

AN ISING SIMILARITY REGRESSION MODEL FOR MODELING MULTIVARIATE BINARY DATA

Zhi Yang Tho*, Francis K. C. Hui and Tao Zou*

The Australian National University

Abstract: Understanding the dependence structure between response variables is an important component in the analysis of correlated multivariate data. This article focuses on modeling dependence structures in multivariate binary data, motivated by a study aiming to understand how patterns in different U.S. senators' votes are determined by similarities (or lack thereof) in their attributes, e.g., political parties and social network profiles. To address such a research question, we propose a new Ising similarity regression model which regresses pairwise interaction coefficients in the Ising model against a set of similarity measures available/constructed from covariates. Model selection approaches are further developed through regularizing the pseudo-likelihood function with an adaptive lasso penalty to enable the selection of relevant similarity measures. We establish estimation and selection consistency of the proposed estimator under a general setting where the number of similarity measures and responses tend to infinity. Simulation study demonstrates the strong finite sample performance of the proposed estimator, particularly compared with several existing Ising model estimators in estimating the matrix of pairwise interaction coefficients. Applying the Ising similarity regression model to a dataset of roll call voting records of 100 U.S. senators, we are able to quantify how similarities in senators' parties, businessman occupations and social network profiles drive their voting associations.

Key words and phrases: Conditional dependence, Ising model, Lasso, model selection, multivariate data, pseudo-likelihood.

1. Introduction

The study of the dependence structure in correlated multivariate data has drawn much attention in recent years, due to the increasing accessibility of multi-response data across many disciplines such as finance (Tsay, 2013), economics (McElroy and Trimbur, 2023) and ecology (Hui, Müller and Welsh, 2023). For a multivariate response vector $\mathbf{y} = (y_1, \dots, y_p)^\top$, the dependencies across any set of two components y_j and $y_{j'}$ for $j, j' = 1, \dots, p$ can be modeled by specifying a joint probabilistic distribution for $\mathbf{y}$. Two prominent examples are Gaussian graphical models (Whittaker, 1990) and Ising models (Ising, 1925), which specify a probability density function (pdf) and a probability mass function (pmf) for a continuous and binary response vector $\mathbf{y}$, respectively, where the pdf and pmf of

*Corresponding author. E-mail: (Tho) zhiyang.tho@anu.edu.au; (Zou) tao.zou@anu.edu.au