4	42	40	6	92
(2, 1)	2	3	3	2	10
(2, 2)	1	4	2	3	10
$n_{x \cdot}$	50	51	48	51	200

Table 9. Alcohol (*A*), Cigarette (*C*), and Marijuana (*M*) use for high school seniors.

Alcohol use	Cigarette use	Marijuana use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

whether they had ever used alcohol, cigarettes, or marijuana. Denote the variables by *A* for alcohol use, *C* for cigarette use, and *M* for marijuana use. By Pearson's Chi-squared test (*A*, *C*) and *M* are not independent. The estimations of functional dependence between *M* and (*A*, *C*) are $\hat{\mu}(M|(A, C)) = 0.3097$, $\hat{\mu}((A, C)|M) = 0.2776$ and $\hat{\mu}((A, C), M) = 0.2893$.

Acknowledgments. We would like to thank Professor Hung T. Nguyen for introducing this interesting topic to us and referees for their valuable comments and suggestions which greatly improve this paper.

References

1. Agresti, A., Kateri, M.: *Categorical Data Analysis*, 2nd edn. Springer, Hoboken (2011)
2. Boonmee, T., Tasena, S.: Measure of complete dependence of random vectors. *J. Math. Anal. Appl.* **443**(1), 585–595 (2016)
3. Dette, H., Siburg, K.F., Stoimenov, P.A.: A copula-based non-parametric measure of regression dependence. *Scand. J. Stat.* **40**(1), 21–41 (2013)
4. Genest, C., Quesada Molina, J., Rodríguez Lallena, J.: De l'impossibilité de construire des lois à marges multidimensionnelles données à partir de copules. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique* **320**(6), 723–726 (1995)
5. Lancaster, H.: Measures and indices of dependence. In: Kotz, S., Johnson, N.L. (eds.) *Encyclopedia of Statistical Sciences*, vol. 2, pp. 334–339. Wiley, New York (1982)
6. Li, H., Scarsini, M., Shaked, M.: Linkages: a tool for the construction of multivariate distributions with given nonoverlapping multivariate marginals. *J. Multivar. Anal.* **56**(1), 20–41 (1996)
7. Nelsen, R.B.: *An Introduction to Copulas*, 2nd edn. Springer, New York (2006)
8. Rényi, A.: On measures of dependence. *Acta Math. Hung.* **10**(3–4), 441–451 (1959)
9. Schweizer, B., Wolff, E.F.: On nonparametric measures of dependence for random variables. *Ann. Stat.* **9**(4), 879–885 (1981)
10. Shan, Q.: *The measures of association and dependence through copulas*. Ph.D. dissertation, New Mexico State University (2015)
11. Shan, Q., Wongyang, T., Wang, T., Tasena, S.: A measure of mutual complete dependence in discrete variables through subcopula. *Int. J. Approx. Reason.* **65**, 11–23 (2015)
12. Siburg, K.F., Stoimenov, P.A.: A measure of mutual complete dependence. *Metrika* **71**(2), 239–251 (2010)

13. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Stat. Univ. Paris **8**, 229–231 (1959)
14. Tasena, S., Dhompongsa, S.: A measure of multivariate mutual complete dependence. Int. J. Approx. Reason. **54**(6), 748–761 (2013)
15. Tasena, S., Dhompongsa, S.: Measures of the functional dependence of random vectors. Int. J. Approx. Reason. **68**, 15–26 (2016)
16. Trutschnig, W.: On a strong metric on the space of copulas and its induced dependence measure. J. Math. Anal. Appl. **384**(2), 690–705 (2011)